Lecture Notes in Electrical Engineering 301

James J. (Jong Hyuk) Park
Albert Zomaya
Hwa-Young Jeong
Mohammad Obaidat
Editors

Frontier and Innovation in Future Computing and Communications

Springer

Lecture Notes in Electrical Engineering

Volume 301

Board of Series Editors

Leopoldo Angrisani, Napoli, Italy
Marco Arteaga, Coyoacán, México
Samarjit Chakraborty, München, Germany
Jiming Chen, Hangzhou, P.R. China
Tan Kay Chen, Singapore, Singapore
Rüdiger Dillmann, Karlsruhe, Germany
Gianluigi Ferrari, Parma, Italy
Manuel Ferre, Madrid, Spain
Sandra Hirche, München, Germany
Faryar Jabbari, Irvine, USA
Janusz Kacprzyk, Warsaw, Poland
Alaa Khamis, New Cairo City, Egypt
Torsten Kroeger, Stanford, USA
Tan Cher Ming, Singapore, Singapore
Wolfgang Minker, Ulm, Germany
Pradeep Misra, Dayton, USA
Sebastian Möller, Berlin, Germany
Subhas Mukhopadyay, Palmerston, New Zealand
Cun-Zheng Ning, Tempe, USA
Toyoaki Nishida, Sakyo-ku, Japan
Federica Pascucci, Roma, Italy
Tariq Samad, Minneapolis, USA
Gan Woon Seng, Nanyang Avenue, Singapore
Germano Veiga, Porto, Portugal
Junjie James Zhang, Charlotte, USA

For further volumes:
http://www.springer.com/series/7818

About this Series

''Lecture Notes in Electrical Engineering (LNEE)'' is a book series which reports the latest research and developments in Electrical Engineering, namely:

- Communication, Networks, and Information Theory
- Computer Engineering
- Signal, Image, Speech and Information Processing
- Circuits and Systems
- Bioengineering

LNEE publishes authored monographs and contributed volumes which present cutting edge research information as well as new perspectives on classical fields, while maintaining Springer's high standards of academic excellence. Also considered for publication are lecture materials, proceedings, and other related materials of exceptionally high quality and interest. The subject matter should be original and timely, reporting the latest research and developments in all areas of electrical engineering.

The audience for the books in LNEE consists of advanced level students, researchers, and industry professionals working at the forefront of their fields. Much like Springer's other Lecture Notes series, LNEE will be distributed through Springer's print and electronic publishing channels.

James J. (Jong Hyuk) Park
Albert Zomaya · Hwa-Young Jeong
Mohammad Obaidat
Editors

Frontier and Innovation in Future Computing and Communications

Springer

Editors
James J. (Jong Hyuk) Park
Department of Computer Science and Engineering
Seoul National University of Science and Technology (SeoulTech)
Seoul
Korea, Republic of South Korea

Albert Zomaya
School of Information Technologies
University of Sydney
Sydney, NSW
Australia

Hwa-Young Jeong
Humanitas College
Kyung Hee University
Seoul
Korea, Republic of South Korea

Mohammad Obaidat
Computer Science and Software Engineering
Monmouth University
West Long Branch, NJ
USA

ISSN 1876-1100 ISSN 1876-1119 (electronic)
ISBN 978-94-017-8797-0 ISBN 978-94-017-8798-7 (eBook)
DOI 10.1007/978-94-017-8798-7
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2014932532

© Springer Science+Business Media Dordrecht 2014
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Message from the FCC 2014 General Chairs

The 2014 FTRA International Symposium on Frontier and Innovation in Future Computing and Communications (FCC 2014) is an event of the series of international scientific conferences. This conference takes place in Auckland, New Zealand, January 13–16, 2014. The FCC-14 will be the most comprehensive conference focused on the various aspects of advances in Future Computing and Communications. The FCC-14 will provide an opportunity for academic and industry professionals to discuss the latest issues and progress in the area of the future computing and communications. In addition, the conference will publish high-quality papers which are closely related to the various theories and practical applications in FCC. Furthermore, we expect that the conference and its publications will be a trigger for further related research and technology improvements in this important subject.

The papers included in the proceedings cover the following topics: Mobile and ubiquitous computing, Dependable, reliable and autonomic computing, Security and trust management, Multimedia systems and services, Networking and communications, Database and data mining, UI/UX for human-centric computing, Grid and scalable computing, Embedded system and software, Artificial intelligence, Distributed and parallel algorithms, Web and Internet computing, and IT policy and business management.

Accepted and presented papers highlight new trends and challenges of Future Computing and Communications. The presenters showed how new research could lead to novel and innovative applications. We hope you will find these results useful and inspiring for your future research.

We would like to express our sincere thanks to the Program Chairs: Byung-Gyu Kim (SunMoon University, Korea), Naveen Chilamkurti (La Trobe University, Australia), Jason C. Hung (Oversea Chinese University, Taiwan), and Mohamed Gaber (Robert Gordon University, UK), all Program Committee members and all the additional reviewers for their valuable efforts in the review process, which helped us to guarantee the highest quality of the selected papers for the conference. Our special thanks go to the invited speaker who kindly accepted our invitations, and helped to meet the objectives of the conference: Prof. Neil Y. Yen, The University of Aizu, Japan.

We cordially thank all the authors for their valuable contributions and the other participants of this conference. The conference would not have been possible without their support. Thanks are also due to the many experts who contributed to making the event a success.

January 2014
James J. Park, SeoulTech, Korea
Albert Zomaya, University of Sydney, Australia
Young-Sik Jeong, Dongguk University, Korea
Mohammad S. Obaidat, Monmouth University, USA
FCC 2014 General Chairs

Message from the FCC 2014 Program Chairs

On behalf of the Organizing Committee, it is our pleasure to welcome you to the 2014 FTRA International Symposium on Frontier and Innovation in Future Computing and Communications (FCC 2014) in Auckland, New Zealand. The main goal of this international symposium is to provide a platform for researchers, scientists, and engineers from academia and industry working in the area of Ubiquitous wired and wireless networks, Intelligent sensor network/RFID, Ubiquitous/pervasive platform and middleware, Mobile and ubiquitous multicasting, Wireless communication technology, Multimedia communication, and Future Internet. This year, lots of papers were submitted and results based on the review obtained, the program committee has finally accepted excellent papers for presentation at the conference. The technical program for this year cover a range of topics such as Air Interface, sensor networks, ad-hoc networks, wireless sensor networks, wireless computing, multimedia signal processing, and vehicular communications.

First of all, we would like to express our sincere thanks to General Chairs: James J. Park (SeoulTech, Korea), Albert Zomaya (University of Sydney, Australia), Young-Sik Jeong (Dongguk University, Korea), and Mohammad S. Obaidat (Monmouth University, USA). Also, we would like to thank all authors who submitted their papers for consideration and our congratulations to all those who have had their papers accepted after a rigorous peer review selection process. We express our sincere gratitude to the program committee members and additional reviewers who all worked hard in reviewing the submitted papers to ensure we have a high-quality program. We would like to thank the following sponsors for their great support: Korea Information Processing Society, FTRA, and KITCS. After closing the conference, all accepted and presented papers in FCC-14 will be

published by Post Proceeding-Lecture Notes in Electrical Engineering (LNEE)—Springer (indexed by EI and SCOPUS). Expected accept rate for SIs from full paper is about 30 ~ 40 %. We hope you will find this symposium interesting and we wish you all a nice and enjoyable stay in beautiful Auckland, New Zealand.

January 2014
Byung-Gyu Kim, SunMoon University, Korea
Naveen Chilamkurti, La Trobe University, Australia
Jason C. Hung, Oversea Chinese University, Taiwan
Mohamed Gaber, Robert Gordon University, UK

Organization

General Chairs

James J. Park, SeoulTech, Korea
Albert Zomaya, University of Sydney, Australia
Young-Sik Jeong, Dongguk University, Korea
Mohammad S. Obaidat, Monmouth University, USA

General Vice-Chairs

Raj Buyya, University of Melbourne, Australia
Hwa-Young Jeong, Kyung Hee University, Korea

Program Chairs

Byung-Gyu Kim, Sunmoon University, Korea
Naveen Chilamkurti, La Trobe University, Australia
Jason C. Hung, Oversea Chinese University, Taiwan
Mohamed Gaber, Robert Gordon University, UK

FTRA Advisory Board (Selected Committee)

Alan Chin-Chen Chang, National Chung Cheng University, Taiwan
C. V. Ramamoorthy, University of California at Berkeley, USA
Habib F. Rashvand, University of Warwick, UK
Han-Chieh Chao, National Ilan University, Taiwan
Hong Shen, University of Adelaide, Australia
Jianhua Ma, Hosei University, Japan

Laurence T. Yang, St. Francis Xavier University, Canada
Sethuraman Panchanathan, Arizona State University, USA
Shu-Ching Chen, Florida International University, USA
Wanlei Zhou, Deakin University, Australia
Yi Pan, Georgia State University, USA
Young-Sik Jeong, Dongguk University, Korea
Makoto Takizawa, Seikei University, Japan
Doo-Soon Park, SoonChunHyang University, Korea
C. S. Raghavendra, University of Southern California, USA
Hamid R. Arabnia, The University of Georgia, USA
Ivan Stojmenovic, University of Ottawa, Canada
Yunho Kim, Mokwon University, Daejeon, Korea

Publicity Chairs

Eunyoung Lee, Dongduk Women's University, Korea
Weiwei Fang, Beijing Jiaotong University, China
Anna Kang, Donguk University, Korea
Young-Ae Jung, Sunmoon University, Korea
J. H. Abawajy, Deakin University, Australia
Ford Lumban Gaol, Bina Nusantara University
Jiankun Hu, RMIT University, Australia

Local Arrangement Chairs

Sang-Ki Seok, SeoulTech, Korea
Su Wan Park, ETRI, Korea
Taek Young Youn, ETRI, Korea

Invited Speaker

Neil Y. Yen, The University of Aizu, Japan

Program Committee

Akshaye Dhawan, Ursinus College, USA
Amiya Nayak, University of Ottawa, Canada
Angel D. Sappa, Universitat Autonoma de Barcelona, Spain
Bartosz Ziolko, AGH University of Science and Technology
Bok-Suk Shin, The University of Auckland, New Zealand
Changsun Shin, Sunchon National University
Chao-Chin Wu, National Changhua University of Education, Taiwan
Chao-Tung Yang, Tunghai University, Taiwan
Christian Fensch, Heriot-Watt University, UK
Dakshina Ranjan Kisku, Asansol Engineering College, India
Dariusz Frejlichowski, West Pomeranian University of Technology
Debi P. Dogra, IIT Bhubaneswar, India
Debzani Deb, Winston-Salem State University, USA
Deqing Zou, Huazhong University of Science and Technology
Dhavy Gantsou, Universite de Valenciennes, France
DongSan Jun, ETRI, Korea
Eleni Karatza, Aristotle University of Thessaloniki, Greece
Gaurav Singh, CSIRO, Australia
Guandong Xu, University of Technology Sydney, Australia
Guohua Jin, AMD, USA
Hae Young Lee, ETRI, Korea
Hai Jin, Huazhong University of Science and Technology
Hangzai Luo, East China Normal University, China
Hari Om, Indian School of Mines University, India
HongJun Jia, Cognex, USA
Hsing-Lung Chen, National Taiwan University of Science and Technology, Taiwan
Ishfaq Ahmad, University of Texas at Arlington, USA
J. Guha, IIT Bhubaneswar, India
Jianxun Liu, Hunan University of Science and Technology, China
Jong-Kook Kim, Korea University, Korea
JoongHoon Kim, KAIST, Korea
Jungong Han, Civolution Technology, The Netherlands
Jun-Won Ho, Seoul Women's University, Korea
K. C. Lai, National Taichung University, Taiwan
Keqin Li, State University of New York, USA
Kilhung Lee, Seoul National University of Science
Kostas Psannis, University of Macedonia, Greece
Krishna Reddy, Torrent University, Italy
Kye-Shin Lee, University of Akron, USA
Laurence T. Yang, St. Francis Xavier University, Canada
Leonel Sousa, INESC/IST, Portugal

Long Wang, IBM Thomas J. Watson Research Center, USA
Luis Javier Garcia Villalba, Universidad Complutense de Madrid (UCM), Spain
Lukas Ruf, Consecom AG, Switzerland
Mario Doeller, University of Applied Science, Germany
Masaru Miyazaki, NHK, Japan
Maytham Safar, Kuwait University, Kuwait
Michal Wozniak, Wrocław University of Technology, Poland
Ning Zhou, University of North Carolina, USA
Oliver Sinnen, University of Auckland, New Zealand
Quanqing Xu, Data Storage Institute, A*STAR, Singapore
Rachid Anane, Coventry University, UK
Rajkumar Kannan, Bishop Heber College, India
Ralph Duncan, CloudShield Technologies, USA
Reinhard Klette, The University of Auckland, New Zealand
Sagarmay Deb, University of Southern Queensland, Australia
Se-Hak Chun, Seoul National University of Science
Seung-Ho Lim, Hankuk University of Foreign Studies
Shi-Jinn Horng, National Taiwan University of Science and Technology, Taiwan
Sokratis Katsikas, University of Piraeus, Greece
Subramanian Ramanathan, ADS Center, Singapore
Tae-Jung Kim, ETRI, Korea
Teng Li, Baidu Inc., China
Teofilo F. Gonzalez, UC Santa Babara, USA
Toshihiro Yamauchi, Okayama University, Japan
Vargas-Vera Maria, Universidad Adolfo Ibanez, Chile
Wee Siong, Institute for Infocomm Research
Weifeng Chen, California University of Pennsylvania
Wen-Chung Shih, Asia University, University
Wenguang Chen, Tsinghua University, China
Yijie Han, University of Missouri Kansas City, USA
Yu Su, Ohio State University, USA
Yungchin Fang, Dell, USA

Message from the KCIC 2013 General Chairs

KCIC 2013 is the KITCS and FTRA eighth event of the series of scientific conference. This conference takes place on December 27–28, 2013, in Seoul Korea. The aim of the KCIC 2013 was to provide an international forum for scientific research in the technologies and application of Multimedia and Ubiquitous Engineering. It was organized by the Korea Information Technology Convergence Society in cooperation with Future Technology Research Association International.

The papers included in the proceedings cover the following topics: Ubiquitous Networks and Wireless Communications and Digital Convergence and Internet of Thing/M2M and Multimedia Applications and Services and Big data and Social Computing and Smart environments and U-city and Cloud Computing and Smart Grid and Vehicular communications and Intelligent and Pervasive Applications and Security and Trust Computing and IT Management and Service and Intelligent Transportation Systems and Bioinformatics and Bio-Inspired Computing and Hybrid of IT/BT, IT/NT and U-healthcare and Vehicle/Ship-build/Construction and UI/UX and Gestures Recognition/Voice Recognition/Face Recognition and Hand Gesture Recognition Interaction and Users tracking sensor Technology and Space User Interface Technology and Location-based services Accepted and presented papers highlight new trends and challenges of IT Convergence. The presenters showed how new research could lead to novel and innovative applications. We hope you will find these results useful and inspiring for your future research.

We would like to express our sincere thanks to Program Chairs: Hwa-Young Jeong (Kyung Hee University, Korea), Deok Gyu Lee (ETRI, Korea), all Program Committee members and all the additional reviewers for their valuable efforts in the review process, which helped us to guarantee the highest quality of the selected papers for the conference.

We cordially thank all the authors for their valuable contributions and the other participants of this conference. The conference would not have been possible without their support. Thanks are also due to the many experts who contributed to making the event a success.

January 2014
James J. Park, SeoulTech, Korea
Young-Sik Jeong, Dongguk University, Korea
KCIC 2013 General Chairs

Message from the KCIC 2013 Program Chairs

Welcome to the 2013 winter KITCS and FTRA Convergence IT Conference (KCIC 2013 Winter), which will be held in Seoul, Korea on December 27–28, 2013. KCIC 2013 will the most comprehensive conference focused on the various aspects of IT convergence. KCIC 2013 will provide an opportunity for academic and industry professionals to discuss recent progress in the area of IT convergence. In addition, the conference will publish high-quality papers which are closely related to the various theories and practical applications in IT convergence. Furthermore, we expect that the conference and its publications will be a trigger for further related research and technology improvements in these important subjects.

For KCIC 2013, we received many paper submissions, after a rigorous peer review process, we accepted 38 articles with high quality for the KCIC 2013 proceedings, published by the Springer. All submitted papers have undergone blind reviews by at least four reviewers from the technical program committee, which consists of leading researchers around the globe. Without their hard work, achieving such a high-quality proceeding would not have been possible. We take this opportunity to thank them for their great support and cooperation. We would like to sincerely thank the following invited speaker who kindly accepted our invitations, and, in this way, helped to meet the objectives of the conference: Prof. James J. Park, SeoulTech University, Korea and Young-Sik Jeong, Dongguk University, Korea. Finally, we would like to thank all of you for your participation in our conference, and also thank all the authors, reviewers, and organizing committee members.

January 2014
Hwa-Young Jeong, Kyung Hee University, Korea
Deok Gyu Lee, ETRI, Korea
KCIC 2013 Program Chairs

Organization

General Chairs

James J. Park, SeoulTech, Korea
Young-Sik Jeong, Dongguk University, Korea

Program Chairs

Hwa-Young Jeong, Kyung Hee University, Korea
Deok Gyu Lee, ETRI, Korea

Publicity Chairs

Byung-Gyu Kim, Sunmoon University, Korea
Anna Kang, Donguk University, Korea
Young-Ae Jung, Sunmoon University, Korea

Publication Chairs

Gangman Yi, Gangneung-Wonju National University, Korea
Sung-Ki Kim, Sunmoon University, Korea

Program Committee

EungNam Ko, Baekseok University, Korea
Jin Kwak, SoonChunHyang University, Korea

Gihwon Kwo, Kyonggi University
Joon-Min Gil, Catholic University of Daegu, Korea
Dong-Gun Kim, Kyung Hee University, Korea
Cheonshik Kim, Sejong University, Korea
SungSuk Kim, Seokyung University, Korea
Hanmin Jung, KISTI, Korea
Min Choi, Chungbuk National University, Korea
Namje Park, Jeju National University, Korea
Seung-Ho Lim, Hankuk University of Foreign Studies, Korea
Young-Ho Park, Sookmyung Women's University, Korea
Su-wan Park, ETRI, Korea
Daesung Moon, ETRI, Korea
HeonChang Yu, Korea University, Korea
HwaMin Lee, SoonChunHyang University, Korea
Soon-seok Kim, Halla University, Korea
Young-Gab Kim, Catholic University of Daegu, Korea
Jae Saeng Kim, Kimpo University, Korea
Jin-Mook Kim, Sunmoon University, Korea
Hyunsung Kim, Kyungil University, Korea
Huy Kang Kim, Korea University, Korea
Won Woo Ro, Yonsei University, Korea
Ki-Seo Park, Daelim University, Korea
Ki-Chun Bang, Namseoul University, Korea
Sa-kwang Song, Korea Institute of Science and Technology Information, Korea
Chang Sun Shin, Sunchon University, Korea
Jae Sup Oh, Kyung Hee University, Korea
Jaesoo Yoo, Chungbuk National University, Korea
Taejong Yoo, Sangmyung University, Korea
Eun-Jun Yoon, Kyungil University, Korea
Sung-Wook Lee, Korea Institute of Science and Technology Information, Korea
Eun-Ser Lee, Andong University, Korea
Eunyoung Lee, Dongduk Women's University, Korea
Chang Yong Lee, Kongju University, Korea
HooJin Lee, Hansung University, Korea
Doo Hun Jung, Korea Institute of Science and Technology Information, Korea
Young Yoon Cho, Sunchon University, Korea
Se-Hak Chun, SeoulTech, Korea
Sung Phil Choi, Korea Institute of Science and Technology Information, Korea
Ilkwon Cho, National Information Society Agency, Korea
Unho Choi, Seoul National University, Korea
Han Yong Choi, Hanbuk University, Korea

Contents

1 **SA-TCP: A Secure and Adaptive TCP for Wireless Ad Hoc Networks** 1
Waleed S. Alnumay, Uttam Ghosh and Pushpita Chatterjee

2 **IPAS: User Test Phase and Evaluation** 11
Sadiq Almuairfi, Prakash Veeraraghavan and Naveen Chilamkurti

3 **Coding DNA Sequence Watermarking** 21
Suk-Hwan Lee, Won-Joo Hwang, Eung-Joo Lee and Ki-Ryong Kwon

4 **Exclusive Region Based Scheduling for MB-OFDM UWB Using Hidden Pilot Sequences** 29
Byung Wook Kim and Byung-Gyu Kim

5 **The Digital Video Frame Obfuscation Method in Consideration of Human Visual System for Intellectual Property Protection and Authentication** 37
Heehyung Bu, Jonghwa Kim and Sungho Kim

6 **Fuzzy-Based Resource Reallocation Scheduling Model in Cloud Computing** 43
Jaekwon Kim, Taeyoung Kim, Minoh Park, Youngshin Han and Jongsik Lee

7 **An Efficient Algorithm for High Utility Sequential Pattern Mining** 49
Jun-Zhe Wang, Zong-Hua Yang and Jiun-Long Huang

8 **Simulated Annealing for the Design of Location Management System and Relation Map** 57
Si-Hwan Jang, Sung-Soo Kim and Sanghyun Joo

9 A Simple, Flexible and Adaptable Solution to Enable Information Exchange for Specific Business Processes 65
Mert Biçakçi and Mehmet Ulas Çakir

10 User Authentication with Distributed Cluster Formation in Wireless Sensor Networks. 75
Dong-Hoon Kim, Youn-Sik Hong and Ki Young Lee

11 Comparative Study of Cyber Security Characteristics for Nuclear Systems 87
Jinsoo Shin, Hanseong Son and Gyunyoung Heo

12 Optical Computing for Digital Signal Process in Incoherent Fiber System Using Positive Realization 95
Kyungsup Kim, Jaecheol Ryou and Woo-Tak Jung

13 Optimization of Non-fuzzy Neural Networks Based on Crisp Rules in Scatter Partition 103
Keon-Jun Park, Byun-Gon Kim, Kwan-Woong Kim, Jung-Won Choi and Yong-Kab Kim

14 LBG-Based Non-fuzzy Inference System for Nonlinear Process 111
Keon-Jun Park, Kwan-Woong Kim, Byun-Gon Kim, Jung-Won Choi and Yong-Kab Kim

15 An Efficient Embedded System for the Detection and Recognition of Speed-Limit Signs 119
Hsing-Lung Chen, Ming-Sze Chen and Shu-Hua Hu

16 Future Enterprise Computing Looking into 2020. 127
Yung-Chin Fang, Yuxiang Gao and Cindy Stap

17 Implementation of a Multi-core Prototyping System for a Video-Based Fire Detection Algorithm 135
Yong-Hun Park, Myeongsu Kang and Jong-Myon Kim

18 A Novel Spectrum Sensing Scheme Based on Wavelet Denoising with Fuzzy for Cognitive Radio Sensor Networks. 145
Hiep Vu-Van and Insoo Koo

19 Accelerating a Bellman–Ford Routing Algorithm Using GPU. . . . 153
In-Kyu Jeong, Jia Uddin, Myeongsu Kang, Cheol-Hong Kim and Jong-Myon Kim

20 Early Fire Detection Using Multi-Stage Pattern Recognition Techniques in Video Sequences 161
Dongkoo Shon, Myeongsu Kang, Junsang Seo and Jong-Myon Kim

21 Energy Efficient Backoff-Free Fragment Retransmission Fragmentation Scheme for IEEE 802.11 WLAN 169
Prosper Mafole, Yuki Manabe, Teruaki Kitasuka and Masayoshi Aritsugi

22 Fault Classification of an Induction Motor Using Texture Features of Vibration Signals 177
Won-Chul Jang, Myeongsu Kang and Jong-Myon Kim

23 An Efficient Audit Scheme for Cloud Storage 185
Ohmin Kwon, Dongyoung Koo, Yongjoo Shin and Hyunsoo Yoon

24 High-Performance Cluster Estimation Using Many-Core Models 193
Junsang Seo, Myeongsu Kang, Cheol-Hong Kim and Jong-Myon Kim

25 Dynamic Processing of Floating Continuous Query for Cold Chain Management 203
Bonghee Hong, Kyungmin Lee and Yongxu Piao

26 A New Approach for Task Scheduling Optimization in Mobile Cloud Computing 211
Pham Phuoc Hung, Tuan-Anh Bui and Eui-Nam Huh

27 Study on Encryption of Wireless Network-Based Mobility Energy-Enterprise Security Management 221
Young-Hyuk Kim, Il-Kown Lim, Jae-Pil Lee, Jae-Gwang Lee, Hyun Namgung and Jae-Kwang Lee

28 Secure and Efficient Deduplication over Encrypted Data with Dynamic Updates in Cloud Storage 229
Dongyoung Koo, Junbeom Hur and Hyunsoo Yoon

29 Tool Selection: A Cloud-Based Approach 237
Chirpreet Singh, Qun Shao, Yuqian Lu, Xun Xu and Xinfeng Ye

30 The Self-induction System for Fuel-Efficient Driving Using OBD II 247
WonAng Jang, JinGwan Park, JuWan Kweon, ChangYong Lee and DoHoon Lee

31 Intuitive Hand Motion Recognition and NUI/NUX Framework 253
Gwanghyung Lee, Dongil Shin and Dongkyoo Shin

32 Evaluating Skin Hereditary Traits Based on Daily Activities 261
Jehyeok Rew, Young-Hwan Choi, Daehoon Kim, Seungmin Rho and Eenjun Hwang

33 A New Fast Encoding Algorithm Based on Motion Activity for High Efficiency Video Coding (HEVC) 271
Jong-Hyeok Lee, Kalyan Goswami and Byung-Gyu Kim

34 Smart Network Services Using Ticket Issuance in the TSM (Trusted Software Monitor) System 279
Su-Wan Park and Jeong Nyeo Kim

35 Macro Modeling Approach for Semi-digital Smart Integrated Circuits 289
Zaniar Hoseini, Kye-Shin Lee and Byung-Gyu Kim

36 Hybrid Performance with Pixel Values' Transition and Curve Fitting for Improved Stereo Matching 297
Kwangmu Shin, Daekeun Kim and Kidong Chung

37 ActiveSTB: Scalable Scheme Over Wireless Networks 305
Heung Ki Lee, Kyung Jin Ahn, Jaehee Jung and Gangman Yi

38 Rhythm-Based Authentication Scheme on Smart Devices 313
Jae Dong Lee, Won Min Kang, Chang Hyun Moon, Hwa-Young Jeong, Young-Sik Jeong, Jai-Jin Jung and Jong Hyuk Park

39 Cloud Services Aided e-Tourism: In the Case of Low-Cost Airlines for Backpacking 321
Qingguo Zhou, Jason C. Hung, Jun Hu, Huaming Chen, Rui Zhou, Ji Qi, Lei Yang and Xuan Wang

40 Modified Cat Swarm Optimization Algorithm for Feature Selection of Support Vector Machines 329
Kuan-Cheng Lin, Yi-Hung Huang, Jason C. Hung and Yung-Tso Lin

41 Using Affective Computing to Detect Emotions with Color 337
Min-Feng Lee, Guey-Shya Chen and Jen-Chieh Wang

42 Bucket Index Ordering Problem in Range Queries 347
Taek-Young Youn, Nam-Su Jho, Ku-Young Chang and Dowon Hong

43 Signal Interference of Ubiquitous Wireless Networks on Data Throughput 357
Jie Zhang, Goo Yeon Lee, Hwa Jong Kim and Yong Lee

44 Privacy-Preserving Data Collection with Self-Awareness Protection 365
Kok-Seng Wong and Myung Ho Kim

45 User-Centric Security Service Model for Cloud Storage 373
SeongMin Yoo, PyungKoo Park, JinSeop Shin, HoYong Ryu and JaeCheol Ryou

46 Continuity Labeling Technique of Multiple Face in Multiple Frame 383
Hyun-jung Kim, Il-Yong Weon and Chang-hun Lee

47 Audio Classification Using GA-Based Fuzzy C-Means 393
Myeongsu Kang and Jong-Myon Kim

48 Anonymous Editing Prevention Method for Depth-Image-Based Rendering Stereoscopic Video 401
Ji-Won Lee, Hae-Yeoun Lee and Sang-Hyun Joo

49 Different Knowledge Modeling Approach for Clinical Decision Supporting Service in u-Health Context 409
Jeong Ah Kim

50 Formal Verification for Inter-Partitions Communication of RTOS Supporting IMA 415
Sachoun Park and Gihwon Kwon

51 3D Face Representation Using Inverse Compositional Image Alignment for Multimodal Face Recognition 423
Kanghun Jeong, Hyeonjoon Moon and Sanghoon Kim

52 Simulating Synthesized Automata from Decentralized Specifications in LTL 431
Ryoungkwo Kwon and Gihwon Kwon

53 Peers-Based Location of Mobile Devices 439
Mordechai Guri, Anatoly Krasner and Yuval Elovici

54 Platform Design for Data Sharing 447
Yun-Hee Son, Min-A Kang, Woo-Seung Jo, Kyu-Jin Choi and Kyu-Chul Lee

55 Reasonable Resolution of Fingerprint Wi-Fi Radio Map for Dense Map Interpolation 453
Wonsun Bong and Yong Cheol Kim

56 Sensor Based Ubiquitous Application Modeling Using Process Definition Language 461
Sang-Young Lee

57 Smart Device Locking System Provides Handy and Improved Security 467
Hyun-Woo Kim and Young-Sik Jeong

58 A Study on the Efficient Management of Android Apps Using a Whitelist 473
Uijin Jang, Changjin Suh and Hyungmin Lim

59 A Workflow Scheduling Technique to Consider Task Processing Rate in Spot Instance-Based Cloud 483
Daeyong Jung, JongBeom Lim and Heonchang Yu

60 Enhanced Beacon Scheduling of IEEE802.15.4e DSME 495
Sung-wook Nam and Kwang-il Hwang

61 Building Recognition Guided by Wi-Fi Based Localization 505
Wonsun Bong and Yong Cheol Kim

62 Context Aware Based Access Control Model in Cloud Data Center Environment 515
Yun Sang Byun and Jin Kwak

63 Whitelist Generation Technique for Industrial Firewall in SCADA Networks 525
Dong-Ho Kang, Byoung-Koo Kim, Jung-Chan Na and Kyoung-Sun Jhang

64 Design of Industrial Firewall System Based on Multiple Access Control Filter 535
Byoungkoo Kim, Dong-Ho Kang, Jung-Chan Na, Beom-Hwan Chang and Tai-Myoung Chung

65 The Improvement for Performance of Inter-VTS Data Exchange Format Protocol in VTS System 545
Namje Park

66 Abnormal Behavior Detection Technique Based on Big Data 553
Hyunjoo Kim, Ikkyun Kim and Tai-Myoung Chung

67 Rotational Bridge Priority Masking for Building Equal-Cost Multi-tree 565
Changjin Suh and Uijin Jang

68 Linear SVM-Based Android Malware Detection 575
Hyo-Sik Ham, Hwan-Hee Kim, Myung-Sup Kim and Mi-Jung Choi

69 A Study on Authentication Mechanism to Ensure the Reliability of Content Distribution Using Log Monitoring and Encoding of OPSs 587
Byoung-Soo Koh, SangHo Seo and ByungOk Jung

70 Virtual Agent Learning for NUI/NUX-Based Games 597
Daxing Jin, Seoungjae Cho, Yunsick Sung, Kyungeun Cho and Kyhyun Um

71 An Analysis of the Korea National DNS Using Big Data Technology 605
Euihyun Jung, Joonhyung Lim and Juyoung Kim

72 Development of Bio-Signal Based Continuous Intensity Wearable Input Device 615
Jinuk Kim, Jihye Kwon, Subin Im, Seongyoon Lee and Jaehyo Kim

73 Speaker Adaptive Real-Time Korean Single Vowel Recognition for an Animation Producing 633
Sun-Min Whang, Bok-Hee Song and Han-Kyung Yun

74 Social Network Service Real Time Data Analysis Process Research 643
Yu-Jong Jang and Jin Kwak

75 A Database Design for Pre-ordering System Based on Prediction of Customer Arrival Time 653
Aziz Nasridinov, Sangeun Yoon and Young-Ho Park

76 Behavior Execution Method for a Robot Using an Online Development Approach 661
Seungyoub Ssin, Seoungjae Cho, Kyhyun Um and Kyungeun Cho

77 Smart Logistics Service Model Based on Context Information . . . 669
KyungHun Kwak, NamJin Bae and YongYun Cho

78 Auto-configuration of the Initial Contour Using Level-Set Method for Lung Segmentation on Chest CT Images 677
Seung-Hoon Chae, Hae-Min Moon, Yongwha Chung and Sung Bum Pan

79 Suction Based Wall Climbing Robot for Edge Movement 687
Jeesoo Lee and Sanghoon Kim

80 Improving Data Hiding Capacity Based on Hamming Code 697
Cheonshik Kim and Ching-Nung Yang

81 Analysis of the Dynamic Characteristics of a Torque-Assist System Using the Model-Based Design Method 707
Kim Gyoung Eun and Kim Byeung Woo

82 Improved Branch Prediction for Just-in-Time Decompression of Canonical Huffman Bytecode Streams 719
Chisoon Jeong, Su Hang and Bernd Burgstaller

83 Effective Path Modification Using Intuitive Gestures 731
Hongzhe Liu, Seoungjae Cho, Yulong Xi, Kyungeun Cho and Kyhyun Um

84 DSRC Based Self-Adaptive Navigation System: Aiming Spread Out the Vehicles for Alleviating Traffic Congestion 739
Hohyeon Jeong, Jung-Won Lee, Jaehoon (Paul) Jeong and Eunseok Lee

85 A Two-Tier Routing Algorithm Based on Geographical Zone in VANET 751
Seung-Hwan Lee and Seung-Joon Seok

86 Analysis of Vehicle Surroundings Monitoring System Based on Vehicle-to-Vehicle Communication 763
Hong Cho and Byeong-Woo Kim

87 The Energy-Efficient Probabilistic Routing in MANETs 773
Sang-Hyun Park and Jung-Ryun Lee

88 Functional Annotation of Proteins by a Novel Method Using Weight and Feature Selection 785
Jaehee Jung, Heung Ki Lee and Gangman Yi

89 Gaze Behavior Analysis System Based on Objects Using Trajectory Information 799
Jung-Min Oh, Sangjin Hong and Nammee Moon

90 The Crack Width Calculation Program for Structural Concrete Members Based on a Limit State Design 809
HwaMin Lee, DooSoon Park and GiYeol Lee

91 Novel Real-Time Automobile Detection Algorithm for Blind Spot Area 821
Seung-Hun Yang, Gwang-Soo Hong, Beak Ryong and Byung-Gyu Kim

92 Environment Control Model of Intelligent Plant Factory Using Context Based of Sensing Data 833
NamJin Bae, KyungHun Kwak and YongYun Cho

93 Implementation of an Integrated Log Analysis System Through Statistics-Based Prediction Techniques 843
Kwangman Ko

94 The Comparison of Pricing Schemes for Cloud Services 853
Se-Hak Chun, Byong-Sam Choi, Young-Woong Ko and Soon-Hak Hwang

95 Platform Design for Co-creation 863
Soon Goo Hong, Hyung Rim Choi, Hyun Jong Kim, Chang-hyun Park and Hyun Mi Lee

96 Towards the Implementation of an Ubiquitous Car Park Availability Detection System 875
Ian K. T. Tan, Poo Kuan Hoong, Chee Ken Hong and Low Zhi Wen

97 Improvised Classification Model for Cloud Based Authentication Using Keystroke Dynamics 885
T. Senthil Kumar, Abhijit Suresh and Aadarsh Karumathil

Author Index 895

Subject Index 899

Chapter 1
SA-TCP: A Secure and Adaptive TCP for Wireless Ad Hoc Networks

Waleed S. Alnumay, Uttam Ghosh and Pushpita Chatterjee

Abstract In this paper, we propose a secure and adaptive transmission control protocol (SA-TCP) for wireless ad hoc networks that utilizes network layer information to detect various types of packet losses and adjusts the value of *congestion window* (*CWND*) dynamically according to the dynamic network conditions. SA-TCP operates normally to collect the samples of *CWND* and takes the mean from these samples to set for the future samples. It also adjusts the value of *congestion window limit* (*CWL*) according to the dynamic network environment. A low complexity identity-based public key cryptography has been integrated with SA-TCP to make it suitable in highly vulnerable wireless ad hoc networks. A secret session key is generated on-fly between source and destination nodes to secure the three-way handshaking process of SA-TCP. Simulation results are given to show that SA-TCP outperforms the popular *New Reno* and *ATCP* in terms of throughput in different network scenarios.

Keywords WANET · TCP · Congestion window · Security

W. S. Alnumay
Computer Science Department, King Saud University, Riyadh, Saudi Arabia
e-mail: wnumay@ksu.edu.sa

U. Ghosh (✉)
Deptartment of E&ECE, Indian Institute of Technology, Kharagpur 721302, India
e-mail: uttamg@iitkgp.ac.in

P. Chatterjee
SRM Research Institute, Bangalore, India
e-mail: pushpita.c@res.srmuniv.ac.in

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_1, © Springer Science+Business Media Dordrecht 2014

1.1 Introduction

Performance of transmission control protocol (TCP) for wireless ad hoc networks (WANETs) is an active area of research and it has received a lot of attention for its applicability in military operations, disaster rescue missions and other civilian applications. TCP performs well in wired environment where packet losses are mainly due to congestion in the network, but in case of WANETs, a marked degradation in the performance of TCP has been shown, where packet losses are due to channel errors or link failures between mobile nodes [1]. The main reason is the inability of TCP to distinguish the various causes of packet losses. The newer versions of TCP, such as Tahoe, Reno and New Reno [2] also exhibit similar properties [3].

A number of transport layer protocols have been proposed in the literature [1, 4–9] to enhance TCP's performance in mobile ad hoc environments. However, it may be noted that when a packet drop occurs, most of the protocols are unable to detect the actual reason of packet loss (congestion, link failure, lossy channels or high bit error rates). In WANET, due to dynamic nature of the network and variable number hops between source and destination, the fairness and efficiency get degraded. It is also seen that the existing protocols are not good enough in setting its parameters like congestion window, congestion window limit, round trip time and the retransmission timeout timers.

The existing transport layer protocols lack certain protection from dynamic internal and external malicious activities and hence relies on separate security mechanisms. Existing secure routing protocols though ensure the correctness of the route discovery, may not provide secure end-to-end data delivery at transport layer. The major security threats [10] associated with transport layer protocols in WANET are such as *SYN flooding*, *session hijacking* and *ACK storm*. Therefore, following are some important goals required to be met while designing a transport layer protocol for wireless ad hoc networks: (a) Maximizing the throughput per connection; (b) Providing both reliable and unreliable connections as per the requirements of application layer; (c) Providing adaptability to the dynamics of the network such as rapid change in topology and changes in the nature of wireless links; (d) Utilizing available bandwidth efficiently; (e) Handling resources such as battery power and buffer sizes efficiently in resource-constraint environment; (f) Proper utilization of information from the lower layers of the protocol stack to improve network throughput; (g) Maintaining end-to-end semantics; and (h) Considering the security aspects as regards WANETs to make itself complete and self-sufficient.

In this paper, we have modified New Reno protocol for wireless ad hoc networks so that it can take care of the above issues. We call this modified protocol as *Secure and Adaptive TCP* (SA-TCP) and present the proposed protocol in next section.

1.2 System Model and Key Distribution

Here, we assume SDRP [11] as the ad hoc routing protocol used in the network layer. We also assume that *Explicit Congestion Notification* (ECN) [12] is enabled and the public/private key pairs are distributed to each node prior to the network starting its operation. The public (PK_A)/private (SK_A) key pair of a node A are distributed using the following technique [13, 14]: Let us assume that Q_1 and Q_2 be the two groups of a prime order O and P_1 and P_2 be the generators of Q_1. A bilinear pairing is a map $e : Q_1 \times Q_1 \rightarrow Q_2$ which has the following properties:

- Bilinear: $e(xP_1, yP_2) = e(P_1, P_2)^{xy}, \forall x, y \in Z_O$;
- Non-degenerate: There exist $P_1, P_2 \in Q_1$ such that $e(P_1, P_2) \neq 1$;
- Computable: There exist an efficient algorithm to compute $e(P_1, P_2) \forall P_1, P_2 \in Q_1$.

The security parameters Q_1, Q_2, e, H_1, H_2, P_1, $P_{pub} = mP_1$ are publicly known by all the nodes in the network. Here, the network master key m is shared by a set of n nodes. Threshold cryptography technique is used where t nodes in a network is required to construct the master key m. H_1 and H_2 are the two cryptographic hash functions such that $H_1 : \{0, 1\}^* \rightarrow Q_1$ and $H_2 : \{0, 1\}^* \times Q_1 \rightarrow Z_O$. Node A with identifier $\mathrm{ID_A}$ (i.e., hardware address of A), will have a public key $PK_A = H_1(\mathrm{ID}_A)$ and a private key $SK_A = mPK_A$. In addition, shared key (K) between nodes A and B is generated in a non-interactive fashion [15]: $K_{AB} = e(SK_A, PK_B) = e(PK_A, SK_B)$. The above key distribution is necessary for authentication of nodes in the ad hoc network to nullify the attacks.

1.3 SA-TCP: Secure and Adaptive Transmission Control Protocol

In this section, we present the proposed modification to TCP New Reno, named SA-TCP, for wireless ad hoc networks. SA-TCP handles network congestion in three phases: *slow start*, *congestion avoidance* and *congestion detection*. In *slow start* and *congestion avoidance* phases, SA-TCP enhances the performance of TCP New Reno by adjusting the values of congestion window (*CWND*) and congestion window limit (*CWL*) according to dynamic network conditions. Therefore, SA-TCP minimizes the chance of congestion and MAC layer collisions in the network.

In congestion avoidance phase of SA-TCP, whenever there is a change in size of *CWND*, it is taken as a sample and this sample value of $(t - 1)$-th congestion window size has been denoted here as $X(t - 1)$. SA-TCP operates in normal mode (i.e., similar to original TCP New Reno) for collecting samples of earlier *CWND* sizes $X(t - i)$ (where $i = 1, 2, 3, \ldots L$). It calculates the value of current congestion window $X(t)$ from these L samples using the following Eq. (1.1):

$$X(t) = \frac{1}{L}\sum_{i=1}^{L} X(t-i) \tag{1.1}$$

SA-TCP also sets the sizes of future *CWND* sizes in case the network condition remains unchanged. The proposed algorithm collects such L samples while working in the normal mode and creates the future congestion window sizes. Here, the value of L is an important factor to predict the current congestion window size. Using a smaller value of L may give inaccurate prediction of the network condition, whereas a larger value of L may make SA-TCP behave like TCP New Reno. We use a least-square method [16, 17] to compute the value of L as follows:

$$S_R = \sum_{i=1}^{L} (CWL - CWND_i)^2 \tag{1.2}$$

Here $CWND_i <= CWL$, and $CWND_i$ is the i-th sample of the congestion window that is predicted. We select the value of L corresponding to the minimum sum of residuals S_R. These ensure that the value of congestion window *CWND* will not exceed the value of congestion window limit *CWL* [17].

SA-TCP also reduces the chance of MAC layer collisions by restricting the value of the congestion window according to the equation [18] $BDP_{Max/Min} = N/k$, wherein $BDP_{Max/Min}$ is the upper/lower bound of bandwidth delay product, $(1/8) \leq k \leq (1/4)$ and N is the number of round-trip hop-count from source to destination nodes on the path. SA-TCP source node obtains the value of N through *RREP* message of SDRP [11] routing protocol.

In *congestion detection* phase, SA-TCP detects the congestion in the network using the *explicit congestion notification* (ECN) [12] bit. When a node on an active route detects the congestion, it sends an ACK segment to the source node with ECN bit enabled. On receiving the ACK segment, the source node reduces both sizes of present *CWL* and *CWND* to half its size and sets this reduced *CWND* size as the slow start threshold (*ssthreshold*) value. Thereafter it sets the value of *CWND* size to 1 and enters into the slow start phase [17].

Apart from segment loss due to congestion, segment may be lost due to either lossy channel or a link failure between the source and destination nodes. In order to detect a link failure on an active route, SA-TCP uses the *route error* (RERR) message of SDRP [11] routing protocol. Here, it uses the concept of P-TCP [17] while setting up its parameters after the route is re-established. The source node stops transmitting segments to the destination once it detects a link failure and waits till the route is re-established. When the route is re-established, the source node obtains the value of round-trip hop-count (N) from *route reply* (RREP) message of [11] routing protocol and calculates the value of congestion window limit *CWL* for the new route. Further, if the size of congestion window *CWND* is higher than the size of *CWL*, SA-TCP source node reduces the size of *CWND* to half. Otherwise the source node keeps the value of *CWND* same as was at the time of link failure. At this point, SA-TCP enters back to its original phase of

congestion avoidance or *slow start* as it was at the time of link failure. The above condition is necessary to ensure that the size of *CWND* never exceeds the size of *CWL* even when the route is re-established. Hence SA-TCP minimizes the chance of congestion and collisions in the network.

1.3.1 Security

This section describes a mechanism to secure the three-way handshaking connection establishment and connection termination processes of SA-TCP. Here, source (S) and destination (D) nodes of the end-to-end connection to be established first exchange their IDs (ID_S and ID_D) and random numbers (r_S by source node and r_D by destination node) to each other in signed (σ_S) $RREQ_S$ and signed (σ_D) $RREP_D$ messages respectively. Each intermediate node I on the path also generates a random number r_I and sends it to source and destination nodes in signed (σ_I) $RREQ_I$ and $RREP_D$ messages respectively. The signature scheme proposed in [13] is used here for the purpose of authentication. Thereafter node S generates a session secret key $K_{SD} = e(r_S SK_S, R_S PK_D)$, wherein SK_S is the private key of node S, PK_D is the public key of node D and $R_S = \sum_{i=1}^{J} r_i$ (here J is the number of intermediate nodes between source and destination nodes). Subsequently S starts three-way handshaking connection establishment process with D similar way to the scheme given in [19]. It first generates an initial sequence number (ISN_S) from a monotonically increased random number (M_S) and a hash value of source port address ($PORT_S$), destination port address ($PORT_D$), ID_S, ID_D and K_{SD}. S also computes an authentication tag (δ_S) on $\mathrm{SYN}(\mathrm{ISN_S})$ segment using a HMAC function and a session secret key K_{SD}. This $\mathrm{SYN}(\mathrm{ISN_S})$ segment is sent along with the tag δ_S by S to D.

On receiving the $\mathrm{SYN}(\mathrm{ISN_S}) + \delta_S$ from S, D computes the session secret key $K_{DS} = e(r_D SK_D, R_D PK_S)$, wherein SK_D is the private key of D, PK_S is the public key of S and $R_D = \sum_{i=1}^{J} r_i$. It also generates the authentication tag (δ_G) on received $\mathrm{SYN}(\mathrm{ISN_S})$ segment using the hash function and K_{DS}. If the generated tag (δ_G) and the received tag (δ_S) are same, then D concludes that S is authenticated. On successful authentication, D generates ISN_D and an authentication tag (δ_D) on $\mathrm{SYN}(\mathrm{ISN_D}) + \mathrm{ACK}(\mathrm{ISN_S} + 1)$ segment in the similar way to S. This $\mathrm{SYN}(\mathrm{ISN_D}) + \mathrm{ACK}(\mathrm{ISN_S} + 1)$ segment is sent along with δ_D by D to S. On receiving $\mathrm{SYN}(\mathrm{ISN_D}) + \mathrm{ACK}(\mathrm{ISN_S} + 1) + \delta_D$, S computes the authentication tag (δ_G) on received $\mathrm{SYN}(\mathrm{ISN_D}) + \mathrm{ACK}(\mathrm{ISN_S} + 1)$ segment using the session secret key K_{SD}. If both of the tags (generated δ_G and received tag δ_D) are same then S concludes that D is authenticated. S also sends $\mathrm{ACK}(\mathrm{ISN_D} + 1) + \delta_S$ segment to D. D generates the tag (δ_G) on received $\mathrm{ACK}(\mathrm{ISN_D} + 1)$ segment and verifies it with the received tag (δ_S). If both tags are same, D finishes the three-way handshaking connection establishment process and also allocates resource for S and starts sending of data messages along with authentication tags. SA-TCP secures

the three-way handshake connection termination process in the same way as connection establishment process. It may be noted here that an intermediate node sends ECN in the ACK segments and authentication tags whenever the node detects congestion.

Analysis The proposed SA-TCP sends all the segments along with authentication tag and ID of the sender. Here, an ID of a network node can be spoofed by an attacker but it may not be able to generate the valid authentication tags. This is due to the fact that it is difficult for the attacker to know the private keys and the master key of a network. The master key is shared between n nodes in the network where it may only get disclosed if $t + 1$ nodes gets compromised. Hence the chances of SYN flooding and segment forging attacks are very low in SA-TCP. The proposed protocol generates the initial sequence number (*ISN*) from a monotonically increasing random number (R) and a hash value of source port ($PORT_S$), destination port ($PORT_D$), source ID $\mathrm{ID_S}$, destination ID $\mathrm{ID_D}$ and a session secret key. Therefore, it is also difficult for the attacker to guess the *ISN* and hijack a session or creating ACK storm in the network.

It can also be seen that SA-TCP does not require any certificate or large sized public key for authentication. It uses identity-based cryptography where the public key of a node is generated from its ID (i.e., 48-bit hardware address). Therefore, SA-TCP only requires to send the ID with the segments thereby eliminating the need of sending the public key. Thus, we find that the proposed protocol has low overhead. However, the session secret key has to generate for each session, which increases the protocol overhead slightly. The following equation shows that both source–destination pairs generate the same session secret key:

$$\begin{aligned} K_{SD} &= e(r_S SK_S, R_S PK_D) = e(SK_S, PK_D)^{r_S R_S} = e(sPK_S, PK_D)^{r_S R_S} \\ &= e(PK_S, PK_D)^{s r_S R_S} = e(R_D PK_S, r_D sPK_D) = e(R_D PK_S, r_D SK_D) \\ &= K_{DS} \end{aligned}$$

Where $R_S = \sum_{i=1}^{J} r_i = R_D$.

1.4 Simulation

In order to compare the performances, we have simulated the proposed SA-TCP along with TCP New Reno [2] and Ad hoc TCP (ATCP) [8] using NS-2 (version-2.34) simulator on a computer running Linux Cent OS 5. *Throughput* is chosen as the performance comparison metric. We have considered two different simulation scenarios as follows:

Scenario 1—*Dynamic Random Topology:* In this scenario, we adopt *random way-point* mobility model where we vary the speed of the nodes between 0 and 25 m/s. The pause time is set to 0 s for acquiring continuous motion of the nodes

in the simulations. This scenario is simulated for 50 number of mobile nodes in a network area of size 670 m × 670 m.

Scenario 2—*String Topology:* We vary the hop distance between source and destination to show the effect in a string topology for the protocols under consideration. This scenario is simulated for 11 number of nodes in a static network of area size 1,500 m × 1,500 m.

The following simulation parameters are kept common for the above mentioned scenarios. The radio transmission range of each node is set to 250 m. *Two-ray ground* propagation model is used and IEEE 802.11 standard has been used as the Medium Access Control layer protocol. *Constant bit rate* (CBR) traffic generator is used of packet size 512 bytes with 4 packets/s traffic generation rate. We have set 1,000 s as the total simulation time for each set of simulations.

1.4.1 Simulation Results

Scenario 1—*Dynamic Random Topology:* The impact of node mobility on throughput is shown in Fig. 1.1a. From the figure we see that the throughput decreases with increased node mobility for all the protocols under consideration. This is due to the increase number of link failures with increased mobility of nodes resulting in segment losses. It can also be seen that the proposed SA-TCP gives better throughput compared to New Reno and ATCP. This is due to the fact that New Reno cannot detect link failures and thereby retransmits duplicate segments unnecessarily. It also mistakes these link failures as congestions in the network and enters into slow start phase. Even though ATCP detects the link failure, the protocol sets *CWND* to 1 and gradually increments it to the normal size. On the other hand, on detecting a link failure, SA-TCP enters back into the original phase (either congestion avoidance or slow start) as it was at the time of link failure and computes *CWND* according to the new route condition.

Scenario 2—*String Topology:* Fig. 1.1b shows the effect of the number of hop distances from source to destination nodes on the network throughput. It can be seen that the throughput decreases with the increased number of hop distances for New Reno, ATCP and the proposed SA-TCP. This is due to the number of MAC layer collisions keeps on increasing with the number of hops. SA-TCP performs better because it sets the mean value of *CWND* periodically. Further, it tries to reduce the collisions by adjusting the value of *CWND* according to the network condition and limits the congestion window below a certain value. On the other hand, as no precaution is taken to limit the congestion window by New Reno and ATCP, unnecessary collisions takes place in the network, which in turn triggers retransmissions leading to a domino effect [18].

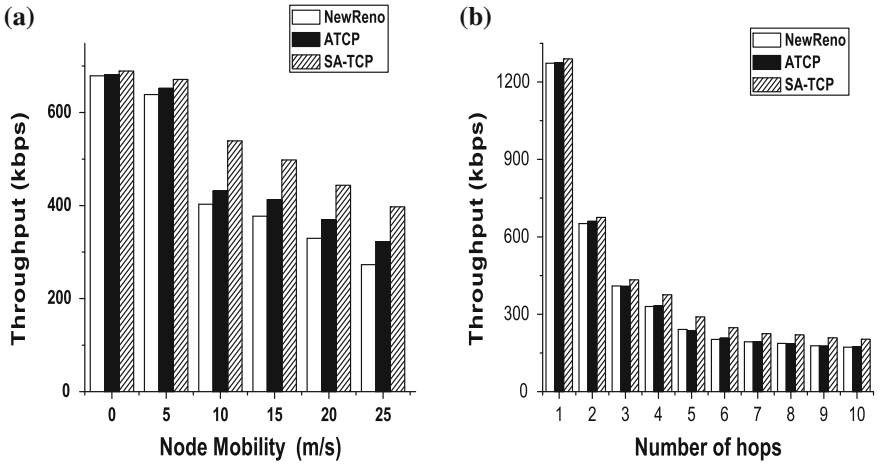


Fig. 1.1 **a** Throughput versus node mobility in dynamic random topology; **b** Throughput versus hop distance between source and destination in string topology

1.5 Conclusion

In this paper, we have proposed a secure and adaptive TCP, called SA-TCP for wireless ad hoc networks that uses network layer information for detecting different types of packet losses. It adjusts the values of *congestion window* (*CWND*) and *congestion window limit* (*CWL*) from their previous state values so that the chances of congestion and collisions in the network are minimized. A low complexity identity-based public key cryptography has been incorporated with the proposed protocol where the public key of a node is evaluated from its ID and TCP segments are sent along with *MAC* tags for authentication. Simulation results show that SA-TCP is efficient in terms of throughput compared to New Reno and ATCP for wireless ad hoc networks.

References

1. Holland G, Vaidya N (1999) Analysis of TCP performance over mobile ad hoc networks. In: Proceedings of the 5th annual ACM/IEEE international conference on mobile computing and networking. ACM, USA, pp 219–230
2. Hoe JC (1996) Improving the start-up behavior of a congestion control scheme for TCP. SIGCOMM Comput Commun Rev 26:270–280
3. Xu S, Saadawi T (2002) Performance evaluation of TCP algorithms in multi-hop wireless packet networks. J Wirel Commun Mob Compu 2:85–100
4. Chandran K, Ragbunathan S, Venkatesan S, Prakash R (1998) A feedback based scheme for improving TCP performance in ad-hoc wireless networks. In: Proceedings of 18th international conference on distributed computing systems, pp 472–479, May 1998

5. Sundaresan K, Anantharaman V, Hsieh H-Y, Sivakumar R (2005) ATP: a reliable transport protocol for ad hoc networks. IEEE Trans Mob Comput 4(6):588–603
6. Singh H, Saxena S, Singh S (2004) Energy consumption of TCP in ad hoc networks. Wirel Netw 10:531–542
7. Kopparty S, Krishnamurthy SV, Faloutsos M, Tripathi SK (2002) Split TCP for mobile ad hoc networks. In: Proceedings of the IEEE global communications conference (GLOBECOM 2002), pp 138–142
8. Liu J, Singh S (2001) ATCP: TCP for mobile ad hoc networks. IEEE JSAC 19:1300–1315
9. Kim D, Toh C-K, Choi Y (2000) TCP-bus: improving TCP performance in wireless ad hoc networks. In: Proceedings of IEEE international conference on communications (ICC), vol 3, pp 1707–1713
10. Wu B, Chen J, Wu J, Cardei M (2007) A survey of attacks and countermeasures in mobile ad hoc networks. In: Xiao Y, Shen XS, Du D-Z (eds) Wireless network security. Springer, US
11. Ghosh U, Datta R (2013) SDRP: A secure and dynamic routing protocol for mobile ad hoc networks. IET Networks
12. Ramakrishnan KK, Floyd S, Black D, Ramakrishnan GK (2001) The addition of explicit congestion notification (ECN) to IP
13. Ghosh U, Datta R (2012) A novel signature scheme to secure distributed dynamic address configuration protocol in mobile ad hoc networks. In: IEEE WCNC, pp 2700–2705, Apr 2012
14. Ghosh U, Datta R (2011) A secure dynamic ip configuration scheme for mobile ad hoc networks. Ad Hoc Netw 9(7):1327–1342
15. Sakai R, Ohgishi K, Kasahara M (2000) Cryptosystems based on pairings. In: The 2000 symposium on cryptography and information security, pp 26–28
16. Stoica P, Friedlander B, Soderstrom T (1986) Least-squares, yule-walker, and overdetermined yule-walker estimation of ar parameters: a monte carlo analysis is of finite-sample properties. Int J Control 43(1):13–27
17. Ghosh U, Datta R (2013) P-TCP: a prediction based secure transmission control protocol for wireless ad hoc networks. IETE J Res 59:364–375
18. Chen K, Xue Y, Nahrstedt K (2003) On setting TCPN++s congestion window limit in mobile ad hoc networks. In: Proceedings of IEEE international conference on communications (ICC n++03), Alaska, May 2003
19. Ghosh U, Datta R (2011) Identity based secure AODV and TCP for mobile ad hoc networks. In: Proceedings of ACM ACWR, pp 339–346

Chapter 2
IPAS: User Test Phase and Evaluation

Sadiq Almuairfi, Prakash Veeraraghavan and Naveen Chilamkurti

Abstract User authentication is one of the most important topics in information security. A text-based strong password scheme can provide a certain degree of security. However, as strong passwords are difficult to memorize, users often write them down on a piece of paper or even save them in a computer file. An image-based authentication scheme has been proposed as a possible alternative solution to text-based authentication, motivated particularly by the fact that humans can remember images better than text. Recently, many networks, computer systems and electronic-commerce environments have tried using a graphical technique for user authentication. All graphical and image algorithms comprise two different aspects: usability and security. Unfortunately, none of the graphical algorithms are able to cover both these aspects at the same time. In this paper, we evaluate the usability and security of different authentication schemes and compare them with our proposed scheme, which is the Implicit Password Authentication System (IPAS) by an experiment and a questionnaire survey.

Keywords Authentication · IPAS · Graphical password · Security · Usability

S. Almuairfi (✉) · P. Veeraraghavan · N. Chilamkurti
Department of Computer Science and Computer Engineering, La Trobe University, Melbourne 3086, Australia
e-mail: sadiqjafar@students.latrobe.edu.au

P. Veeraraghavan
e-mail: p.veera@latrobe.edu.au

N. Chilamkurti
e-mail: n.chilamkurti@latrobe.edu.au

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_2, © Springer Science+Business Media Dordrecht 2014

2.1 Introduction

In recent years, computer and network security have been seen as extremely important issues. A key factor in security research is authentication which is the determination of whether a user should be allowed access to a given system or resource. In this respect, the password is a common and widely authentication method still used currently.

Although traditional alphanumeric passwords are used widely, they have problems such as being hard to remember, vulnerable to guessing, dictionary attacks, key-loggers, shoulder-surfing and social engineering [1]. In addition to these types of attacks, a user may tend to choose a weak password or record his password. This may further weaken the authentication schemes. As an alternative to the traditional password-based scheme, the biometric system was introduced. This relies upon unique features unchanged during the life time of a human, such as finger prints, iris etc. The major problem of biometric as an authentication scheme is the high cost of additional devices needed for identification process [2]. The false-positive and false-negative rate may also be high if the devices are not robust. Biometric systems are vulnerable to replay attack (by the use of a sticky residue left by a finger on the device), which reduces the security and usability levels. Thus, recent developments have attempted to overcome biometric shortcomings by introducing token-based authentication schemes which relies on the use of a physical device such as smartcards or electronic-key.

Graphical-based password techniques have been proposed as a potential alternative to text-based techniques, supported partially by the fact that humans can remember images better than text [3]. Therefore, graphical-based authentication schemes have higher usability than other authentication techniques. It is also difficult to break graphical passwords using normal attacks such as dictionary attack, brute force and spyware which have affected text-based and token-based authentication [4]. Thus, the security level of graphical-based authentication schemes is higher than other authentication techniques as we proved in our experiment.

In this paper, we start with an overview of current authentication schemes in Sect. 2.2. Then, Sects. 2.3 and 2.4 describe the IPAS user test phase which is the core of this paper. Sections 2.5 and 2.6 explain IPAS usability and security levels from the user's perspective and provides user feedback. Finally, we present the conclusion and future directions in Sect. 2.7.

2.2 Related Work

In general, graphical password techniques can be classified into two: recognition-based and recall-based graphical techniques [5].

2.2.1 Recognition-Based Systems

In recognition-based systems, a group of images is displayed to the user and accepted authentication requires a correct image being clicked or touched in a particular order. Some examples of recognition-based system are Awase-E system, AuthentiGraph, and Passfaces system, one may refer to [2] for more details.

Although a recognition-based graphical password seems to be easy to remember, which increases the usability, it is not completely secure. It needs several rounds of image recognition for authentication to provide a reasonably large password space, which is tedious [6]. Also, it is obvious that recognition-based systems are vulnerable to replay attack and mouse tracking because of the use of a fixed image as a password. Thus, we consider these drawbacks in our IPAS system, which overcomes the problems of recall-based schemes too.

Pure Recall-Based Technique. With these types of systems, users need to reproduce their passwords without any help or reminder by the system. Draw-A-Secret technique and Passdoodle system are common examples of pure recall-based techniques. We have evaluated and discussed pure recall-based systems in our paper [2].

Cued-Recall Based Technique. With this technique, the system provides hints which help users to reproduce their passwords with high accuracy. These hints are presented as hot spots (regions) within an image. The user has to choose some of these regions to register their password and they have to choose the same region following the same order to log into the system. The user must remember the "chosen click spots" and keep them secret. There are many implementations, such as the Blonder algorithm and the PassPoint scheme which we covered in [1, 2].

2.3 IPAS Overview and its Advantages

IPAS is an intelligent imaged-based authentication framework that falls under the "What you know" type of authentication. In a traditional "what you know" type of authentication, either the password phrase or the image (in the case of an image-based password) is static and the user is expected to return the exact phrase (in the case of a textual password) or "click" on the exact location in an image (in the case of an image-based password) to the server for authentication. Since the user is expected to provide the same static information during every authentication cycle, it has a number of security pitfalls, including replay attack and shoulder surfing. IPAS is the most generalized authentication scheme that eliminates all the security pitfalls.

IPAS creates an intelligent user personalization profile for every individual user during the registration stage. The profile is based on the information (like pastime

hobbies, likes/dislikes, favourite movie stars etc.) provided by the user at the time of registration. The authentication server may have several images and each image may have more than one clickable area. Each clickable area may represent an object and have several text attributes associated with it. During the authentication process, the system presents to the use one or more pieces of information in an implicit image form. The user is expected to "click" on the correct area the represents the "expected information" in an implicit form to complete the authentication process. For example, if the user likes apples, the system may present an image of "Sir Isaac Newton" or a Macintosh computer. Since different information is expected from the user every time they log in through different images implicitly representing the information, IPAS is immune to the problems suffered by the "What you know" type of authentication schemes. Since IPAS may be configured to present to the user static information every time they log in (including static textual images), it may also act like other static authentication schemes.

2.3.1 IPAS Domains

Registered IPAS users can be categorized, in relation to their interaction behaviour, into two domains: region/state domain and word space/distance domain.

Various kinds of users in a system can, at times, have a different perspective of their needs. For instance, the perception of someone in America may not be the same as someone who is in Australia. Therefore, the system will consider the area/region where the user is during the creation of the user's password and will link their information to the region domain. As a result, the objects in an authentication image will be related to the user's region. For example, if the keyword or the information is opera, then users from Australia may easily recognize the Sydney Opera House; whereas users from Europe may more easily recognize the London Coliseum Theatre than the Sydney Opera house (some may not even recognize it at all). Thus, the system under this domain will be country/region dependent which should be considered during the production roll-out of the module.

In the space domain, we measure implicit level between word and object in an image. For example, let the chosen keyword be *apple*. An image which contains an *apple pie* will represent a closer relation to the word *apple*, while an image of *Steve Jobs* will represent a more distant relation to the word *apple*. The semantic variants of this concept can be developed quickly and with little effort by the system developer. We test these domains during our IPAS experiment and survey as below.

2.4 IPAS User Test Phase

2.4.1 Methodology

In our experiment, we tested IPAS using a number of participants from different backgrounds and with different experience in using authentication techniques. Participants include IT students, nurses, bankers, and miscellaneous workers. All participants engaged in the IPAS experiment and then answered our survey questionnaire.

2.4.2 IPAS Experiment

In our experiment, the participants were asked to read a given story/keywords, however, in the real-life implementation of IPAS, this would be provided by the user. The participants were then asked to match the keywords (password) with the object(s) in an image. The objects in the image implicitly represent one or more keywords. Three different stories and images were distributed randomly to the participants. We fetch the keywords of one story which was used in the IPAS experiment as follows:

Keywords selected in advance are: "*Unkempt Hair, Skyscrapers, Homeless, Dollar, Eagle, and Rat*".

Training Phase. At the beginning of the experiment, we informed participants of its main purpose and what they would do before, during, and after the experiment. Then, we explained the definitions of the different authentication schemes: text, biometric, token, graphical and image-based password schemes with simple examples and figures. After this, an overview of IPAS, the experiment steps, and the survey questions were discussed. We also showed them several images and the way to link keywords with the object in the image.

We clarified the difference between IPAS and PassPoint as follows: PassPoint uses one image with different click points (passwords) for all users during the authentication process, while IPAS uses a new image with different objects to represent the password every time. The IPAS image is changed with every authentication process to avoid replay attacks or shoulder surfing. Also, IPAS is a one-to-many relation which means one word (password) is represented by many implicit objects within the image while the Question Hint scheme is a one-to-one relation between the password and its question. For example, the word *apple* in IPAS could be represented as apple pie, apple juice, Steve Jobs etc. while in the Question Hint scheme; the word *apple* will be the direct answer for a question "What is your favourite fruit?". In this stage of the IPAS experiment, we answered any questions and concerns which were raised by the participants.

Experiment Phase. To begin with, we asked the participants about the circled objects in the given image as in Fig. 2.1. The most common answers were direct or

Fig. 2.1 IPAS sample image with implicit objects

explicit words *comb* and *drink bottle*. After this, we asked the user to read the above story and keywords and then to implicitly match the keywords and objects with the given image by marking the correct object. Most users were able to implicitly link the keywords *unkempt hair* and *skyscrapers* with the correct objects within the image which means that they were able to recall the password correctly. After this, the users were asked to answer the survey questions which compared IPAS with other well-known authentication schemes from a usability and security perspective.

The motivation for this survey is multifold. First, we analyze the authentication schemes based on usability and security parameters. Then, we compare IPAS with current authentication schemes in relation to usability and security issues.

2.4.3 Survey Questions

In this section, we give an overview of our survey questions. In the section on demographic information, we ask two questions: one about age and the other about gender. Then, the survey is divided into two main categories: general information about authentication schemes and IPAS evaluation after performing the experiment. The last part of the survey asks users to make comments and suggestions.

The category on general information about authentication schemes included the three following questions: Question# 1 "How frequently do you deal with the below authentication schemes? Tick one box for each scheme". Response options for this question were "Always", "Usually", "Often", "Rarely", and "Never". The second question was "How do you evaluate the below authentication schemes from an ease-of-use perspective?" Response options for this question were a rating from 1 to 5 where 1 is easy and 5 is difficult. The third question was "How do you evaluate the below authentication schemes from a security perspective?" Response

options for this question were a rating from 1 to 5 where 1 is not secure at all and 5 is strongly secure.

The IPAS evaluation category includes six questions which cover the main and minor factors as follows:

(1) How do you evaluate IPAS from an easy-to-use perspective? (2) How do you evaluate IPAS from a security perspective? (3) What do you think about the story in the IPAS experiment? (4) Do you prefer a close or a distant relation between the key word and object in the image? (5) What kinds of stories do you prefer? (6) Do you like stories in your native language? (This question is for non-English speakers.)

The response options for questions 1 and 2 were similar to those for questions 1 and 2 in category one. For the third question about the IPAS story, the participants were required to respond with a rating from 1 to 5 where 1 is too short and 5 is too long to indicate their opinion about the length of the IPAS story. In question 4, the participants were asked about the relationship between keywords and the object in the image. Response options for this question were a rating from 1 to 5 where 1 indicates a direct/explicit relationship and 5 is an indirect/implicit relationship. To enhance IPAS domain objectives and to evaluate the effects of the minor factors, we asked the participants in question 5 about the type of story they preferred. The responses for this question were Regional, International, or Both. The last question in this category was a specific question to evaluate the language in the IPAS story, especially for non-English speakers. The responses for this question were either yes or no.

At the end of the survey, we included an open-ended question for participants' comments and suggestions about the IPAS experiment and the survey on the authentication schemes.

2.4.4 Survey Analysis

The responses to the demographic information showed that 29 % of the participants were female and 71 % were male. In relation to the age of the participants, 54 % were aged from 18 to 30 years, 45 % were aged from 31 to 50 and 1 % were more than 50 years old. The authentication schemes which were included in our survey are: the test-based authentication scheme, the biometric-based authentication scheme, the token-based authentication scheme, the graphical-based authentication scheme, and the image-based authentication based.

Our first question asked how frequently the participants used the authentication schemes, with five response options being always, usually, often, rarely, and never. In response to the question as to how frequently the participants used the authentication schemes, the survey respondents answered 75 % always, 14 % usually, 6 % often, and 2 % for each rarely and never, respectively.

The participants indicated that the second most popular authentication scheme was the token-based scheme with 22 % responding always, 15 % usually, 14 %

often, 16 % rarely, and 33 % never. The biometric authentication scheme was the third most popular, with 24 % or respondents reporting (always and usually) above average, 16 % average (Often), and 60 % below average. Finally, graphical-based passwords and image-based password were the least popular of the authentication schemes with only 7 % of respondents indicating they had dealt with image-based passwords and 86 % indicating they had either rarely or never used an image-based authentication scheme.

The results indicate that a text-based password is still the most common form of authentication preferred by most participants. Conversely, an image-based password is almost unknown to most of the participants, therefore there needs to be greater awareness of and research into this type of authentication scheme.

The next questions were designed to evaluate the usability and security of the current authentication schemes in order to compare them with IPAS, which is the aim of our survey.

2.5 Usability of IPAS

Usability refers to the degree to which a system is easy to learn, use, and meets the user's needs. Even though many authentication schemes have been proposed to improve password strength, the responses to the usability of these systems indicate that yet another improved scheme needs to be introduced. Therefore, in our survey, we use a scale to study user's feedback on the different authentication schemes.

The scale in this area was divided into five stages: easy, a little easy, acceptable, a little difficult, and difficult. In this question, we include IPAS and other authentication schemes to measure which is considered to be an easy method of authentication by participants.

The five abovementioned stages were revised into three stages as follows:

Acceptable = Average, Easy, a little easy = Above Average, Difficult, a little difficult = Below average.

As a result, a total of 81 % (63 % easy and 18 % a little easy) of participants said that it was easy to use a text-based password, 13 % indicated it was acceptable (average), and only 6 % considered a text-based password scheme difficult to use as a way of authentication. Usability of biometric, token-based, and graphical-based authentication schemes are 46, 48, and 38 % respectively. Thus, most users consider them as difficult to be used.

The responses to image-based password schemes were as follows: 30 % indicated that they are easy to use, 21 % indicated acceptable, and 49 % considered them difficult to use as an authentication scheme. In general, image-based passwords were considered difficult to use by the majority of participants.

In relation to IPAS, most participants (54 %) indicated it was easy to use, 25 % indicated it was acceptable, and 21 % indicated it was difficult to use as an authentication scheme. Therefore, according to the participants' feedback, IPAS is easier to use than other image-based authentication schemes.

The majority of participants (81 %) indicated that text-based password schemes were the most usable, IPAS was the second most usable authentication scheme (54 %), the token-based scheme was the third most usable at (48 %), followed by the biometric (46 %), graphical (38 %) and image-based schemes, (30 %) respectively.

2.6 Security of IPAS

In this question, we analyze participants' opinions regarding the security level of different authentication schemes. The response options to this question were "Not Secure at all", "Less Secure", "Acceptable", "Secure", and "Strongly Secure" for each scheme. Also, we included IPAS with other authentication schemes to measure which authentication scheme was considered the most secure scheme.

The five abovementioned stages were revised into the three stages as follows: *Acceptable = Average, Secure and Strongly Secure = Above Average, Not secure at all & less secure = below average.*

We selected above average (secure and strongly secure) values to evaluate the security level of the given schemes. Most of the participants (65 %) considered biometric-based authentication scheme as the most secure scheme, with the second most secure authentication scheme being IPAS (58 %), followed by the image-based password (55 %), token-based (53 %), and graphical-based (52 %). Only 27 % of the participants thought that the text-based password scheme was a secure scheme.

The responses showed that overall, the participants felt that IPAS was the second most secure of the authentication schemes. Furthermore, they indicated that the image-based password scheme had an acceptable level of security which implies that password entropy in image-based password schemes is longer than text-based password schemes. Therefore, IPAS achieved the required balance of usability and security to be used as a primary authentication scheme.

2.7 Conclusion and Future Directions

In this paper, we compared IPAS with other authentication schemes by performing two experiments and asking participants to answer a questionnaire. We then explained the usability and security of IPAS from the users' point of view. In our subsequent papers, we will try to present the dynamic IPAS as a new version of IPAS with more features to obtain a better balance between usability and security.

References

1. Almuairfi S, Veeraraghavan P, Chilamkurti N (2011) IPAS: implicit password authentication system. In: Advanced information networking and applications (WAINA), 2011 IEEE workshops of international conference on advanced information networking and applications
2. Almuairfi S, Veeraraghavan P, Chilamkurti N (2013) A novel image-based implicit password authentication system (IPAS) for mobile and non-mobile devices. Math Comput Model 58(1–2):108–116. ISSN 0895-7177, 10.1016/j.mcm.2012.07.005
3. Xiaoyuan S, Ying Z et al (2005) Graphical passwords: a survey. In: 21st annual computer security applications conference, pp 463–472
4. Wells J, Hutchinson D, Pierce J (2008) Enhanced security for preventing man-in-the-middle attacks in authentication, data entry and transaction verification. In: Australian information security management conference. Paper 58
5. Xiaoyuan S, Ying Z et al (2005) Graphical passwords: a survey. In: Computer security applications conference, 21st annual
6. Pierce JD, Wells JG, Warren MJ, Mackay DR (2003) A conceptual model for graphical authentication. In: 1st Australian information security management conference, 24 Sept. Perth, Western Australia, paper 16

Chapter 3
Coding DNA Sequence Watermarking

Suk-Hwan Lee, Won-Joo Hwang, Eung-Joo Lee and Ki-Ryong Kwon

Abstract This paper presents a coding sequence watermarking on lifting based DWT domain and brings up the availability of frequency domain watermarking for DNA sequence. The proposed scheme allocates codons to numerical codes by the permutation table of histogram rank of amino acid and substitutes a sub-sequence of codons to a sequence that has DWT coefficients that is optimum to a watermark bit, among a set of DWT coefficients with the same amino acids. From experimental results, we verified that the proposed scheme has the robustness to until a combination of 10 % point mutations, 5 % insertion and deletion mutations and also the amino acid preservation and the security.

Keywords DNA Watermarking · Coding DNA sequence · DNA copyright protection

S.-H. Lee (✉)
Department of Information Security, Tongmyong University, Busan, Republic of Korea
e-mail: skylee@tu.ac.kr

W.-J. Hwang
Deparment of Information and Communications Engineering, Tongmyong University, Busan, Republic of Korea
e-mail: ejlee@tu.ac.kr

E.-J. Lee
Department of Information and Communications Engineering, Inje University, Gimhae, Republic of Korea
e-mail: ichwang@inje.ac.kr

K.-R. Kwon
Department of IT Convergence and Application Engineering, Pukyong National University, Busan, Republic of Korea
e-mail: krkwon@pknu.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_3, © Springer Science+Business Media Dordrecht 2014

3.1 Introduction

With the rapid development of BIT (Biology Information Technology), the need for DNA security technique has been recognized for the security of private genetic information and huge information of DNA storage as well as the copyright protection and authentication of GMO. For this reasons, DNA cryptography [1] and DNA steganography [2] have been researched from the late 90s and DNA watermarking [3–5] has been issued in the recent times.

The watermark can be embedded in non-coding DNA sequence (nCDS) or coding DNA sequence (CDS) and the embedding schemes depend on two regions. The non-coding sequence makes the signal processing for the watermark embedding be facilitated because it does not encode for protein sequence. While the non-coding DNA should not be indiscriminately referred as a junk DNA, much of this DNA has no known biological function yet. Nonetheless, the watermarking in non-coding sequence has followed from the assumption that it is just 'junk DNA'. Following the assumption, the non-coding sequence can be substituted overall to arbitrary sequence. This produces the robust problem. Therefore, the non-coding sequence is suitable to the medium for DNA steganography or fragile watermarking. The coding sequence is encoded to the protein sequence. The watermark in this sequence must not produce the change of protein profile, which is called as the amino acid preservation. Generally, it can be solved by the codon redundancy.

In this paper, we propose the availability of DWT based watermarking for coding DNA sequence and investigate the performance of mutation robustness, amino acid preservation, and security in a view of signal processing. Unlike DWT based DNA signal analysis, the coding sequence watermarking must process DWT and IDWT (inverse DWT) with the amino acid preservation. This makes the watermark embedding be not easy. Thus, the coding sequence that is obtained from IDWT of watermarked DWT coefficients must preserve the amino acid sequence before watermarking. But it is difficult to preserve the amino acid sequence by the general procedure of frequency domain watermarking; the series of DWT step, embedding step, and IDWT step. This problem is our motive of this paper. The proposed scheme process a number of sub-sequences of codons for the embedding target a watermark bit. From experimental results, we verified that the proposed scheme has more high performance of the robustness under specific range of small scale mutations and the security than DNA-crypt algorithm and also the amino acids are preserved definitely. And also, we verified that the position and level of DWT coefficients for target embedding affect the robustness because of the support region of wavelets.

3.2 Proposed DNA Sequence Watermarking

CDS watermarking has following main requirements of mutation robustness, amino acid preservation, and security. Goals of requirements are similar as them of multimedia watermarking. (1) The watermark likes as noise on a sequence. If the watermark makes the amino acid sequence be changed, similar as missense mutations, the resulting proteins will be nonfunctional. Thus, the process of watermark embedding means the intended silent mutation. The amino acid preservation corresponds to the invisibility of multimedia watermarking. (2) DNA watermarking is to verify the watermark in the pirated DNA sequence for the copyright protection or ownership verification of DNA. CDS watermarking must be designed to be robust considering all cases of mutations. (3) The watermark cannot be extracted or removed easily by anyone even though both the watermarking algorithm and biological experiment condition are known.

Our embedding scheme focus on the potential method of DWT based watermarking with the amino acid preservation and this full process is shown in Fig. 3.1. Codons are mapped to numerical codes, which are determined by the permutation table of histogram rank of amino acids. We divide a full codon sequence C into sub-sequences BC_i with N_B codons and gather sub-sequences of all synonymous codons, $\{h^{-1}(\mathrm{A}_i)\}$, of BC_i and calculate DWT coefficients, $\{g(h^{-1}(\mathrm{A}_i))\}$, of them. Then, we find a codon sub-sequence BC'_i whose DWT coefficients are optimum for the embedding condition and substitute BC_i to BC'_i. The length of watermark bits depends on the number of codon sub-sequence BC_i. The following are the details of each stage.

Four bases $\mathrm{b} = \{\mathrm{T}, \mathrm{C}, \mathrm{A}, \mathrm{G}\}$ are coded to four numbers $b = \{0, 1, 2, 3\}$ and each codon $\mathrm{c}_k = (\mathrm{b}_{3k-2}\mathrm{b}_{3k-1}\mathrm{b}_{3k})$ is mapped to an integer c_k; $c_k = 4^2 \times b_{3k-2} + 4^1 \times b_{3k-1} + 4^0 \times b_{3k}$, $c_k \in [0, 63]$. Inversely, three base numbers b_{3k-2}, b_{3k-1}, b_{3k} are obtained easily from a codon integer c_k as follows; $b_{3k-2} = \left\lfloor \frac{c_k}{4^2} \right\rfloor$, $b_{3k-1} = \left\lfloor \frac{c_k \% 4^2}{4^1} \right\rfloor$, $b_{3k} = (c_k \% 4^2) \% 4$.

We use Lifting based integer DWT transform for considering the amino acid preservation and the robustness and embeds the watermark in DWT coefficients in a different way. We divide a full codon numerical sequence C with $N/3$ codons (except for start and stop codons) into sub-sequences BC_i with $N_B = 2^3 \ll N/3\,(s \geq 1)$ codons, $BC_i = (c_{i1}c_{i2}\ldots c_{iN_B})$. The number of codon sub-sequences is $floor((N-2)/N_B$. We generate the watermark bits $W = \{w_\mathrm{i} \in [0,1] | i \in [1, N_W]\}$ with this number, which $N_W = floor((N-2)/N_B)$ and embed watermark bits in codon sub-sequences one by one. A codon numerical sub-sequence BC_i is transformed to DWT on $log_2 N_B$ levels. Thus, BC_i is decomposed to a approximation coefficient L and detail coefficients on levels from l to 1, $H_i H_{i-1} \cdots H_1$; $X_i = F(BC_i) = \{L_i H_i H_{i-1} \cdots H_1\}$, where $H_k = \left\{x_{k,1}x_{k,2} \cdots x_{k,N/2^k}\right\}$. We obtain a set $\boldsymbol{\Omega}_i$ of numerical sub-sequences of synonymous codons of BC_i and a set $\boldsymbol{X}_i$ of DWT coefficients of them. $|\boldsymbol{\Omega}_i|$ is the number of numerical sub-sequences of

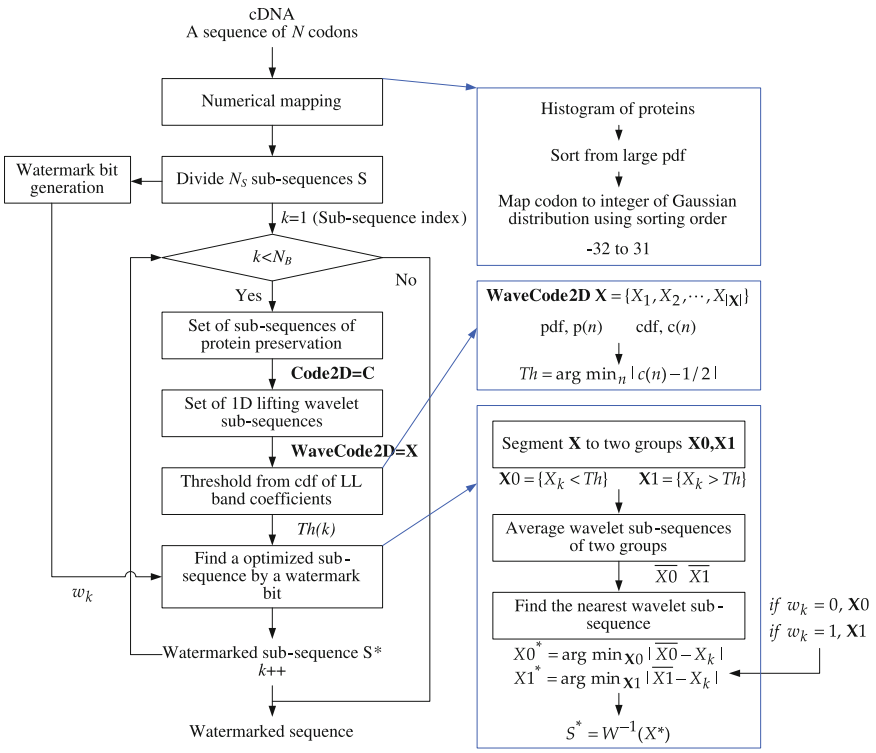


Fig. 3.1 Proposed embedding process with two steps; Codon numerical mapping, Watermark embedding on DWT domain

synonymous codons and it is the multiple of synonymous codons $|\mathrm{A}_k|$ of each amino acid; $|\boldsymbol{\Omega}_i| = \prod_{k=1}^{N_B} |\mathrm{A}_k|$.

As six-fold generated codons are changed to fourfold, the range of $|\boldsymbol{\Omega}_i|$ is $1 \le |\boldsymbol{\Omega}_i| \le 4^{N_B}$. DWT coefficients that are available for bit embedding may be numerous depending on the embedding method. We set the embedding condition for approximation coefficient L_i of codon numerical sub-seqeunce *BCi* by using the threshold Th_i as follows.

$$\begin{cases} L_i' \le Th_i, & \text{if } \mathrm{w}_i = 0 \\ L_i' > Th_i, & \text{if } \mathrm{w}_i = 0 \end{cases} \quad \text{where} \quad Th_i = \arg\ \min_n |cdf(n) - 1/2 \tag{3.1}$$

The threshold Th_i is determined to 1/2 of *cdf*(*n*), which is the cdf (cumulative distribution function), *cdf*(*n*)($n \in Z$), of approximation coefficients among a set of DWT coefficients $\boldsymbol{X}_i$. We obtain a group $\boldsymbol{X}_i'$ of all transform coefficients X_{ij}' with approximation coefficient L_i' that satisfies the embedding condition according to a watermark bit w_i, $\boldsymbol{X}_i = \{X_{i1}', X_{i2}', \cdots\}$ where $L_{ij}' \in X_{ij}' = F(BC_{ij}')$, and obtain a

group of codon numerical sub-sequences of $\boldsymbol{\Omega}_i' = F^{-1}(\boldsymbol{X}_i')$. We find a codon numerical sub-sequence BC_{ij}^* that is close to average sub-sequence $\overline{BC}_{ij}'$ of $\boldsymbol{\Omega}_i'$

$$BC_{ij}^* = \arg\ \min_{\boldsymbol{\Omega}_i'} \left| \overline{BC}_i' - BC_{ij}' \right| \forall BC_{ij}' \in \boldsymbol{\Omega}_i' \tag{3.2}$$

and then substitute BC_{ij}^*or an original codon numerical sub-sequence BC_i. As above method, $N_B N_B$its of watermark are embedded into N_B number of codon sub-sequences one by one.

The watermark should be extracted from the DNA sequence of a pirated gene. The extracting process is performed by using a key, which is the permutation table of histogram rank of amino acids, and is similar as the embedding process. The frameshift in coding sequence can be generated due to deletions or insertions. Before the extracting process, we realignment based on a watermarked coding sequence by pairwise alignment algorithm. The realignment process deletes any inserted codons and fills any deleted codons with codons.

3.3 Experimental Results

We experimented based on in silico by using DNA sequences provided in NCBI and compared our algorithm with DNA-Crypt algorithm. The more a number of codons in sub-sequence increases, the more it takes long time and high memory to find an optimum sub-sequence. We embedded the watermark by using four cases of approximation coefficients and detail coefficients of (4,3) and (8,3) respectively and evaluated the performance of four cases, which are called by L band (4,3), L band (8,3), H band (4,3), and H band (8,3). We used the lifting based integer wavelet transform with Cohen-Daubechies-Feauveau (CDF) biorthogonal wavelets for DWT and IDWT. DNA-Crypt experiment embedded 2bits into fourfold degenerated codons and used WDH(5) for mutation correction code because it is more robust than 8/4 Hamming code. All experiments were performed 1,000 repetition times in each experiment condition.

DNA watermarking is similar as the intended silent mutation. Therefore, we analyzed the change rate of codons and amino acids before and after the watermarking. In all experiments, no amino acids were changed in both our scheme and DNA-Crypt. This is a natural result because two schemes embed the watermark for the amino acids to be preserved. Under the similar capacity, our scheme produced more silent mutations from 3.7 to 5.4 times than DNA-Crypt. This is reason that all codons in support region of wavelet filters are influenced by the watermark. Thus, our scheme embeds the watermark into the variation of codons in support regions.

For evaluating the mutation robustness, we applied point mutation, insertion, and deletion at the same time to all bases except for start codon and stop codon in watermarked sequences. Thus, we set arbitrarily the point mutation rate to $2r$, the insertion rate to r, and also the deletion rate to r and then varied the mutation rate

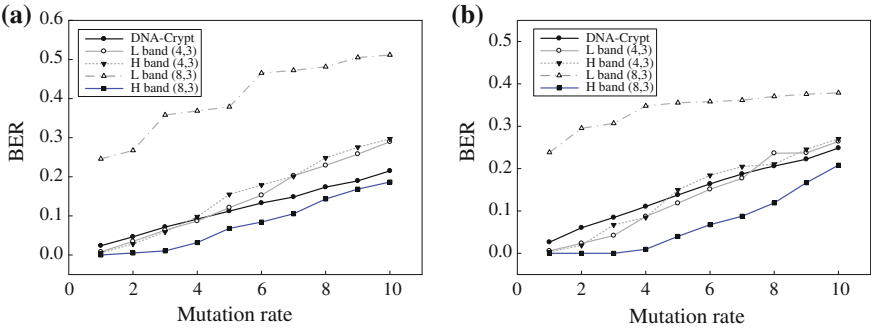


Fig. 3.2 BERs of the extracted watermarks in mutated codon sequences. (**a**) B. Subtilis and (**b**) S. Cerevisiae

from 1 to 10. For recovering the frameshift by insertions or deletions, we re-alignmented the sequence by Smith-Waterman algorithm [6] using the reference of the watermarked sequence and then extracted the watermark. The robustness evaluation was used as BER (bit error rate) of the extracted watermark. BERs of the extracted watermark in mutated sequences are shown in Fig. 3.2. H band (8,3) method has higher BERs than other methods, which means that it is not effective for the watermark embedding. L band (4,3) and H band (4,3) methods have higher BER than DNA-Crypt in high mutation rate, which means that they are effective for the embedding in low mutation rate. H band (8,3) method has lower BER than other methods, which means that it is very effective for the watermark embedding. These results verified that when the mutation rate is low, the order of robustness is H band (8,3), H band (4,3), L band (4,3), DNA-Crypt, and L band (8,3) and when the mutation rate is high, the order of robustness is H band (8,3), DNA-Crypt, H band (4,3), L band (4,3), and L band (8,3).

3.4 Conclusions

We have introduced a lifting based DWT watermarking scheme based on for coding DNA sequence. The proposed scheme converts codons to numerical codes by using the permutation table of histogram rank of amino acids, which is a watermark key, and embeds the watermark in L band or H bands of low levels in DWT domain of codon numerical sub-sequences. The embedding process should be dissimilar as the general process of forward DWT, watermark embedding, and inverse DWT for preventing the changeability of amino acids by watermarked DWT coefficients. Our experiments embedded the watermark three times repeatedly in L band and H bands of low levels of sub-sequences of 4 and 8 codons, so-called L band (4,3), L band (8,3), H band (4,3), and H band (8,3) methods. The results of experiments confirmed that our methods have the amino acid

preservation and also the security and specially H band (8,3) method has more the mutation robustness and capacity than DNA-Crypt and other methods. We recommend H band (8,3) method for embedding the robust and secure watermark.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2011-0023118), and by the framework of international cooperation program managed by National Research Foundation of Korea (2012K2A1A2032979), and by the Busan Metropolitan City, Korea, under the 2013 Brain Busan 21 program grants.

References

1. Gehani A, LaBean T, Reif J (2004) DNA-based cryptography. Aspects of molecular computing. Lecture notes in computer science, vol 2950/2004, pp. 34–50
2. Risca VI (2001) DNA-based steganography. Cryptologia 25(1):37–49
3. Heider D, Barnekow A (2011) DNA Watermarking: challenging perspectives for biotechnological applications. Curr Bioinform 6(3):375–382
4. Heider D, Barnekow A (2007) DNA-based watermarks using the DNA-Crypt algorithm. BMC Bioinform 8:176
5. Lee S-H, Kwon S-G, Kwon K-R (2012) Robust DNA watermarking based on coding dna sequence. J IEEK 49(2):123–132
6. Smith T, Waterman M (1981) Identification of common molecular subsequences. J Mol Biol 147:195–197

Chapter 4
Exclusive Region Based Scheduling for MB-OFDM UWB Using Hidden Pilot Sequences

Byung Wook Kim and Byung-Gyu Kim

Abstract This paper deals with an exclusive region (ER)-based scheduling scheme in multiband orthogonal frequency division multiplexing (MB-OFDM) ultra-wideband (UWB) with hidden pilot sequences. Deriving a generalized radius for the ER using information of statistical topology, a novel scheduling approach that uses a radius for the ER is introduced. Numerical results show that scheduling that is based on proposed ER radius can always outperform the time division multiple access (TDMA) scheme with respect to network throughput and the proposed approach to the radius of the generalized ER provides better scheduling performance than the radius of the conventional method.

Keywords MB-OFDM UWB · Hidden pilot sequence · Exclusive region · Scheduling

4.1 Introduction

Due to its wide bandwidth and low complexity, multiband-orthogonal frequency division multiplexing (MB-OFDM) ultra-wideband (UWB) radios provide high-rate systems that are capable of supporting multiple users with variable rates and can be used in wireless personal area networks (WPANs) for various applications [1–3]. For the increasing demand on high-rate multimedia services, a hidden pilot

B. W. Kim (✉)
School of Railway and Electrical Engineering, Kyungil University, 50 Gamasilgil, Hayangup, Gyeongsan-si, Gyeongsangbuk-do, Republic of Korea
e-mail: bwkim@kiu.kr

B.-G. Kim
Department of Computer Science and Engineering, Sunmoon University, 100 Kalsan-ri, Tangjeong-myeon, A-san city, Chungnam, Republic of Korea
e-mail: bg.kim@mpcl.sunmoon.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_4, © Springer Science+Business Media Dordrecht 2014

based scheme is a promising candidate due to its capability of high bandwidth efficiency and frequency diversity [4].

The network scheduling based on exclusive region (ER) [5, 6] have been introduced because their utilization improves network throughput. When the ER is set around the receiving device of the desired link, the interfering devices in the same time slot that are transmitting concurrently should be outside the ERs of other links, to ensure that concurrent transmissions are beneficial. However, the report of performance of the conventional work on ER-based scheduling pertains only to networks based on Voronoi diagrams. In addition, the problem of ER-based scheduling for the schemes of high bandwidth efficiency has not yet been considered in the literature.

Here, ER-based scheduling with hidden pilot sequences, which can be applied to MB-OFDM UWB, is introduced. The general expression of the optimal ER radius, which is based on statistical topology, for the hidden pilot based scheme is derived. Simulation results show that the use of the proposed scheme provides better overall throughput performance than other scheduling schemes.

4.2 System Model

Consider MB-OFDM UWB systems with the network configuration that is defined in the IEEE 802.15.3 standard. A central controller has the function of performing control functions, including network scheduling, at the MAC layer. Between the central controller and all the other devices in the network, control data are exchanged, while information data are transmitted through peer-to-peer links where transmitters and receivers are directly connected.

Since wireless links are considered in this system, the channel estimation has a direct effect on the overall performance of the system. In comparison to the conventional channel estimation methods [7, 8], a hidden pilot based scheme [4] provides frequency diversity and less interference between the data and pilot symbols, while maintaining high efficiency in bandwidth utilization.

Herein, our purpose is to determine how to schedule concurrent transmissions so as to maximize the network throughput. The devices in a piconet can communicate with one another by using the same time-frequency code (TFC). Therefore, when a new device wants to join the piconet, it should detect the TFC number of that piconet. Because the preamble pattern associates with the TFC number, the TFC number of the interfering piconet can be obtained by recognizing the preamble pattern.

It is important to note that the data rate in UWB communications is proportional to the received signal to interference plus noise ratio (SINR). According to the Shannon theory, the channel capacity of the hidden pilot-aided scheme can be presented as follows:

$$\begin{aligned} R_l &= W \log_2(1 + SINR) \\ &= W \log_2\left(1 + \frac{p_{l(s)}}{N_l + p_{l(b)} + I_l}\right), \end{aligned} \tag{4.1}$$

where $p_{l(s)}$ is the received symbol power of of the link l, $p_{l(b)}$ is the received hidden pilot power of of the link l, R_l is the channel capacity of link l, N_l is the one-sided spectrum level of white Gaussian noise, W is bandwidth, and I_l is the interference term. Note that the UWB system has extremely high bandwidth, i.e., $W \to \infty$. Therefore, the channel capacity can be simplified as

$$\begin{aligned} R_l &= W \log_2\left(1 + \frac{p_{l(s)}}{N_0 W + p_{l(b)} + I_0 W}\right) \\ &\approx \frac{p_{l(s)}}{N_0 + \frac{p_{l(b)}}{W} + I_0} \log_2 e, \end{aligned} \tag{4.2}$$

where I_0 is the power spectral density of I_l. It is important to note that channel capacity is a form of the linear function of SINR and can be adjusted proportionally to the SINR of the receiver to meet the requirement for the bit error rate (BER).

4.3 Proposed Exclusive Region-Based Scheduling

The considered network topology consists of UWB devices that are distributed according to a two-dimensional Poisson process in the plane. By considering a random connection model in which each pair of points of a two-dimensional Poisson process is randomly connected, the links between transmitting devices and receiving devices are designed. In order to implement scheduling each link to time slots, the generalized ER radius for concurrent transmissions should be derived.

Here, the formation of networks that have a statistical topology, rather than the fixed topology illustrated by the Voronoi diagram [5], is presented. Given that the data rate of UWB communications is proportional to the SINR, the data rate of a desired receiver l is

$$R_l = k' \frac{p_{l(s)}}{N_0 + \frac{p_{l(b)}}{W} + \frac{zq}{W}\bar{M} \int_{-\pi}^{\pi} \int_{s_l}^{\infty} p(r, r_c, \theta, \theta_c) r^{-\alpha} dr d\theta}, \tag{4.3}$$

where k' is a scaling constant, s_l is the radius of the ER of lth link, r is the sender-receiver distance, q is the transmit symbol power, z is the multiuser interference (MUI) factor, which represents the average overlap between any two TFCs, and $\bar{M}$ is the number of interferers.

To express the received power, an attenuation model [9], which has been confirmed as a suitable indoor propagation model for UWB-WPAN, is used.

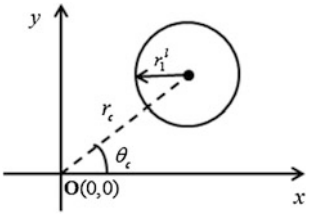


Fig. 4.1 Exclusive region-based coverage of a receiver

The received symbol and pilot power is set to $p_{l(s)} = q\sigma_s^2 r_l^{-\alpha}$ and $p_{l(b)} = q\sigma_b^2 r_l^{-\alpha}$ respectively, where r_l is the sender-receiver distance of l-th link, and α is the path-loss exponent.

As depicted in Fig. 4.1, the location of the designated receiver is presented by (r_c, θ_c). $p(r, r_c, \theta, \theta_c)$ denotes the probability density function (PDF) of the distribution of the sensors, which shows the network topology where the center of the network is the designated receiver. Given that Eq. (4.3) uses a data rate equation with statistical term, it can be applied to any kind of network. The general solution for the optimal ER radius based on statistic topology in uniform networks is considered. Given that the number of concurrent transmitting links is M and the PDF of the uniformly distributed network is a constant, i.e., $p(r, r_c, \theta, \theta_c) = 1/M$, Eq. (4.3) can be rewritten as

$$R_l \approx k' \frac{p_{l(s)}}{N_0 + \frac{p_{l(b)}}{W} + \frac{zq}{W}\int_{s_l}^{\infty} r^{-\alpha} dr} = k' \frac{p_{l(s)}}{N_0 + \frac{p_{l(b)}}{W} + \frac{zqs_l^{-\alpha+1}}{W\,\alpha-1}}, \tag{4.4}$$

where $\bar{M} \approx M$. For simplicity, s_l is replaced to s. Then the total throughput considering M concurrent transmissions is given by

$$\sum_{M} R_i \approx k' \frac{M\,\bar{p}_{(s)}}{N_0 + \frac{\bar{p}_{(b)}}{W} + \frac{zqs^{-\alpha+1}}{W\,\alpha-1}} = k' \frac{A\,\bar{p}_{(s)}}{s^2\left(N_0 + \frac{\bar{p}_{(b)}}{W}\right) + \frac{zqs^{-\alpha+3}}{W\,\alpha-1}}, \tag{4.5}$$

where $\bar{p}_{(s)}$ is the average received symbol power, $\bar{p}_{(b)}$ is the average received pilot power, and A is the radius of a network. Here, it is assumed that $\mathrm{M} \approx A^2/s^2$ for simplicity. The maximum throughput can be obtained by minimizing the denominator of Eq. (4.5). Thus, the optimal ER radius is given by

$$s^* = \left(\frac{2(\alpha-1)\left(N_0 W + \bar{p}_{(b)}\right)}{zq(\alpha-3)}\right)^{\frac{1}{-\alpha+1}}. \tag{4.6}$$

Compared to the conventional approach using an approximated diagram, the optimal ER radius from the PDF of the network topology is more practical for a real network. Since the radius is a function of the path-loss exponent, the level of background noise, the average level of received pilot signal and the MUI factor, an acceptable computational complexity is required to obtain the radius of ER.

Table 4.1 Simulation parameters

Bandwidth	500 MHz	Normalized pilot power	0.7
Center frequency	5.092 GHz	Shadowing parameter	4.3
Transmit power	−14 dBm	Path-loss exponent	4
Noise power per Hz	−78 dBm/Hz	Nakagami factor	4

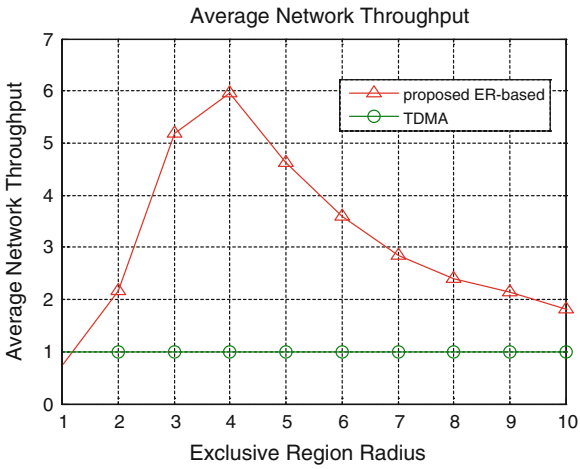


Fig. 4.2 Network throughput performance in the uniformly distributed network

4.4 Simulation Results

In this section, simulation results are presented to verify the performance of the proposed scheduling scheme. The considered UWB network topology consists of 500 devices, deployed randomly in a $20 \times 20\,\mathrm{m}^2$ device field. For concurrent communications, 80 devices are selected at random in each scheduling cycle, half of which were sources and the rest of half destinations. The relevant parameters are listed in Table 4.1. Since the TFC is considered, the MUI factor z was set to 0.3. It is assumed that the network topology remained fixed during each scheduling cycle. The UWB fading model [5], which takes into account both small-scale and large-scale fading, is used. To evaluate the effect of the proposed algorithm, normalized network throughput, which normalizes the network throughput to the value of time division multiple access (TDMA) scheduling, is used. Each simulation was performed 1,000 times with different random seeds to attain the average performance results.

Figure 4.2 shows a comparison of the performance of proposed ER scheduling and TDMA scheduling in the uniformly distributed networks. Note that the performance of the proposed ER-based scheme in Fig. 4.2 is based on Eq. (4.6) where s^* is set to the values of the x axis. It is observed that the proposed ER-based scheme shows the best performance when s^* in Eq. (4.6) is set to 3–4 m. It means

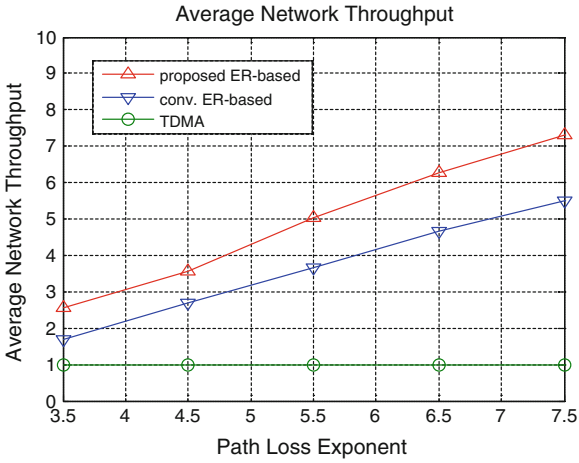


Fig. 4.3 Network throughput performance versus the path-loss exponent in the uniformly distributed network

that the average of the optimal ER radius of each device under the given network is 3–4 m. Note that this value of the ER radius presenting the maximum gain of throughput is close to that of the optimal ER radius derived in Eq. (4.6).

Figure 4.3 presents a comparison of normalized network throughput of the scheduling strategies according to the path loss exponent for the offices and the indoor environments in the uniformly distributed network. It can be seen that the scheduling with our proposed ER radius solution yields better performance than the conventional ER-based scheme in [5]. Note that our ER solution is robust to any kind of device locations by using a function of general expression of the PDF rather than using a fixed device formation. In addition, the figure shows that the proposed scheduling algorithm performs much better than the TDMA schemes with respect to average network throughput.

4.5 Conclusions

For the various wireless services with high-rate demands in WPAN, we proposed a scheduling scheme with ER radius for hidden pilot-aided MB-OFDM UWB systems. A generalized solution for the concurrent transmission with ER radius, which is based on the statistical topology, is derived. The simulation results show that scheduling with our generalized ER radius outperforms scheduling with TDMA and a conventional ER radius.

References

1. See: http://www.ieee802.org/15/pub/TG3a.html."
2. Liu KH, Ling X, Shen X, Mark JW (2008) Performance analysis of prioritized MAC in UWB-WPAN with bursty multimedia traffic. IEEE Trans Veh Technol 57:2462–2473
3. IEEE P802.15: Multiband OFDM physical layer proposal for IEEE 802.15 task group 3a (Doc. Number P802.15-03/268r3) (2004)
4. Kim BW, Jung SY, Kim JT, Park DJ (2008) Hidden pilot based precoder design for MIMO-OFDM systems. IEEE Commun Lett 12:657–659
5. Liu KH, Cai L, Shen X (2008) Exclusive-region based scheduling algorithms for UWB WPAN. IEEE Trans Wirel Commun 7:933–942
6. Cai LX, Cai L, Shen X, Mark JW (2010) REX: a randomized exclusive region based scheduling scheme for mmWave WPANs with directional antenna. IEEE Trans Wirel Commun 9:113–121
7. Fan J, Yin Q, Wang W, Ding L (2009) Pilot-aided channel estimation schemes for OFDM systems with cyclic delay diversity. IEEE Vehicular Technology Conference pp 1–5
8. Bolcskei H, Heath RW, Paulraj AJ (2002) Blind channel identification and equalization in OFDM based multiantenna systems. IEEE Trans Signal Process 50:96–109
9. Ghassemzadeh SS, Tarokh V (2003) UWB path loss characterization in residential environments. In: Proceeding of IEEE RFIC symposium pp 501–504

Chapter 5
The Digital Video Frame Obfuscation Method in Consideration of Human Visual System for Intellectual Property Protection and Authentication

Heehyung Bu, Jonghwa Kim and Sungho Kim

Abstract This paper is to propose the digital video frame obfuscation method for intellectual property protection and authentication by using the DC component of an intra frame and the motion vector of an inter frame at digital video encoding. The proposed method is based on characteristics of the HVS (human visual system) which is sensitive at a low frequency and a middle frequency. This method makes a video signal distorted, so that the video is normally displayed only when a suitable authentication signal is used.

Keywords Digital video frame obfuscation · Intellectual property protection · Authentication signal · Human visual system

5.1 Introduction

Recently, as the digital contents industry grew, publication and distribution of digital contents have been progressing rapidly. However, this rapid growth causes problems as well such as unauthorized access, falsification, and copy. In this respect, techniques for DRM (digital rights management) are required for quality assurance and intellectual property protection of digital contents [1, 2].

H. Bu (✉) · J. Kim
Department of Computer Engineering, Mokpo National University,
Jeonnam 534-729, South Korea
e-mail: hhbu@mokpo.ac.kr

J. Kim
e-mail: jhkim@mokpo.ac.kr

S. Kim
Department of Computer Engineering, Kyungpook National University,
Deagu 702-701, South Korea
e-mail: shkim@knu.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_5, © Springer Science+Business Media Dordrecht 2014

The research on DRM technology in the field of the digital video mainly focuses on digital video watermarking.

The existing methods for the digital video watermarking are usually of using DCT coefficient, motion vector, and GOP structure [3, 4]. However, most watermarking technologies have a lack of at least one more for videos with a variety of brightness and colors. Therefore, the existing methods unavoidably make a loss on one hand.

To return to the beginning, the digital video watermarking is defined as technique to watermark the information of intellectual property rights into the digital video contents. This technique is able to prove ownership by extracting a watermark when a dispute takes place over the intellectual property rights. However, the existing digital video watermarking cannot be a solution to the problems such as degradation of video quality, increase of compressed bit-stream, attack by removal of high frequency, and recompression.

For that reason, this paper proposes the digital video frame obfuscation method. This method, based on the ownership information, can make a video displayed normally only when a user possesses the authentication signal.

The proposed method can be a solution to proof of the signature even by attacking an unauthorized user inversely.

In Sect. 5.2, there is explanation about signature and m-sequence. In Sect. 5.3, the digital video frame obfuscation method is described in detail. In Sect. 5.4, the experimental result is showed. In Sect. 5.5, the conclusion is made.

5.2 Authentication Signal Consisting of Signature Signal and M-sequence

The authentication signal used in this paper is made of a signature signal and m-sequence. The signature signal consists of a header with a block size and a signature encrypted by RSA. RSA cryptosystem performs encryption and decryption operations [5]. M-sequence is a bit-stream generated by the linear feedback shift register, as a sequence characterized by little correlation between sequences [6].

The authentication signal is generated by combination of the two arrays of the signature signal and m-sequence. Figure 5.1 shows the procedures of signature signal generation from the information on the intellectual property rights and signature letter creation from the generated signature signal inversely. At the stage of signature signal generation, a header of 8 bits has a block size for reading the signature signal as block unit at the detection stage. The detection stage of signature needs owner's private key because the information of intellectual property rights can be detected only by owner's private key.

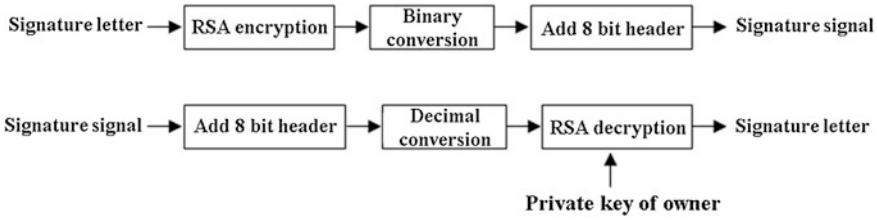


Fig. 5.1 Block diagram of generation of signature signal and detection of signature

5.3 The Digital Video Frame Obfuscation Method Using DC Coefficient and Movement Vector

This paper proposes the digital video frame obfuscation method using DC coefficient of an intra frame and movement vector of an inter frame. The proposed method minimizes computation costs by using key parts to respond sensitively to the HVS. The HVS has characteristics that are sensitive at the low frequency domain including major information of an image and the middle frequency domain including edges, but insensitive relatively at the high frequency.

The Fig. 5.2 shows that HVS responds sensitively when spatial frequency is 5–10 cycles per degree but responds insensitively relatively at the high frequency domain.

The video signal is distorted by the XOR operation between the DC components of an intra frame and the authentication signal, and the XOR operation between the movement vectors of an inter frame and the authentication signal. The DC components are included at the low frequency and movement vectors are included at the middle frequency.

Figure 5.3a is a diagram for the basic MPEG encoder scheme adding an authentication signal to the XOR operation except the estimation and compensation of movement. In the proposed encoder, the first XOR operation is performed before the different value of the actual DC coefficients is put into a bit-stream. The second XOR operation is performed before the sign of the different movement vector values is put into a bit-stream. As stated above, a final outcome is the distorted bit-stream of the compressed video.

In the proposed decoder of Fig. 5.3b, first all the distorted bit-stream of the compressed video comes in, DC sizes of an intra frame are used for the variable length decode process, and then the different values of the DC components are used for the XOR operation with the authentication signal. Also, at an inter frame, the absolute different values of movement vectors are used for variable length decoding process and then the signs of the different values are used for XOR operation with the second authentication signal.

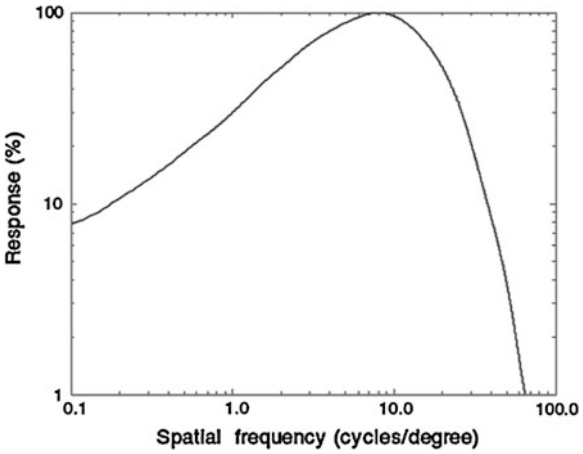


Fig. 5.2 The spatial frequency response of the human visual system

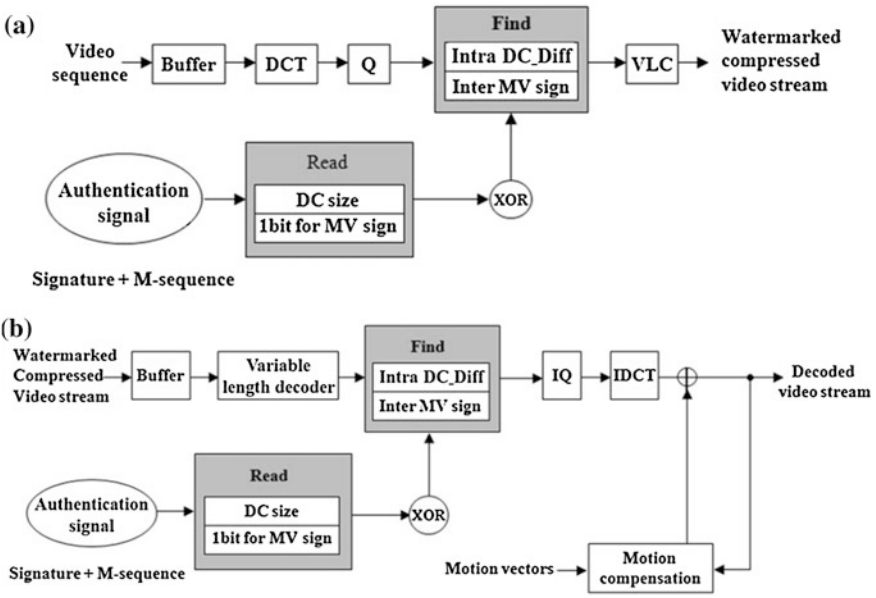


Fig. 5.3 **a** Block diagram of the proposed encoder, **b** Block diagram of the proposed decoder

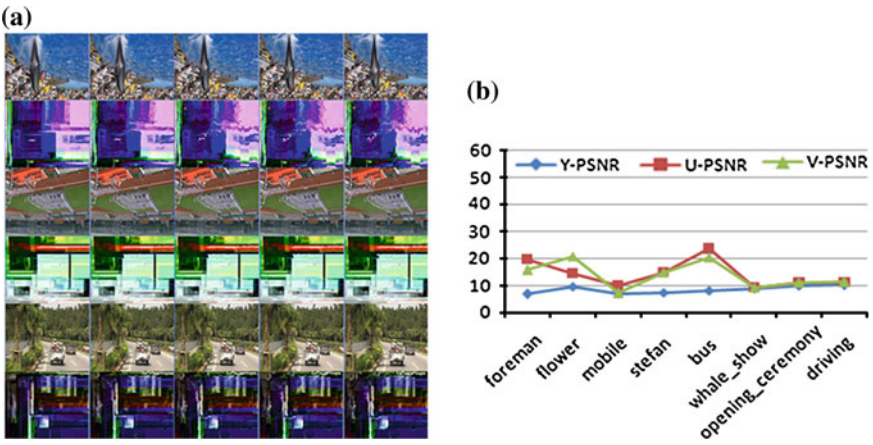


Fig. 5.4 **a** Three SD videos from DRM (*each top*: authorized use, *each bottom*: unauthorized use), **b** Average PSNR of eight test videos distorted

5.4 Experimental Results

The experiment checks whether the performance is normal or not by comparing the PSNR values for the original compressed video when the authentication signal is used or not. MPEG-4 VM was used and the implementation environment was Intel® Core™ 2 Duo CPU E8400 @ 3.00 GHz, 3.26 GB RAM, and Windows 7. Test videos were CIF (352 × 288) 5 and SD (720 × 480) 3 [7, 8].

In Fig. 5.4a, XOR operation was performed with the authentication signal and not at decoder. As a result, it could be confirmed that in case of using (a) suitable authentication signal, the videos were displayed normally and in case of not using an authentication signal, the videos were displayed in a severely distorted manner. The PSNR (peak signal to noise ratio) was used for quality assessment for test videos as objective quality. The MSE (mean square error) method is used for computation of errors. The equations are as follows [9].

$$\text{PSNR}\,[\text{dB}] = 10\log_{10}\frac{255^2}{\text{MSE}}\,[dB] \tag{5.1}$$

$$\text{MSE} = \frac{1}{\text{M}\times\text{N}}\sum_{\text{i}=0}^{\text{M}-1}\sum_{\text{j}=0}^{\text{N}-1}(\text{x}(\text{i},\text{j})-\text{x}'(\text{i},\text{j}))^2 \tag{5.2}$$

In the Eq. (5.2) above, M and N are the number of rows and columns in the video frames. x(i, j) is a pixel value of the original frame and x′(i, j) is a pixel value of the test video frame.

In Fig. 5.4b, the values of U and V PSNR in CIF videos are relatively higher than those in SD videos. This occurs when the value of an authentication signal

Table 5.1 Comparison of the proposed method and the existing method

	The proposed method	The existing method
Applied part	DC coefficient, movement vector	DC coefficient
Bit-stream	Remained	Increased
Signature information	Available	None

appears zero in a row. However, videos were displayed abnormally as intended originally due to the effect appearing in a small part only within a frame.

As a result, when an appropriate authentication signal was used for the entire video signal, Y, U, and V PSNR values all lead to its peak and every video was displayed normally. However, when an authentication signal was not used, average PSNR value was 12.25 dB as low value and every video was displayed abnormally in a severely distorted way. When videos were displayed normally, the procedure of signature detection proceeded. According to Fig. 5.1 of Sect. 5.2, owner's signature of intellectual property rights was detected.

Next, in comparison with the existing method [2], disadvantage, besides the Table 5.1, is that information of edges remains as it is, because the edge information usually exists at a middle frequency, not a low frequency including DC coefficient.

5.5 Conclusion

Consequently, the proposed method was able to keep bit-stream, to conduct procedure of authentication normally over the entire video signal, and also to prove intellectual property rights.

References

1. Lee CR (2002) Digital rights management. J TTA 82:91–94
2. Park SH, Choi HJ, Seo YH, Kim DW (2005) Ciphering scheme and hardware implementation for mpeg-based image/video security. J Inst Electron Eng Korea 42(SP 2):27–36
3. Liu F, König H (2010) A survey of video encryption algorithms. Comput Secur 29(1):3–15
4. Langelaar GC, Lagendijk RL, Biemond J (1998) Real-time labeling of mpeg-2 compressed video. J Vis Commun Image Represent 9(4):256–270
5. Menezes Alfred J, Van Oorschot PC, Vanstone SA (1997) Handbook of applied cryptography. CRC Press, Boca Raton
6. Golomb SW (1967) Shift register sequences. Holden-Day, San Francisco
7. ISO/IEC 14496-2:2001(2001) Information technology: coding of audio-visual objects—Part2: visual, ISO/IEC JTC1/SC29/WG11
8. http://scien.stanford.edu/pages/labsite/scien_test_images_videos.php
9. Fibush DK (1997) Practical application of objective picture quality measurements. In: IBC97 pp 123–135, International Broadcast Union

Chapter 6
Fuzzy-Based Resource Reallocation Scheduling Model in Cloud Computing

Jaekwon Kim, Taeyoung Kim, Minoh Park, Youngshin Han and Jongsik Lee

Abstract A cloud computing system consists of physical resources for processing large-scale tasks. With a recent trend of rapidly growing data, a cloud computing system needs a processing method to process a large-scale task in a physical resource. Generally, a physical resource divides a requested large-scale task to several tasks. And a processing time of each divided task varies with two factors which are processing efficiency of each resource and distance between resources. Although a resource completes a task, the resource is standing by until all divided tasks are completed. When all resources complete a large-scale task, each resource can start to process a next task. In this paper, we propose a Fuzzy-based Resource Reallocation Scheduling Model (FRRSM). Using fuzzy rule, FRRSM reallocates an uncompleted task to with a resource in considering efficiency and distance factors of the resource. FRRSM is an efficient method for processing a large-scale task or multiple large-scale tasks.

Keywords Cloud computing · Resource reallocation · Fuzzy

J. Kim · T. Kim · M. Park · J. Lee (✉)
Department of Computer and Information Engineering, Inha University, Inchon, South Korea
e-mail: jslee@inha.ac.kr

J. Kim
e-mail: jaekwonkorea@naver.com

T. Kim
e-mail: silverwind@gmail.com

M. Park
e-mail: park-min-o@nate.com

Y. Han
Division of Liberal Arts and Teaching (Information and Communication), Sungkyul University, Anyang, South Korea
e-mail: hanys@sungkyul.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_6, © Springer Science+Business Media Dordrecht 2014

6.1 Introduction

A cloud computing environment handles a large-scale task processing. And the system performance depends upon the resource states and task scheduling [1]. However, the task scheduling issue is well-known as NP-complete problem [2]. The task scheduling issue still has many challenges for researchers, and many studies bring out various solutions to overcome the task scheduling problem [3]. Typically, the scheduling policy applies a partitioning approach for the large-scale task. The cloud computing also divides large tasks into processing units according as the composition of each resource. However, the system efficiency has an influence upon the performance and distance of resources. And those factors also affect on a processing time.

In order to allocate the cloud resource for processing, existing research considers a task size and priority. However, the resource state has less concerned for scheduling. Even if some resource is finished the task, the available resource should wait until the completion of other busy resources. Thus, more study is needed to utilize those available resources. If the system provides a resource reallocation policy for available resources, It may improve a task processing speed and a utilization of the resources. The resource reallocation is the method to allocate the other large-scale task to the available resource. The method considers a network state on the virtualization environments [4]. Hence, the simulation-based verification is needed to adopt on the cloud computing system.

This paper proposes a fuzzy-based resource reallocation scheduling model (FRRSM). The proposed FRRSM includes the reallocation method for early finished available resource. The FRRSM relocates the available resource to other virtual machines and assigns a new large-scale task. In this process, the FRRSM utilizes a fuzzy method. The fuzzy logic for FRRSM handles a queue size of resources, a job size and a network distance as fuzzy input variables. The fuzzy logic for FRRSM handles a queue size of resources, a job size and a network distance as fuzzy input variables. And the logic finds the resource for reallocation with rule-based selection. To prove the performance of FRRSM, we design the virtual cloud computing environment with the DEVS-based modeling and simulation methodology [5].

6.2 Related Works

Some related study is underway to detect and solve the resource hotspot of the VM on the virtualization environments. In order to deploy the VM, Sandpipier [6] utilizes the state of the CPU, RAM and network. These states named ‘volume’ are the basis of the migration and replacement of the VM. However, this volume has focused on the simple viewpoint such as CPU-side states. It does not fit to utilize through the geographically distributed system likes cloud environments. Real time resource reallocation for better resource utilization (RTRRBRU) [7] calculates the

usage of the CPU and memory in real time. RTRRBRU can measure the workload and real time reallocates the VM. This approach may reduce the workload for several resources. However, the reallocation in real time increases the network traffic. Thus RTRRBRU also does not fit to the cloud environments. And the cloud computing includes the dynamic resource allocation policy with the global scheduling method. However, the global scheduling only consider the resource allocation. Therefore, the cloud environments require the dynamic resource reallocation to manage the job finished resource.

6.3 Method

6.3.1 Dynamic Resource Reallocation in Cloud Environment

In the cloud environment, a virtual machine holds several physical machines. And each virtual machine processes its assigned large-scale task. The task managing of the virtual machine basically stores the user's request on the task queue. The manager constructs sub tasks from the task in the queue. The sub task is the MapReduce-based splits of the large-scale task. The manager sends sub tasks to the physical machine, which composed of virtual machines. And the physical machine processes the received sub task. After the processing is complete, the system sends the sub tasks' processing information to the virtual machine. And the sub task managing module merges the gathered information. A dynamic resource reallocation is the method to transfer an available resource between virtual machines. It has no trouble in processing the task on virtual machines. As the heavy-loaded resource may grants more available resources, it may improve the processing speed and time for a plural large-scale task.

6.3.2 Resource Reallocation Algorithm

Figure 6.1 shows the architecture of the FRRSM. FRRSM is composed of five components.

Job Generation listens the user's request for large-scale tasks, and creates the task for processing. Coordinator assigns the generated task to each VM. The VM has one or more physical machines. Resource Detector traces available resources (physical machine). Resource Detector traces available resources (physical machine). When Resource Detector finds some available resource, Monitoring obtains state information of physical machines. And then, Resource Reallocation assigns available resources to other VM using the fuzzy methods. Due to the cloud environment is the geographically distributed, most important factor is the performance of physical resources. And the distance between the physical machine

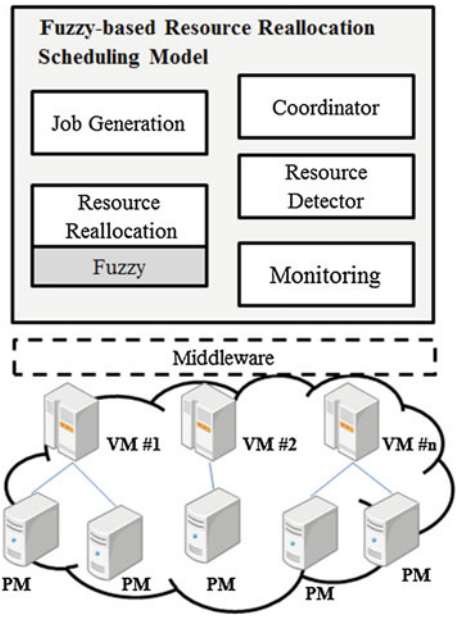


Fig. 6.1 FRRSM architecture

and virtual machine produces some effect on the processing efficiency. Hence, the resource manager should consider the distance and the job finishing time.

This paper utilizes the fuzzy method for resource reallocation. The fuzzy method calculates the queue size, job size and distance of each VM, and estimates the performance of the physical machine. The fuzzy parameter is composed of the queue size (μA), number of job (μB) and distance (μC). The fuzzy logic utilizes these parameters, and infers the Avail_Score (μD) to find a VM for resource reallocation. Each parameter (VL = VeryLow, L = Low, M = Middle, H = High, VH = VeryHigh) is presented as following Fig. 6.2.

Both μA and μB are state information of the queue and job size from the VM. And μC is the calculated distance between each VM. The distance may estimate with the Cartesian coordinate system and the Euclidean distance [6]. In order to infer the output μD, the logic utilizes three fuzzy input variables with 45 rule bases. Table 6.1 shows the 45 rule bases. The fuzzy inference engine is the Mamdani method, and selects the highest scored (μD) VM for resource reallocation.

6.4 Experiment

In order to prove the performance of FRRSM, we design the experiment environment with DEVS methodology. We select the Round-Robin scheduling model (RRSM) for comparison. And the experiment measures an average of job processing time using the RRSM with FRRSM and non-FRRSM.

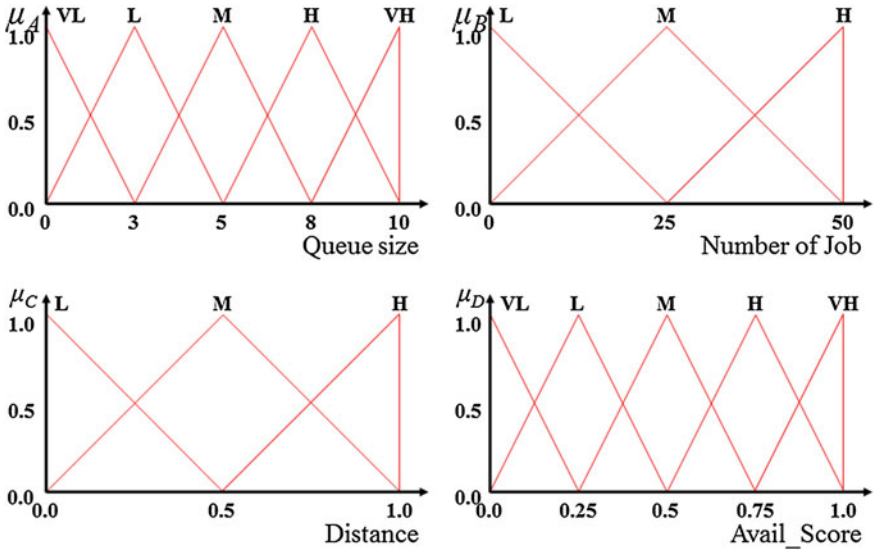
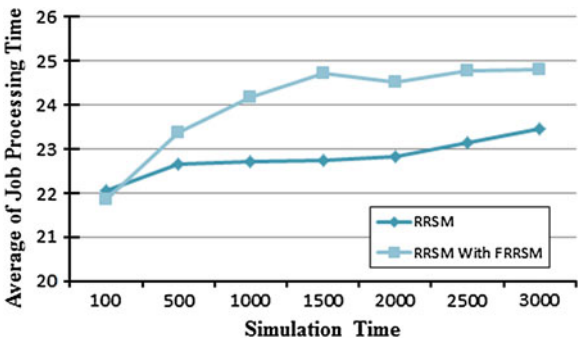


Fig. 6.2 Fuzzy variable

Table 6.1 Rule bases

μ_A, μ_B, μ_C		VL	L	M	H	VH
L	L	VH	VH	H	M	L
	M	VH	H	M	L	VL
	H	H	H	M	L	VL
M	L	VH	H	H	M	L
	M	VH	H	M	L	VL
	H	H	M	M	L	VL
H	L	VH	H	M	L	VL
	M	H	M	L	L	VL
	H	M	M	L	VL	VL

Fig. 6.3 Experiment result



As shown in Fig. 6.3, RRSM records 23.469, and RRSM with FRRSM records 24.814 average of job processing time for 3,000 simulation times. The above result shows that our proposed FRRSM is the effective method for large-scale task processing.

6.5 Conclusions

In this paper, we propose the fuzzy-based resource reallocation model to improve the efficiency of large-scale task processing on the cloud computing. The proposed method reallocates the available resources to other VM. To reallocate the resources, we utilize the fuzzy method. And we prove the effectiveness of our proposed method with the experimental result.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2002751) and this research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2013R1A1A3A04007527).

References

1. Calheiros R, Ranjan R, Buyya R (2011) Virtual machine provisioning based on analytical performance and QoS in cloud computing environments. In: International conference on parallel processing (ICPP), Taipei, Taiwan, pp 295–304
2. Bokhari SH (1987) Assignment problems in parallel and distributed computing. Kluwer Academic Publisher, Berlin
3. Munir EU, Li J, Shi S, Zou Z, Yang D (2008) MaxStd: a task scheduling heuristic for heterogeneous computing environment. Inf Technol 7:679–683
4. Weissman JB, Lee BD (2002) The virtual service grid: an architecture for delivering high-end network services. Concurr Pract Exp 14(4):287–319
5. Zeigler BP et al (1996) DEVS framework for modeling, simulation, analysis and design of hybrid systems in hybrid II. Lecture notes in CS, Springer-Verlag, Berlin, pp 529–551
6. Wood T, Shenoy P, Venkataramani A, Yousif M (2007) Black-box and gray-box strategies for virtual machine migration. In: Proceedings of the 4th USENIX symposium on networked systems design and implementation, pp 229–242
7. Sijin H, Li P, Yike G (2011) Real time elastic cloud management for limited resource. In: Cloud 2011 IEEE international conference on computing (CLOUD), Washington, pp 622–629
8. Tan PN, Steinbach M, Kumar V (2007) Introducton to Data Mining, Addison Wesley, Boston, pp 66–69

Chapter 7
An Efficient Algorithm for High Utility Sequential Pattern Mining

Jun-Zhe Wang, Zong-Hua Yang and Jiun-Long Huang

Abstract High utility sequential pattern mining is to mine sequences with high utility (e.g. profits) but probably with low frequency. In some applications such as marketing analysis, high utility sequential patterns are usually more useful than sequential patterns with high frequency. In this paper, we devise two pruning strategies RSU and PDU, and propose HUS-Span algorithm based on these two pruning strategies to efficiently identify high utility sequential patterns. Experimental results show that HUS-Span algorithm outperforms prior algorithms by pruning more low utility sequences.

Keywords High utility sequential pattern mining · High utility sequences · Utility mining

7.1 Introduction

High utility sequential pattern mining is to find sequences in a sequence database with utility no less than a user-specified minimum utility threshold. However, mining high utility sequential patterns is not an easy task. The major challenge is that utility is not of the downward closure property. Thus, frequent sequential pattern mining algorithms cannot be directly applied to mine high utility sequential patterns. To address this problem, Sequence Weighted Utilization (SWU), which is

J.-Z. Wang · Z.-H. Yang · J.-L. Huang (✉)
Department of Computer Science, National Chiao Tung University, Hsinchu, Taiwan, Republic of China
e-mail: jlhuang@cs.nctu.edu.tw

J.-Z. Wang
e-mail: jzwang@cs.nctu.edu.tw

Z.-H. Yang
e-mail: radiohead0401@gmail.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_7, © Springer Science+Business Media Dordrecht 2014

the upper bound of utility, was incorporated into frequent sequential pattern mining algorithms to facilitate high utility sequential pattern mining. A sequence is regarded as a potential high utility sequence (PHUS) if its SWU is no less than the minimum utility threshold. Generally, SWU-based algorithms find PHUSs and then identify high utility sequences from PHUSs. Thus, high utility sequences can be found by this kind of two-phase algorithm.

However, since SWU usually greatly overestimates the utility of sequences, SWU-based algorithms [1–3] will spend much time to filter out low utility sequences from PHUSs, thus prolonging the execution time. In view of this, we propose two quantitative metrics, say Reduced Sequence Utility (RSU) and Projected-Database Utility (PDU), to prune low utility sequences. With the aid of these two metrics, we then devise algorithm HUS-Span to efficiently mine high utility sequential patterns. To evaluate the performance of algorithm HUS-Span, experiments are conducted on both synthetic and real datasets and results show that algorithm HUS-Span can identify high utility sequences more efficiently than existing algorithms.

The rest of this paper is organized as follows. Section 7.2 introduces the problem definition. Section 7.3 presents the HUS-Span algorithm, while Sect. 7.4 shows the experimental results. Finally, Sect. 7.5 concludes this paper.

7.2 Preliminaries

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of distinct items. Each item $i_j \in$ I is associated with an external utility, denoted as p(i_j) (e.g. unit profit of i_j). A quantitative item (q-item) is a 2-tuple (i, q), where $i \in I$ and q is the internal utility of i (e.g. quantity of i). A quantitative itemset (q-itemset) is a set of q-items $[(i_{b_1}, q_1), (i_{b_2}, q_2), \ldots, (i_{b_m}, q_m)]$. Without loss of generality, each q-item in a q-itemset is listed in alphabetical order. A quantitative sequence (q-sequence) is an ordered list of q-itemsets $< t_1 t_2 \ldots t_k >$, where $t_p (1 \leq p \leq k)$ is a q-itemset. A quantitative sequence database (q-SDB) is a set of tuples (sid, S), where S is a q-sequence and sid is the sequence ID of S. In this paper, we use the prefix "q-" to represent the item, itemset, or sequences with quantity. Given two q-sequences t and t', t is called a q-subsequence of t', denoted as $t \sqsubseteq t'$, if in t', there exists at least one ordered list of q-itemsets which equals to t. The size and the length of a sequence/q-sequence are the numbers of itemsets and items in the sequence/q-sequence, respectively.

Definition 1 The utility of a q-item (i, q), called *q-item* utility, is defined by $u(i, q) = p(i) \times q$.

Definition 2 The utility of a q-itemset $l = [(i_1, q_1)(i_2, q_2)\ldots(i_n, q_n)]$, called q-itemset utility, is defined as $u(l) = \sum_{k=1}^{n} u(i_k, q_k)$.

Definition 3 The utility of a q-sequence s $= <\mathrm{l}_1\mathrm{l}_2\ldots\mathrm{l}_\mathrm{m}>$, denoted as *u(s)*, is defined by $u(s) = \sum_{k=1}^{m} u(l_k)$.

Definition 4 The utility of a q-SDB DB, denoted as u(DB), is defined by $u(DB) = \sum_{\forall S \in DB} u(s)$.

Definition 5 Given a sequence $s = <e_1 e_2 \ldots e_n>$ and a q-sequence $t = <(l_1, q_1)(l_2, q_2)\ldots(l_n, q_n)>$, s matches t iff $e_k = l_k$ for $1 \leq k \leq n$, denoted as $s \sim q$.

Definition 6 The utility of a sequence, *s,* in a q-sequence *t,* denoted as $v(s, t)$, is defined as $v(s, t) = \{u(s') | \ s \sim s' \wedge s' \sqsubseteq t\}$. The utility of a sequence, *s,* in a q-SDB DB, denoted as *v(s),* is defined by $v(s) = \{v(s,t) | t \in DB\}$.

Definition 7 The maximum utility of a sequence *s,* denoted as $u_{\max}(s)$, is defined by $\Sigma \max\{v(s,t) | \ v(s, t) \in v(s) \wedge s \sim s' \wedge s' \sqsubseteq t \wedge t \in DB\}$. Sequence s is called a high utility sequence in a q-SDB if $u_{\max}(s) \geq \xi$, where ξ is a user-specified minimum utility.

Problem Statement. Given a q-SDB *DB* and the threshold parameter ξ, high utility sequential pattern mining is to identify the complete set of high utility sequences in *DB*.

7.3 The Proposed Algorithm

7.3.1 Overview

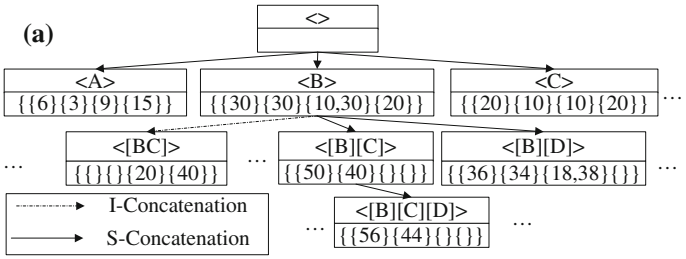
Based on the LQS-Tree[1] used in [2], the proposed HUS-Span algorithm traverses the tree in a depth-first way. Figure 7.1a shows the LQS-Tree of q-SDB *DB* in Table 7.1a. In the LQS-tree, each node *t* consists of its sequence and its utilities $v(t)$. The root is an empty sequence, and HUS-Span algorithm begins the mining process from the root. Algorithm 1 shows the procedure of HUS-Span when visiting a node *s*. HUS-Span first checks whether or not the utility of the projected q-SDB with prefix *s* (*s*-PDB) is less than ξ by the PDU Strategy[2] (Line 1). If so, HUS-Span backtracks to the parent of *s*; otherwise, it scans *s*-PDB in Line 3, to put items into *ilist* or *slist* (Lines 4–5). Then, HUS-Span removes those items unable to form high utility sequences from *ilist* to *slist* by the RSU strategy[3] (Line 6). Then, for each item *i* in *ilist*, HUS-Span performs the I-Concatenation operation to generate child node s' (Line 8), outputs high utility sequences if s' is (Lines 9–10), and calls itself recursively to visit node s' (Line 11). For each item *i* in *slist*, HUS-Span performs the same procedures as that for each item in *ilist* (Lines 12–16).

Similar to [2], each q-sequence in D is encoded using a utility matrix and each matrix is stored in memory. In a utility matrix of q-sequence *s*, each element e_{ij} is a

[1] Due to the page limit, readers can refer to [2] for more details about LQS-Tree.

[2] The details of PDU will be described in Sect. 7.3.3.

[3] The details of RSU will be described in Sect. 7.3.2.



(b)

Item	q-itemset1	q-itemset 2	q-itemset 3	q-itemset 4
A	(0,75)	(0,55)	(0,47)	(9,8)
B	(10,65)	(0,55)	(30,17)	(0,8)
C	(10,55)	(0,55)	(0,17)	(0,8)
D	(0,55)	(8,47)	(0,17)	(8,0)

Fig. 7.1 The LQS-Tree and the Utility Matrix. **a** The LQS-tree for *Q*-SDB. **b** The Utility Matrix of *Q*-sequence S_3 in Table 7.1a

tuple of *q*-item i in the jth q-itemset in s, which is defined by $(u(i, q), u_{rest}(i, s))$, where $u(i, q)$ is the *q*-item utility of i and $u_{rest}(i, s)$ is the utility of remaining items after i in s. $u_{rest}(i, s)$ is used to facilitate the proposed pruning strategies. Figure 7.1b shows the utility matrix of S_3 in Table 7.1a. For a node t in LQS-Tree, the utility of any t' children can be calculated by the utility $v(t)$, and the positions of the last q-items of a q-subsequence matching t in each utility matrix, called ending positions. The pivot of t in q-sequence S_k is the first ending position in S_k' utility matrix. For more details about the concatenation operation and how to calculate a node's children utility, readers can refer to [2].

Algorithm 1: HUS-Span(s, $v(s)$)

Input: A q-SDB D, the miminum utility threshold ξ, and a sequence s with utility $v(s)$.
Output: The set of high utlity sequences.

```
1  if PDU(s) < ξ then
2      return
3  scan s-PDB once to:
4      1. put I-concatenation items into ilist
5      2. put S-concatenation items into slist
6  delete low RSU items from ilist and slist
7  for each item i ∈ ilist do
8      (s', v(s'))←I-Concatenation(s, i)
9      if u_max(s') ≥ ξ then
10         output s'
11     HUS-Span(s', v(s'))
12 for each item i∈slist do
13     (s', v(s'))←S-Concatenation(s, i)
14     if u_max(s') ≥ ξ then
15         output s'
16     HUS-Span(s', v(s'))
```

Table 7.1 A quantitative sequence database and its profit table

(a) A quantitative sequence database

Sid	*q*-sequence
S_1	<[(A, 2), (B, 3)][(C, 4)][(D, 3)]>
S_2	<[(B, 3)][(A, 1), (C, 2)][(E, 1)][(D, 2)]>
S_3	<[(B, 1), (C, 2)][(D, 4)][(B, 3)][(A, 3), (D, 4)]>
S_4	<[(A, 5)][(B, 2), (C, 4), (D, 1)]>

(b) A profit table

Item	A	B	C	D	E
External utility	3	10	5	2	1

Since the utility does not hold the downward closure property, SWU was used for mining high utility sequences for holding the property. Here, we give the definition of SWU of a sequence t.

Definition 8 SWU of a sequence t in DB, denoted as SWU (t), is defined by SWU (t) $= \Sigma_{t \sim s' \wedge s' \sqsubseteq s \wedge s \in D}\ u(S)$.

For example, SWU(([B][D])) $= u(S_1) + u(S_2) + u(S_3) = 62 + 47 + 75 = 184$.

Theorem 1 (Sequence Weighted Downward Closure Property [1, 2]) *Given a sequence t in q-SDB DB, SWU(t) $\geq$ SWU(t') for all $t \sqsubseteq t'$.*

Based on Theorem 1, most of the existing algorithms [1–3] adopted *SWU* to avoid generating a sequence t if *SWU* (t) is less than ξ. However, SWU usually suffers from the problem of overestimating the utility of a sequence too conservatively. Given a sequence p, in p-PDB, those q-items before the first q-item of the first q-subsequence matching p in each original q-sequence, called unprojected q-items, are unable to form p-PDB. However, for a sequence t formed by I-or S-Concatenation (p, i), the utilities of those unprojected q-items, are added to *SWU* (t). For example, in <[C]>-PDB, ([C][D]) matches q-subsequences of S_1, S_2, and S_3. In S_1, the utilities of unprojected q-items [A, 2] and [C, 3] are added to *SWU* ([C][D])) though the two unprojected q-items cannot form any sequences in <[C]>-PDB. This problem becomes worse especially when a q-SDB consists of a large amount of long q-sequences. Such a problem motivates us to design two metrics, RSU and PDU, to obtain the estimated utility closer than SWU and the cumulated utility used in [2], thereby leading to better mining efficiency.

7.3.2 The Reduced Sequence Utility Strategy

To address the problem of SWU, we design another utility metrics called the Reduced Sequence Utility (RSU) to replace SWU. RSU(t, p-PDB) of a sequence t is the sum of utilities of projected q-sequences containing q-subsequences matching t in p-PDB, and p is a prefix of t.

Definition 9 Given a q-SDB DB, The RSU of a sequence t in p-PDB, denoted as RSU (t, p-PDB), is defined by:

$$\text{RSU}\ (t, p-\text{PDB}) = \sum_{\left(t \sim s_p \wedge s_p \sqsubseteq s \wedge s \in DB\right) \wedge \left(p \sim s' \wedge s' \sqsubseteq s_p\right)\ \wedge (i\ \in\ s')\ \wedge\ p \text{is a prefix of } t} \left(u(s') + u_{rest}(i, s)\right), \tag{7.1}$$

where i is the pivot of p in s.

Example 1 In <[C]> -PDB, since <[C][A]> matches only one subsequence of S_3, SWU(<[C][A]>) is $u(S_3) = 75$ whereseas RSU(<[C][A]>)=$u(C,\ 2) + u_{rest}(C1,\ S_3) = 10 + 55 = 65$. We can see that the utility of the unprojected q-item (B, 1) before $C1$ is contributed to SWU(<[C][A]>). In contrast, (B, 1) is not added to RSU(<[C][A]>).

Theorem 2 *Given a q-SDB DB, and two sequence t and t′, where* $t \sqsubseteq t'$ *and p be the prefix of both t and t′. Then, RSU(t, p − PDB)* $\geq$ *RSU (t′, p − PDB).*

For a k-sequence t, a new item i concatenates to t and results in a (k + 1)-sequence t'. If RSU(t') $\geq \xi$, item i is called a high RSU item; otherwise, item i is called a low RSU item. Based on Theorem 2, we can see that if $RSU(t,\ p - PDB) < \xi$, t and its supersequences, whose prefix is p, cannot form high utility sequences. And most importantly, HUS-Span algorithm can prevent from concatenating low RSU but high SWU items to a sequence in some cases that algorithms adopting SWU such as USpan [2] cannot do this. For example, given $\xi = 68$. In <[C]>-PDB, for those algorithms adopting SWU, [A] will be concatenated to ([C]) to form ([C][A]) for SWU(([C][A])) = 75. In contrast, HUS-Span algorithm will not concatenate [A] for RSU $(([C][A])) = 65$.

7.3.3 The Projected-Database Utility Strategy

The cumulative utility of node t and t's offsprings, adopted by the depth pruning strategy in [2], is to check whether or not node t can generate any high utility sequences. It is defined as the sum of utilities of S_k's first q-subsequence s' matching t with all the remaining q-items after s', for all $S_k \in DB$, $t \sim s' \sqsubseteq S_k$. If it is less than ξ, t is unable to generate any high utility sequences. However, for those first q-subsequences matching t without any remaining q-items, their utilities are still added to the cumulative utility of node t and t's offsprings. Given $\xi = 68$, consider node $<[B][C][D]>$, since no q-items exist after the q-subsequences matching $<[B][C][D]>$, the sum of utilities of <[B][C][D]>'s offsprings is 0. The cumulative utility of node $<[B][C][D]>$ and its offsprings, which is $56 + 44 = 100$, is only contributed by the q-subsequences matching ([B][C][D]). Since it is greater than ξ, USpan [2] will scan $<[B][C][D]>$ -PDB, find that no q-items exist in $<[B][C][D]>$ -PDB, and backtrack to the search procedure. To

address the problem, we design another utility metrics, called the Projected-Database Utility (PDU) of a sequence, to replace the metrics used in [2].

Definition 10 Given a q-SDB DB, the utility of the projected q-SDB with prefix p, denoted as PDU (p), is defined by:

$$\mathrm{PDU}(p) = \sum\nolimits_{i \in s' \wedge s' \sim p \wedge s' \sqsubseteq s \wedge s \in DB \wedge u_{rest}(i,s) > 0} (u(s') + u_{rest}(i,s)) \tag{7.2}$$

where i is the pivot of prefix p in s.

Example 2 ([D])-PDB consists of only one q-sequence, which is the q-subsequence <[(D, 4)][(B, 3)][(A, 3), (D, 4)]> of S_3. With the assistance of the utility matrix, PDU(<[D]>) can be easily calculated by adding $u(D2, 4)$ to u_{rest} ($D2$, S_3), which equals to $8 + 47 = 52$. The cumulated utility, which is used in USpan [2], of node <[B][C][D]> and its offsprings is 100. On the other hand, in <[B][C][D]> -PDB, since there are no q-items in it, PDU(<[B][C][D]>) is 0.

Theorem 3 *Given a sequence p and D, if PDU(p) < ξ, it is unnecessary to concatenate any items after each pivot of p in p-PDB.*

Based on Theorem 3, if PDU (t) of a sequence t is less than ξ, HUS-Span can immediately backtrack to the search procedure in some cases that USpan [2], using the cumulated utility, cannot achieve this. In implementation, with the assistance of the utility matrixes and the LQS-tree, to calculate PDU(p), it is unnecessary to construct p-PDB in advance. HUS-Span just needs to check whether the remaining utility of q-items after the pivot of p in each utility matrix is greater than 0. If so, the remaining utility and the utility of pivot are added to PDU(p).

7.4 Performance Evaluation

In this section, the experimental results of the performance of HUS-Span in comparison with USpan [2] are reported. A synthetic dataset and a real dataset are used for experiments. The synthetic dataset is generated by the IBM data generator with parameter D10kC10T2.5S4I2.5N1k. It means that the number of sequences is 10 k, the average number of itemsets per sequence is 10, the average number of items per itemset is 2.5, the average length of maximal sequences is 4, the average length of itemsets in maximal sequences is 2.5, and the number of distinct items is 1 k. The real dataset, called BMS-WebView-1 [4], consists of 59,602 sequences of clickstream data from an e-commerce web site, and contains 497 distinct items. In the two datasets, the internal utility is randomly generated and its range is between 1 and 10. The external utility is randomly generated using a log-normal distribution with parameter (μ, σ^2) = (0, 1.5).

Figure 7.2 shows the execution time of USpan and HUS-Span algorithm. The execution time spent by HUS-Span is shorter than USpan in both the synthetic and real datasets. The reason is that strategy PDU and strategy RSU can filter out more

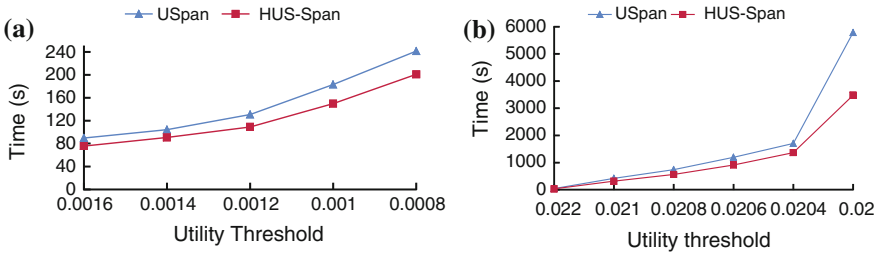


Fig. 7.2 Evaluation of execution time. **a** D10kC10T2.5S4I2.5N1k, **b** BMS-WebView-1

low utility sequences than the strategies used in USpan algorithm, thus improving the performance of HUS-Span algorithm. We can also see that the execution time of both algorithms is increasing when the utility threshold is more and more small. When the utility threshold is small, HUS-Span still outperforms USpan especially in the real dataset.

7.5 Conclusions and Future Work

In this paper, we proposed two pruning strategies, RSU and PDU, to estimate sequence utility. In addition, algorithm HUS-Span, was also proposed to prune low utility sequences with the aid of RSU and PDU. Experimental results showed that algorithm HUS-Span outperforms the prior algorithms such as USpan in both the synthetic and real datasets. For future work, since the utility matrix consumes more memory space especially when the sequence is sparse, we will also try to design an efficient data structure for massive data (e.g. long sparse sequences or a large number of sequences).

References

1. Ahmed CF, Tanbeer SK, Jeong BS (2010) A novel approach for mining high-utility sequential patterns in sequence databases. ETRI J 32(5):676–686
2. Yin J, Zheng Z, Cao L (2012) Uspan: an efficient algorithm for mining high utility sequential patterns. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 660–668
3. Shie BE, Hsiao HF, Tseng VS, Yu PS (2011) Mining high utility mobile sequential patterns in mobile commerce environments. In: Proceedings of the 16th international conference on Database systems for advanced applications, vol 1. pp 224–238
4. Kohavi R, Brodley CE, Frasca B, Mason L, Zheng Z (2000) Kdd-cup 2000 organizers' report: peeling the onion. SIGKDD Explor Newsl 2(2):86–93

Chapter 8
Simulated Annealing for the Design of Location Management System and Relation Map

Si-Hwan Jang, Sung-Soo Kim and Sanghyun Joo

Abstract Location area planning problem means partitioning the mobile network into location areas for minimizing the total cost and many partitioning problem is a NP-hard. In this paper, we use the simulated annealing with a new solution representation to solve NP-hard problem. In our method, we can automatically generate different number of location areas using compact index to obtain the optimal partitions. We compare the results obtained in our method with the existing results available in literature. We show that our methodology is able to perform better than previous methods.

Keywords Location management · Simulated annealing · Location area planning · Compact index

8.1 Introduction

The rapid growth in micro-cellular communication networks has generated a lot of interest among researchers to provide a good quality of service with minimum cost. Location area management is one of the important research fields. One of the

S.-H. Jang (✉) · S. Joo
Next-Generation Visual Computing Research Section, Visual Content Research Department, Creative Content Research Laboratory, Electronics and Telecommunications Research Institute (ETRI), 218 Gajeong-ro,Yuseong-gu, Daejeon 305-700, Republic of Korea
e-mail: jjangshan@etri.re.kr

S. Joo
e-mail: joos@etri.re.kr

S.-S. Kim
Department of Industrial Engineering, Kangwon National University, 192-1, Hyoja 2 Dong, Chuncheon-si 200-701, Republic of Korea
e-mail: kimss@kangwon.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_8, © Springer Science+Business Media Dordrecht 2014

strategies for providing a good quality of service at minimum cost is to partition the network into Location Areas (LA). The minimum cost has two components namely location update cost and paging cost. Location update cost is incurred when the user changes itself from one location area to another network. The paging cost incurred when a call arrives, the search is done only in the location area to find the user [1, 2].

Because of this difficult combinatorial complexity namely NP-hard, many researchers have applied heuristic methods such as genetic algorithms, ant colony optimization, neural networks, taboo search and other methods to get optimal value [1–4].

In this paper, we consider the location area planning problem for minimizing the sum of location update cost and paging cost, by using a simulated annealing. Simulated annealing has been used to solve optimization problems likewise NP-hard [5–7]. In earlier studies, first, the number of location areas is fixed and then, the cells are assigned to the location areas. In proposed method, the initial and neighbor solutions are automatically generated with different number of location areas, and the cells are assigned to the corresponding number of location areas.

In our study, the initial solutions are generated using the genetic algorithm method given in [2]. It is known that the initial solution and neighbor solutions are very important for better performance of simulated annealing. For obtaining good initial solution and neighbor solutions in our simulated annealing study, we use the idea of compact index (CI) given in [3]. The simulated annealing with this good initial solution method was tested on the network examples given in [1, 2]. We are able to obtain better results than the results reported in [1, 2].

8.2 Location Area Planning

In mobility management systems, one of the ways of reducing the total cost is to partition the network into Location Areas. In this process, when a mobile terminal enters a new LA, it updates its location, and as long as it travels within the LA, it never updates its location. On the other hand, whenever there is an incoming call, and since the network does not know the exact location of the user in the current LA, it pages the user in all the cells of the last updated LA [2]. In the location area planning problem, the total cost is given as

$$\text{Cost} = \beta \times N_{LU} + N_P \tag{8.1}$$

In the above equation, N_{LU} is the total number of location updates and N_P is the total number of paging operations. β is a constant and it represents the ratio of a location update to a paging in the network. In our experiment, we have used $\beta = 10$. In a given network, the number of location updates is caused because the user movement and the paging operation are related to incoming calls.

8.2.1 Solution Representation

In our approach, the solution is represented as a string of length n, the number of cells in the network. Each cell is assigned another number that represents the location area the cell belongs to. Consider the following 4 × 4 network.

In the above Table 8.1, the first row, gives the cell number (CN). The second and third row gives two location area configurations LA_1 and LA_2 respectively.

The above each configuration consists of 3 location areas. The compact index is defined as the ratio of number of adjacent boundaries between different location areas and the total number of boundaries in the given network. We can see from Fig. 8.1 that LA_2 is better configuration than LA_1 in terms of compact index.

We now have two configurations LA_1 and LA_2. The question now is which configuration (LA_1 or LA_2), we will use in our simulated annealing as a starting solution. In order to get the good starting solution, we use the compact index.

The compact index for LA_1 is $\frac{16}{33} = 0.4848$ and compact index for LA_2 is $\frac{11}{33} = 0.3333$. We choose the configuration with minimum compact index value. The usage of compact index is used in Simulated Annealing for obtaining the starting solution and also obtaining neighborhood solutions.

The flowchart of Simulated Annealing with compact index and several notations are shown in Fig. 8.2.

8.3 Experiment Results and Discussions

8.3.1 Experiment Results

In our experiment, we have used 5 × 5, 7 × 7 and 9 × 11 networks considered in [1, 2]. In the simulated annealing, the values of parameters used are: T (initial temperature) = 1, t (number of neighborhood search with the same temperature) = 10 and $\alpha = 0.001$ and the experiment was run on Pentium 4, computer (3.4 GHz, 2G RAM).

Each of the above network examples was iterated 50 times using simulated annealing with good initial solution using compact index and present the optimal result obtained in the runs. This is because it is well known that behavior of simulated annealing is stochastic. The results obtained are given below.

5 × 5 network: We are able to obtain the same configuration obtained in [1, 2]. The number of location areas is 3 and the total cost obtained is 26,990. The compact index for this optimal configuration is 0.2143. The computation time to obtain this optimal configuration is 1.8 s and the solution converges after 800 generations (Fig. 8.3).

7 × 7 network: The optimal configuration obtained for this network is shown in Fig. 8.4. The same configuration is obtained in [2] using GA-HNN3. The number

Table 8.1 Example of 4 × 4 network

CN	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
LA_1	1	2	2	3	1	2	2	3	1	2	2	3	1	1	1	3
LA_2	1	1	2	2	1	1	2	2	3	3	3	2	3	3	3	2

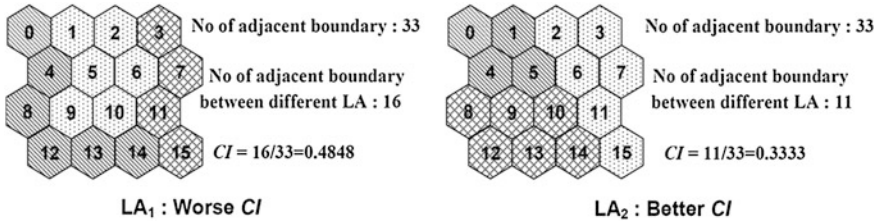


Fig. 8.1 Configurations of LA_1 and LA_2

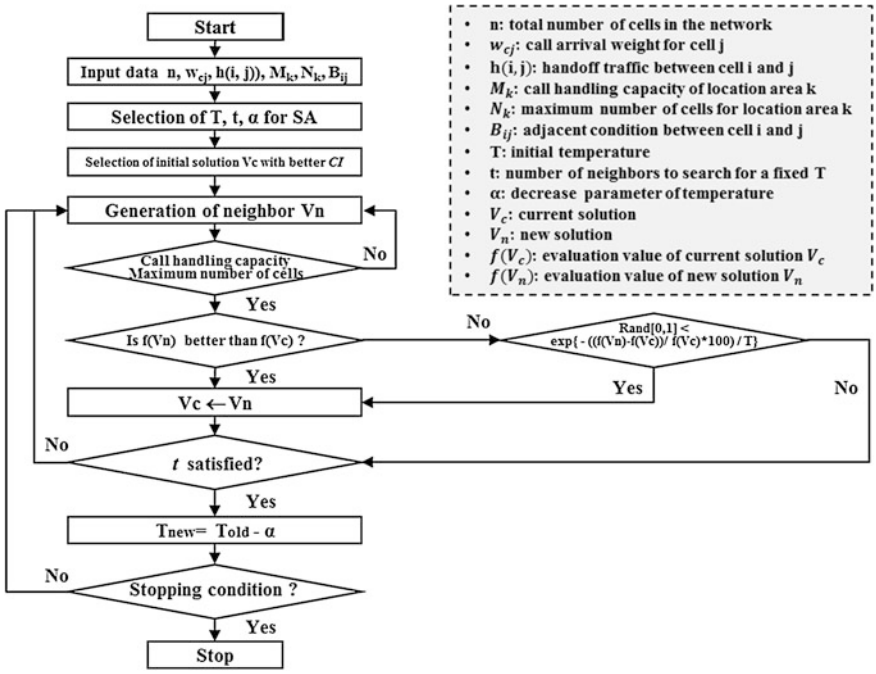


Fig. 8.2 Flowchart of simulated annealing with compact index

of location areas is 8 and the total cost obtained is 60,606. The computation time to obtain this optimal configuration is 5 s and the solution converges after 800 generations. The compact index for this optimal configuration is 0.2857.

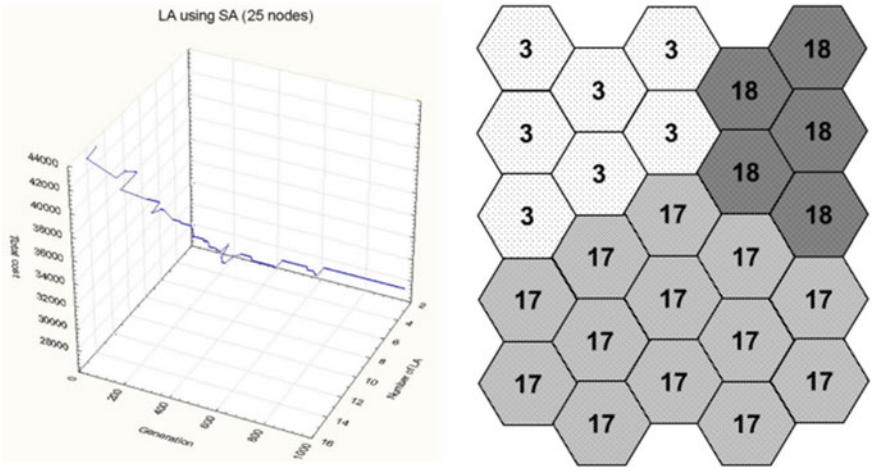


Fig. 8.3 Convergence for the optimal solution in three dimension and optimal configuration obtained for 5 × 5 network

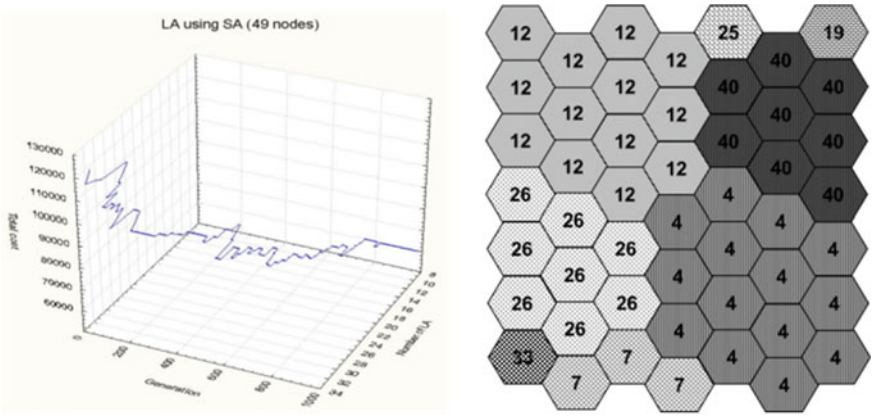


Fig. 8.4 Convergence for the optimal solution in three dimension and optimal configuration obtained for 7 × 7 network

9 × 11 network: The optimal configuration obtained for this network is shown in Fig. 8.5. The number of location areas is 11 and the total cost obtained is 169,273. The computation time to obtain this optimal configuration is 17 s and the solution converges after 1,000 generations. The compact index for this optimal configuration is 0.3023. For the same network results obtained in [1] has 10 location areas and the total cost is 172,669. The number generation is around 1,500 in [1]. From these networks, we can see that our method performs well.

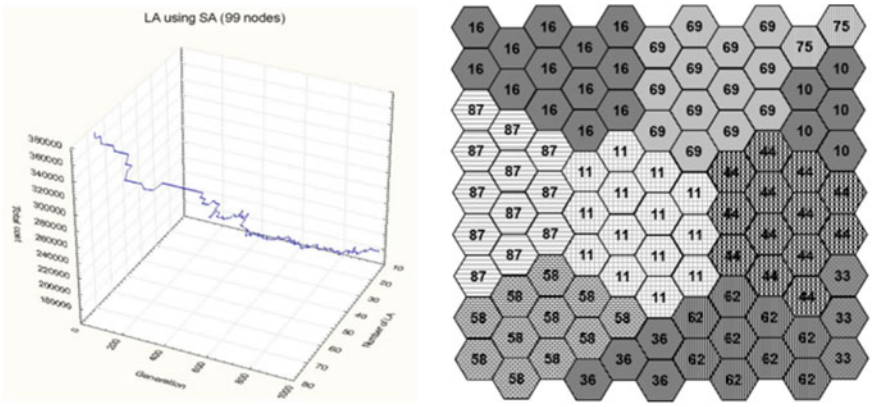


Fig. 8.5 Convergence for the optimal solution in three dimension and optimal configuration obtained for 9 × 11 network

8.3.2 Discussions

In our approach, we partition the given network into optimal number of location areas using the concept of compact index. The compact index varies from 0 to 1. The optimality criterion used in our study is minimizing the total cost. In proposed approach, the different numbers of location areas are automatically generated. The three dimensional representation shows the convergence to the optimal solution with different number of location areas. The only additional computation needed in our approach is the selection of initial solution using of compact index. This additional computation is negligible with respect to that of simulated annealing methodology.

8.4 Conclusions

We have considered the location area planning problem and present simulated annealing method with compact index for obtaining the solution. A description of compact index and its use in obtaining the initial solution and generation of neighborhood solution with different number of location areas in simulated annealing is presented. The test networks considered in earlier studies were used for verification of our approach. A comparison of our approach with existing methods is presented. It is shown that our method is able to obtain better location areas for the test networks, than previous ones, in terms of the cost. In future, we will apply this method to the relation map.

Acknowledgments This work was supported by the Industrial Strategic Technology Development Program, (10044312, Developments of Visual Communication Platform and Authoring Tool for Emotional Expression) funded by the Ministry of Science, ICT and Future Planning (MSIP, Korea).

References

1. Taheri J, Zomaya AY (2007) A simulated annealing approach for mobile location management. Comput Commun 30:714–730
2. Taheri J, Zomaya AY (2007) A combined genetic-neural algorithm for mobility management. J Math Model Algorithm 6:481–507
3. Kim SS, Smith AE, Hong SJ (2008) Dynamic load balancing using an ant colony approach in micro-cellular mobile communications system. In: Michalewicz Z, Siarry P (eds) Advances in metaheuristics for hard optimization. Springer, Berlin
4. Pierre S, Hou′eto F (2002) Assigning cells to switches in cellular mobile networks using taboo search. IEEE Trans Syst Man Cybern Part B: Cybern 32(3):351–356
5. Chen Y-W, Kohatsu T (2007) 3D image reconstruction from limited projections by simulated annealing. The 2nd international conference on innovative computing, Information and Control, Kumamoto, Japan, 2007
6. Kroumou V, Yu J, Shibayama K (2010) 3D path planning for mobile robot using simulated annealing neural networks. Int J Innovative Comput Inf Control 6(7):2885–2899
7. Kim G, Kim SS, Kim I-H, Mani V, Moon J-K (2011) An efficient simulated annealing with a valid solution mechanism for TDMA broadcast scheduling problem. Int J Innovative Comput Inf Control 7(3):1181–1191

Chapter 9
A Simple, Flexible and Adaptable Solution to Enable Information Exchange for Specific Business Processes

Mert Biçakçi and Mehmet Ulaş Çakir

Abstract In the current information age, the variety of data sources that are used to generate meaningful information increases rapidly. Various information systems are developed to process this vast amount of information. Data providers with different technical infrastructures, data semantics and information taxonomies need to exchange data in order to collaborate for a common goal in the joint nature of information systems. Due to current circumstance a new challenge arises: Integration of decoupled systems to consolidate information to produce smart and qualified outputs which are tailored to needs. Unless this challenge is tackled, meaningful results cannot be achieved through a constantly growing set of data. In this paper a new solution is proposed to establish a common framework for facilitating interoperability among legacy and existing information systems using different protocols and to accomplish seamless and in-time integration.

Keywords Early warning systems · Information systems · Geospatial intelligence systems · Integration

9.1 Introduction

Using the same map has been the primary goal of the current and future information systems and interoperability is one of the main issues on the way to success. Despite recent standardization efforts, the interoperability studies realized between information providers and consumers aren't adequate. This is partly

M. Biçakçi (✉) · M. U. Çakir
STM A.Ş, Ankara, Turkey
e-mail: mbicakci@stm.com.tr

M. U. Çakir
e-mail: ucakir@stm.com.tr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_9, © Springer Science+Business Media Dordrecht 2014

because of legacy systems that should be integrated into newly emerging net-centric systems. Also interoperability standards are being defined in recent years. To enable interoperability, the current approach is to rapidly engineer custom-made interfaces between specific systems, with the consequent challenge of software maintenance upon later system upgrades. Without addressing the facilitation of interoperability, the burden and expense of constant reverse engineering of interfaces to ensure compatibility with legacy applications will remain.

An innovative solution to interoperability problem is to provide an integration system that can act as a data/information broker between heterogeneous systems, allowing a much greater degree of flexibility in the integration of legacy systems and supporting a more de-coupled information environment that still provides the necessary environment for the rich sharing of data.

Early warning systems are chosen as an example of information systems integration to represent the current drawbacks of standalone information systems and the benefit of integrating such systems. Current situation in early warning systems are very similar to the scenario that is suggested as a solution. There is no integration of early warning systems that covers all hazards in all places [1]. Also there is a variety of data sources from different systems including automated systems and systems that involve human interaction. Different consumers of these systems need different early warning message types and are interested in hazards that occur in different areas [2]. So there is a necessity to disseminate the early warning information using different protocols enhanced by filtering capabilities (i.e. area of interest, hazard type, etc.).

In this paper, a generic solution to integration of different information systems is described and early warning systems are given as an example of this proposal to solve a real life problem.

9.2 Proposed Solution

As it is stated, there are many information systems that provide, consume and exchange data. Most of these systems either is standalone or has a limited integration capability. The problems of standalone information services can be summarized as:

- Different data sources use different protocols like SMTP (email), XMPP (chat), Web Services, etc. Some data sources even have their own interface. Also some systems involve human interaction. Integration is costly and requires new deployment of software.
- End user systems consume information using different protocols but all information systems do not support variety types of protocols like SMTP (email), XMPP (chat), Web Services, etc.
- End user systems require filtering of provided data like area of interest, hazard type, etc. filtering.

Table 9.1 Data sources

Information system type	Data format
File system	XML files
	MS excel files
	MS access files
	Character delimited text files
ODBC	MS access files
	SQL server databases
	PostgreSQL databases
SMTP	SMTP file attachment (see file system for supported file data formats)
XMPP	XML messages through XMPP message body
Web service	Request/response web services that are based on simple object access protocol (SOAP)
SharePoint services	MOSS 2007 document libraries (see file system for supported file data formats)
Tweet	Web services (RESTful)
RSS feed	RSS
SMS	Messaging service
Pub\Sub	WS-N notification

- End user systems have to use specific software to receive information which can be costly. Instead to consume information, free tools like Google Earth can be used.

The solution proposed to solve above mentioned problems is called Information Integration Core (I^2Core). I^2Core is an integration system that accepts data using variety of protocols to receive data from existing information systems. I^2Core consumes data from new data sources by enabling easy configuration of the system for different kinds of data for rapidly changing environment without having to rebuild the I^2Core infrastructure so that application does not need to be re-deployed. I^2Core disseminates data in different protocols and in different formats. While consuming and disseminating data, I^2Core provides utility functionalities like security, service discovery, auditing, help, etc.

The main objective of the I^2Core is to provide simple, flexible and adaptable mechanisms to enable information exchange for specific business processes. The I^2Core has a robust Integration Middleware that enables dissemination of many information objects received from diverse sources, to consumers, based on the consumers' needs.

I^2Core can receive information through file systems, databases, e-mail, chat, web service, portals, tweets, RSS feeds, pub\sub notification. The variety of information systems that I^2Core supports gives the ability to receive information from many information systems using different protocols and supplying information in different formats. The supported information system types as data sources are summarized in Table 9.1.

Table 9.2 Consumer systems

Information system type	Data format
File system	Graphical files (GML, KML, SHP etc.), MS excel files, MS access files, Characted delimited text files
SMTP	SMTP file attachment
XMPP	XML messages through XMPP
Web service	Request/response web services that are based on simple object access protocol (SOAP).
Tweet	Web services (RESTful)
RSS feed	RSS
SMS	Messaging service
Pub\Sub	WS-N notification

While supporting many protocols and data formats, I^{2}Core also can be configured during run-time. This means that there is no need for development and deployment of a new I^{2}Core to support new information systems as data sources. This saves a big amount of resource and time as there is no new implementation cycle needed. If this was not the case, to support new information systems as data sources, whole lifecycle of system development should be experienced, i.e. requirements analysis, system design, implementation, integration and verification.

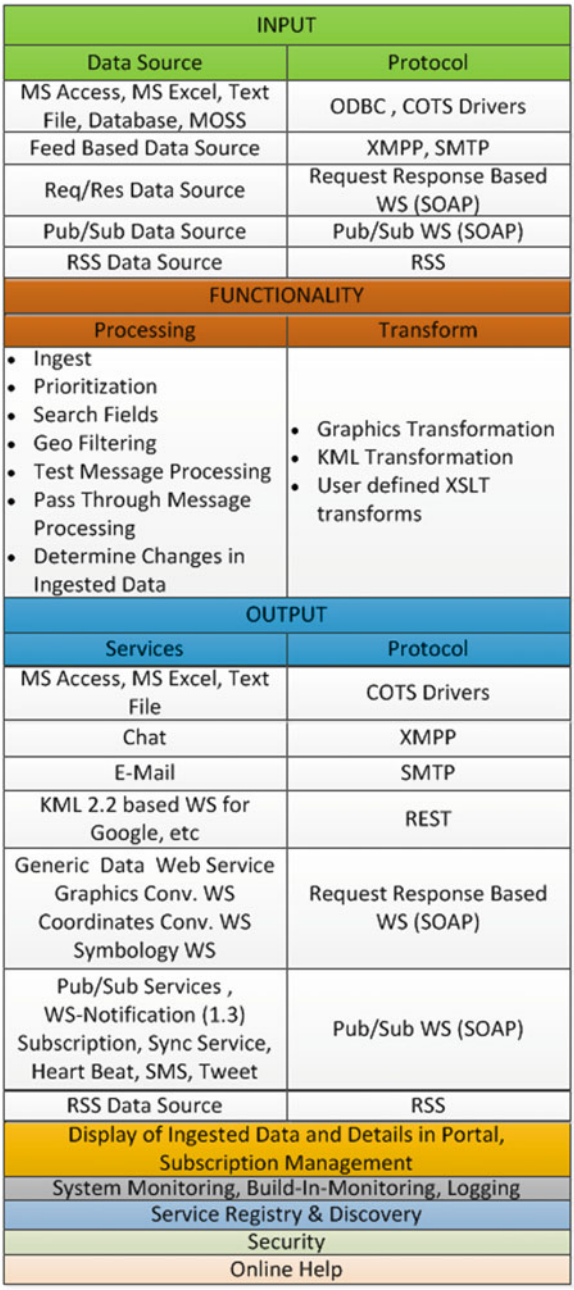
I^{2}Core can disseminate information to consumer systems by supporting different protocols like pub\sub, web services, e-mail, chat, tweet or RSS. This allows many consumer systems and end users to receive information from I^{2}Core without any need for special adapters. The supported information system types as consumers are summarized in Table 9.2.

The Fig. 9.1 shows the most fundamental features of the I^{2}Core.

I^{2}Core features are:

- Provision of information to consumers in a protocol they use (e-mail, chat, web services, portal, etc.),
- Provision of information to consumers in a format they use (graphics (e.g. KML), plain text, native, etc.),
- Provision of information to consumers,
- Optimizing bandwidth through a geographic filtering capability, by moving towards a push/event based architecture,
- Provision of service discovery,
- Provision of system administration, monitoring, error handling and error notification capabilities,
- Provision of service security.

Fig. 9.1 I²Core features



9.3 Concept of Execution

The I²Core is a SOA based system that can act as a broker between heterogeneous systems, allowing a much greater degree of flexibility in the deployment of systems and supporting a more de-coupled information environment that provides for the rich sharing of data. The basic data flow on I²Core is given in Fig. 9.2.

At the highest level, the functionality to be provided through the I²Core is; to ingest data from a variety of data sources and to transform that information into forms that can be consumed by other systems.

In order to achieve this, all necessary configurations are performed in the system through I²Core System Management Application. When data is to be ingested by I²Core, the following steps are performed:

- Data is ingested from data source systems by the adapters of I²Core.
- I²Core validates data against input validation schema by using XSD file which was previously configured. All validation errors are consolidated as a single validation error to be sent to the configured users for Error Notification Service.
- Data is transformed, de-batched, processed and stored in the I²Core database to make it ready for consumption.
- During data processing, I²Core decides:

– If the ingested data is a new data.
– If the ingested data is an update of a previously ingested data.
– If any previously ingested data is deleted from data source.
– If any of these conditions is met, I²Core produces notification and sends it to Publish/Subscribe component.
  - Either while sending data as a notification or when a consumer requests data from I²Core, data is transformed to an output format according to the preferences of the consumer.
  - Graphics Services; which implements graphics standards like GML, SHP, etc. and KML which can be consumed by Google Earth.
  - Generic Data Service; which provides all types of data in all supported formats with matching filtering criteria.
  - Publish/Subscribe Services; which either pushes notifications to notification consumers via Web Service, XMPP, SMTP, tweet, RSS or accumulates notifications in a pull-point which allows request/response type consumers to get notifications on demand.

9.3.1 Sample Execution of the Concept

For the demonstration of usage of I²Core, Early Warning Systems has been chosen as information provider and consumer systems because Early Warning Systems

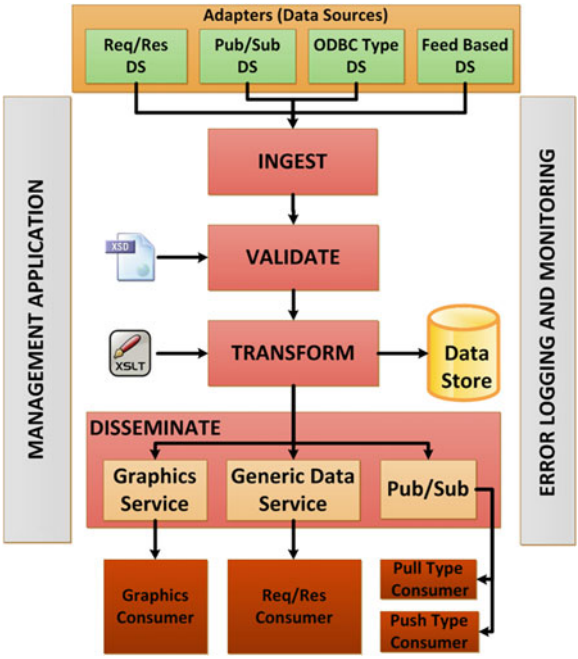


Fig. 9.2 I^{2}Core basic data flow

have all the difficulties that I^{2}Core aims to solve. Also an integrated early warning system is urgently needed [3]. With this demonstration example, in fact a solution to the early warning and geo-spatial intelligence systems integration problem is also provided.

First of all, there are many early warning systems used for different purposes all around the world. These systems work standalone without interacting with other early warning systems. There is no unique system that can receive information from all these variety of early warning systems.

Secondly, there are many consumers (either humans or other systems) of the information provided by the early warning systems. There is no unique early warning system that supports all the protocols and data formats that disseminate early warning information the way that all consumers want.

Furthermore, early warning systems generally require special software and/or hardware to work which are generally expensive. Client installation of system applications or hardware installations should be performed before in order to use these systems. In addition, some special training should be given to the end user before using them which takes time and resource.

Finally, some early warning systems require human interaction. Human interaction means sending and receiving information via e-mail, chat, etc. A few of them support human interaction with this concept.

Usage of I^{2}Core to integrate early warning systems tries to solve the above mentioned problems. First of all, I^{2}Core supports variety of protocols as data source information systems. The complete list of data source information system

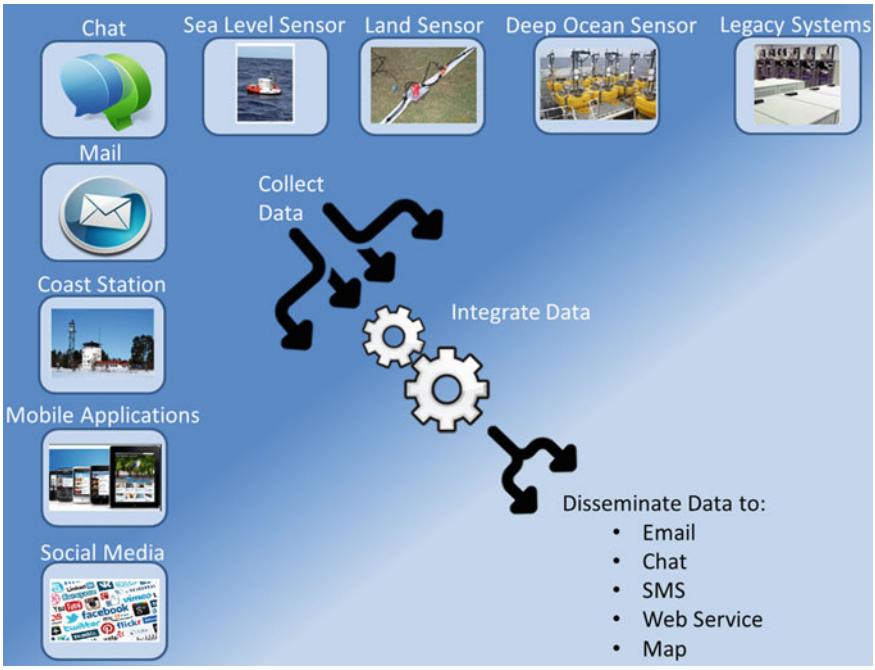


Fig. 9.3 I^{2}Core usage in early warming systems

types were given in Table 9.1. I^{2}Core also supports many data formats as data format can be provided on the fly without new deployment of the system as described in Sect. 9.3—Concept of Execution. Secondly, I^{2}Core supports many protocols for data dissemination. This helps I^{2}Core to support variety of consuming information systems.

Furthermore, I^{2}Core does not require special software and/or hardware to be installed. It just takes a person to do the configuration of an integration case. Afterwards, data source can provide data and consumers can receive data automatically once I^{2}Core is configured. Free tools like Google Earth can be used to show information gathered from I^{2}Core on a map. As a result of this, no specific training is needed to use I^{2}Core. Finally, I^{2}Core supports human interaction in both ways, i.e. it can both receive and send information via human interaction using chat, e-mail, social media applications, etc.

Early warning systems may receive data from many sources like sensors, legacy systems, and humans. Sensors may include sea level, land, deep ocean sensors. Humans may interact with I^{2}Core using chat, e-mail, tweet, etc. As seen in Fig. 9.3, I^{2}Core collects data from variety of data sources, integrates and

consolidates data coming from different source and disseminates data in different formats.

Let's consider the following scenario as an example:

An earthquake sensor in the Pacific Ocean senses an earthquake and via satellite links sends this information in CAP format [4] to I^{2}Core. A policeman sees a fire that is resulting from the earthquake and sends this information via e-mail from his smartphone. A citizen sends a tweet about his injury. I^{2}Core receives all this information and consolidates them. Fire department receives e-mail from I^{2}Core about the fire by filtering on the geographical area that they serve. Hospitals receive chat message about the new injury. Mayor of the city sees the overall damage resulting from the earthquake on the Google Earth. Also, early warning system integrated to I^{2}Core, receives all the information via web service and announces this information via cellular mobile infrastructure.

As seen from the above scenario, I^{2}Core plays an important role as an integration system between data source and consumer early warning systems and end users. Usage of I^{2}Core to integrate early warning systems will save time and resource.

9.4 Conclusions

I^{2}Core can be a solution to problems arising while integrating information systems. It has several advantages over developing custom integration solutions for each information system. The benefits of using such an integration system are:

- Information coming from different data source information systems (manual, automated, legacy, and new) with different protocols can be consolidated.
- Integrating legacy systems into new system with very low cost and requires no new deployment.
- Easily integrating data provider and consumer systems that utilizes different data providing solutions like Web Services, WS-Notification compliant pub/sub protocol, database, e-mail, chat, file system, etc.
- Validating incoming data, transforming data into I^{2}Core's internal data format and exposing data with standard or custom services.
- Disseminating data using geo-filtering capability which allows Consumers to receive data of the geographical areas they are interested in.
- Discovery of services.
- Utilizing security standards (WS-Security) for providing web service security functions.
- Auditing data coming in and out.

Also, a scenario for integration of early warning systems is used to represent the usage of I^{2}Core for solving a real life problem. I^{2}Core proves to be a good solution

for integration of early warning systems. This can be investigated further with the real implementation in another paper.

For future works, new abilities to I^{2}Core can be added. I^{2}Core can be improved to provide multimedia data (photo, audio, and video) to consuming systems and end users. Also some new and custom adapters to I^{2}Core can be added to support more data source information systems.

In this paper, a simple, flexible and adaptable solution to enable information exchange for specific business processes is described as an overview and early warning systems integration is given as an example. Detailed comparison of different integration systems can be subject of another paper.

References

1. Sorensen JH (2000) Hazard warning systems: review of 20 years of progress. Nat Hazards Rev 1(2):119–125
2. Lendholt M, Hammitzsch M (2012) Towards an integrated information logistics for multi hazard early warning systems. Open Environ Eng J 5:27–43
3. Ghosh JK, Bhattacharya D, Samadhiya NK, Boccardo P (2012) A generalized geo-hazard warning system. Nat Hazards 64(2):1273–1289
4. Botterell A (2006) The common alerting protocol: an open standard for alerting, warning and notification. In: Proceedings of the 3rd international ISCRAM conference, Newark, NJ

Chapter 10
User Authentication with Distributed Cluster Formation in Wireless Sensor Networks

Dong-Hoon Kim, Youn-Sik Hong and Ki Young Lee

Abstract Security is an important aspect of any system and its development has enhanced widespread adoption of WSNs. In a typical situation, user wants to access the data directly from a node instead of accessing at the gateway or at the backend of an application system. Therefore, it is necessary that such an access is allowed only to a registered user. We present a user authentication scheme with distributed cluster formation in WSNs. This scheme works as a ring structure and uses one-way communication. In our protocol, three distinct components, gateway, user, and node, can mutually authenticate each other. In addition, we use S/Key algorithm as one-time password for authentication. Our analysis shows that our proposed scheme is as robust as conventional 4-way authentication scheme, while the communication overhead is substantially decreased.

Keywords S/Key algorithm · 3-way chaining authentication · Wireless sensor networks · Cluster formation

D.-H. Kim · K. Y. Lee (✉)
Department of Information and Telecommunication Engineering, Incheon National University, Incheon, Korea
e-mail: kylee@incheon.ac.kr

D.-H. Kim
e-mail: dhkim85@incheon.ac.kr

Y.-S. Hong
Department of Computer Science and Engineering, Incheon National University, Incheon, Korea
e-mail: yshong@incheon.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_10, © Springer Science+Business Media Dordrecht 2014

10.1 Introduction

WSNs have been used in various real-time applications, such as vehicular tracking, habitat monitoring and so on. WSNs can be deployed in unattended environments. Each node is made of low-cost and battery-operated tiny sensing devices that are of limited computation and communication capabilities. The data sensed by each node is routed from a node to its one-hop neighbor in an ad-hoc manner and then finally to the gateway node (also known as *base station*) that has a wired connection with a backend of an application system. The data will be sent either at regular interval or upon event detection. In general, real-time data may no longer be accessed at the gateway or at the centralized application systems. Instead, they could be accessed directly from a node in a WSN. Thus, it is necessary that such an access is allowed only to a registered user. The registered user can login to the network and access data on a continuous basis or whenever is required.

Since sensor nodes have limited resources and computational capabilities, it is desirable for the user authentication scheme to be simple and efficient. To do this, we use a 3-way chaining user authentication schemes for WSNs which works as a ring structure and uses one-way communication. In this scheme, three distinct components, gateway, user, and node, can mutually authenticate each other. Also, we use S/Key algorithm as one-time password (OTP) for authentication in this scheme. At the core of our proposed algorithm is an OTP scheme organized between user and the gateway. The password generated by OTP is valid for only one log-in session. Every time a new log-in session is established, the password table for a dedicated user should be changed according to his (her) log-on data. Thus, our proposed scheme is robust with respect to replay attack and stolen-verifier. These processes use only cryptographic one-way hash functions and exclusive-OR. The computation involved in encrypting and decrypting with an OTP are considered relatively simple and can be achieved using bit-level operations natively provided by modern embedded processors.

During user authentication, heavy traffic may be occurred for exchanging messages among nodes. A sensor node for transmitting a message in a wireless channel consumes energies more than 10-times compared to the case of sleep mode. Thus reducing message overhead is one of the major system design concerns. However, our proposed scheme adopts only 3-way protocol. So it reduces message overheads significantly, and thus it saves overall energy consumption of sensor nodes.

10.2 Related Works

In past years, many user authentication schemes for WSNs have been proposed. Some schemes use password only, other schemes use two factors to authenticate user. Wong et al. [1] proposed a dynamic strong-password based solution for

WSNs which imposes very light computational load and requires simple operations, such as one-way hash function and exclusive-OR operations. More improved dynamic user authentication schemes are presented in [2, 3]. Some security flaws in the Wong et al.'s scheme, such as vulnerability to replay attack and passwords can easily be revealed by any of the sensor nodes, was identified by Tseng et al. [2]. By implementing improved scheme of Wong et al.'s scheme, they provide resistance to replay attack and reduction of password leakage risk. Vaidya et al. [3] pointed out that previous dynamic user authentication schemes do not provide mutual authentication between user and gateway node.

However, Das [4] pointed out that these protocols are vulnerable to many logged in users with the same login-id threat and stolen verifier attack. Das proposed a two-factor user authentication scheme for WSN using only one-way hash function. This scheme resists many logged in users with the same login identity, stolen-verifier, guessing, impersonation and replay attack. Arikumar and Thirumoorthy [5] and Vaidya et al. [6] point out some security pitfalls remain in Das's scheme and propose an improvement such as changing password and providing mutual authentication between gateway and sensor nodes in Das's scheme that provides better security. These improved schemes, however require extra packet transmission for mutual authentication and heavy traffic maybe occurred in the network. In this paper, we propose 3-way authentication scheme for WSNs which provide mutual authentication among the components, user, gateway, sensor node, as well as equivalent level of security efficiently.

10.3 A Position Estimation Using Relative Distance Fingerprinting

In this section, we describe our proposed user authentication with clustered formation scheme in more detail. Before we discuss it, the notations used throughout the paper are summarized as shown in Table 10.1.

Our scheme performs as follows: A user wants to access sensing data of a specific sensor node N. At first, the gateway has to authenticate a user in the Pass 1. To do this, a user sends both his/her identifier (ID) and the one time password (OTP) to the gateway. In the Pass 2, the sensor node has to authenticate the gateway using public key cryptography. For doing this, the gateway G encrypts a request message with a message authentication code (MAC) and the private key PRG and then sends it to the node N. In the Pass 3, only an authorized sensor node can decrypt the request message using the public key PU_G. After that, it has to authenticate the user U using public key cryptography. That is, the sensor node encrypts a response message with the newly generated MAC and the password sent by the gateway G and then sends it to the authenticated user. In a similar fashion, only a registered user can decrypt the message using the public key PU_N. At last, it can authenticate both the gateway and the sensor node.

Table 10.1 The list of notations

Notation	Description	Notation	Description
U	A user	PU_G	Public key of the gateway
G	The gateway of a WSN	PR_N	Private key of a sensor node
N	A sensor node	PU_N	Public key of a sensor node
ID	User identification	‖	Concatenation operator
PW	Password chosen by the user	=?	Verification operator
H()	One-way hash function	Tx	Current timestamp
OTP	One time password	ΔT	Time interval for packet transmission between passes
MAC	Message authentication code		
PW_G	Password stored on the gateway	E()	Encryption function
PR_G	Private key of the gateway	D()	Decryption function

10.3.1 Registration/Password Change Phase Using the S/Key Algorithm

Registration phase will be done when a user wants to register with the WSN in a secure manner. OTP is a kind of dynamic password that is valid for only one login session. Therefore, this scheme does not vulnerable to replay attack. If an attacker happens to obtain an OTP, he/she may log in for one time period, but it becomes useless once that period expires. Notice that the process of generating the *n*-sequenced set of passwords uses only simple hash function. So it does not consume much resource. To get the next password in the series with the previous one, he/she needs to find a way of calculating the inverse of the hash function H. Since H was chosen to be one-way, this is extremely difficult to do. The S/Key algorithm is one of the simplest methods because it requires only one-way hash function and performs exclusive-OR operations that are suitable for WSNs [7]. Figure 10.1 shows the entire process of the S/Key algorithm. The S/Key algorithm is divided into two phases: password generation phase and authentication phase.

In the password generation phase, a user generates his/her *n* OTPs by one-way hash function with PW or previous output of the hash function. After generation of the *n*-sequenced OTPs, both the user's ID and the *n*th OTP, the last generated PW, are to be submitted to the gateway. The gateway stores them as the current reference PW_G. Then the gateway sends the PU_N, the public key of sensor nodes, to the user for secure data transmission. Notice that as the size of *n*-sequenced OTPs increases, more resources and longer execution time should be needed. So determining the size of *n*-sequenced OTPs is a tradeoff between the resource constraints and the robustness of user authentication.

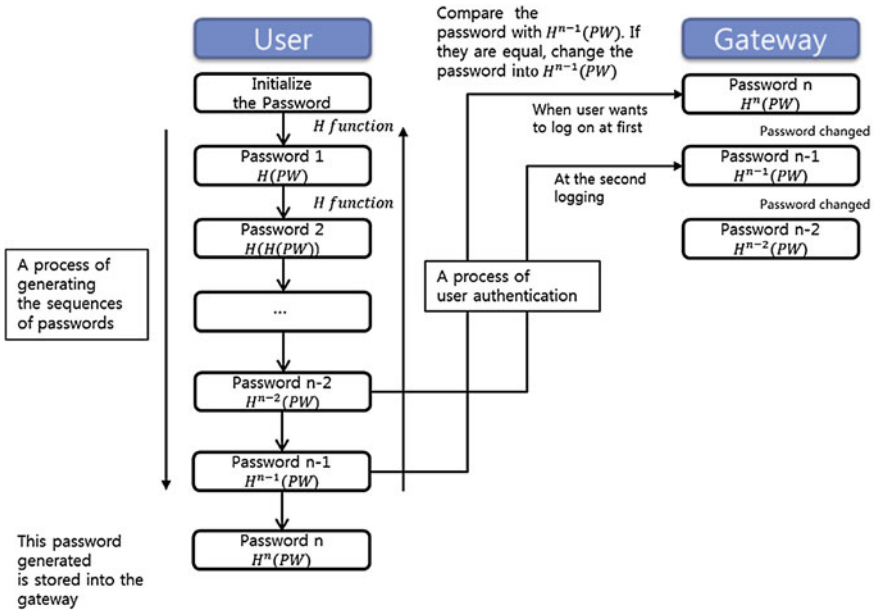


Fig. 10.1 The generation of the set of passwords using the S/Key algorithm

10.3.2 Pass 1: User Authentication by the Gateway

A user sends both ID and PW, the *i*th OTP $H^i(PW)$, to the gateway. At first, the gateway verifies the user ID. After that, it computes a new $PW_G{}^*$ by using H(OTP) and then compares it with the PW_G which is stored in the gateway during the registration phase. If $PW_G{}^*$ is equal to PW_G, then the MAC is created by using the function $H(PW_G||T_1)$. To create a secure request message, the PW_G and MAC, are encrypted using the private key of the gateway PR_G. At the end of this pass, the current PW_G should be changed to the current OTP, $H^i(PW)$ for the purpose of password verification. Each step in the user authentication by the gateway is described as follows:

- Verify the user's ID
- Compute $PW_G^* = H(OTP)$
- Check to see if $PW_G^* = ?PW_G$
- If it is not true, terminate the process and send rejection message to user
- Otherwise, computes the MAC: $MAC = H(PW_G||T)$
- Create an encrypted request message: $E = E_{PR_G}(PW_G||MAC)$
- Send $[E, T_1]$to the nearest sensor node
- Change the current reference: $PW_G = OTP$.

Figure 10.2 depicts the parameter to be transmitted at each pass. For example, in the pass 1, a user ***U*** transmits both ID and OTP to the gateway ***G***, whereas in the

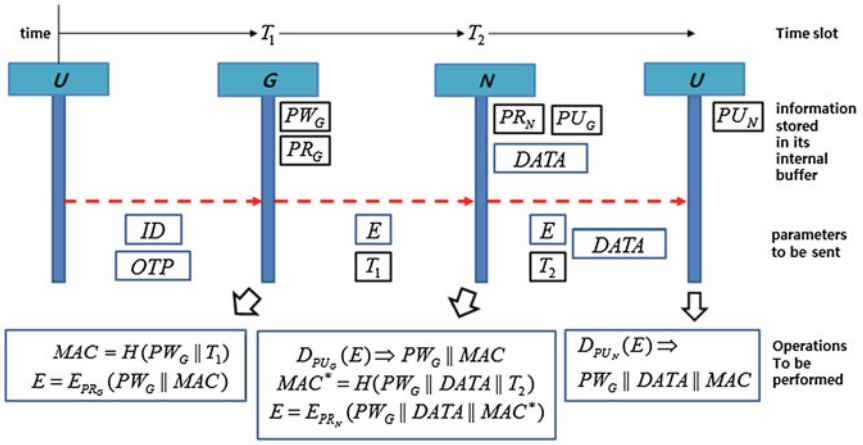


Fig. 10.2 The parameters to be transmitted at each pass

pass 3 the node ***N*** sends the tuple (*E, DATA*, and T_2) to ***U***. In addition, after the registration phase, ***G***, ***N***, and ***U*** already have the information for user authentication.

10.3.3 Cluster Formation During Public Key Cryptography

During user authentication in a WSN, public key cryptography is used for encryption and decryption. To reduce the overheads during key distribution over a WSN, a WSN is partitioned into a set of sub networks, called clusters, as shown in Fig. 10.3.

Each cluster is identified by one cluster-head (CH). A CH is a node that can reach all nodes in the cluster in 1-hop wireless communication. All nodes in each cluster use the common public keys to decrypt or the common private key to encrypt. It can be copied into a specific node from the corresponding CH, when necessary. Thus, given the number n of sensor nodes in a WSN the number of public keys can be reduced to p, where p is the number of clusters in a given WSN and $p \ll n$. In addition, since each cluster has the distinct public keys (or private keys), it is not possible for an attacker who obtains the keys of a cluster to access the data from the node which belongs to another cluster.

The gateway has the number p of private key to be used in the pass 2 as shown in Fig. 10.4. The gateway uses a distinct private key $KEY_{private}^{pass2-CHi}$ assigned to each cluster i to encrypt. All nodes in the same cluster can share a single public key $KEY_{public}^{pass2-CHi}$ to decrypt, which is to be stored in the cluster head. In the pass 3, the user tries to authenticate the node that he wants to access. However, user does not know which cluster contains it. Thus, in the pass 3, all sensor nodes use common private key and also the user has the common public key as shown in Fig. 10.5.

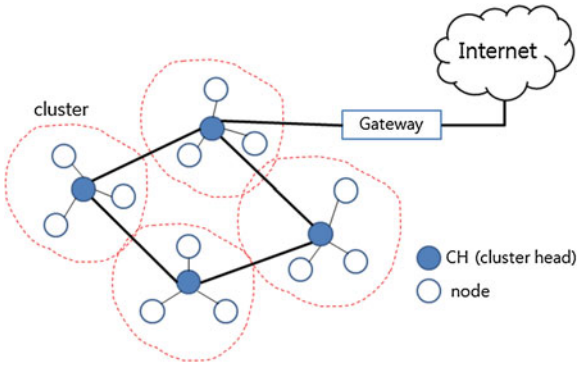


Fig. 10.3 A partitioning WSN into a set of sub networks, called clusters

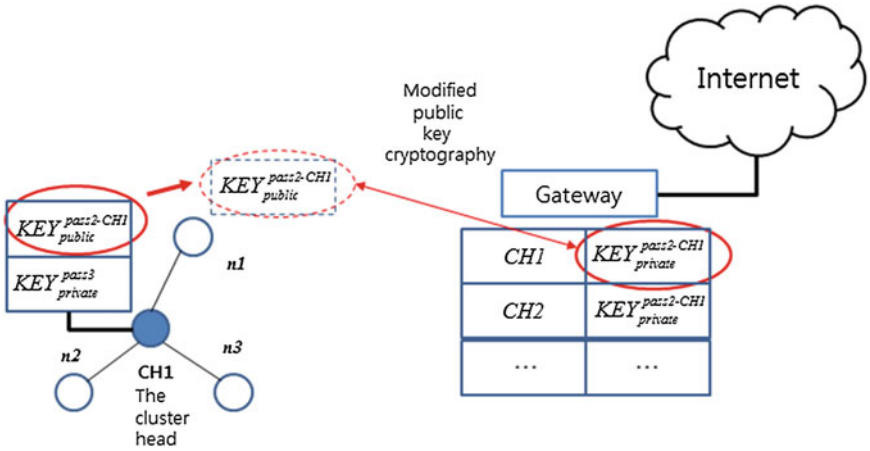


Fig. 10.4 The gateway authentication by the sensor node in the pass 2

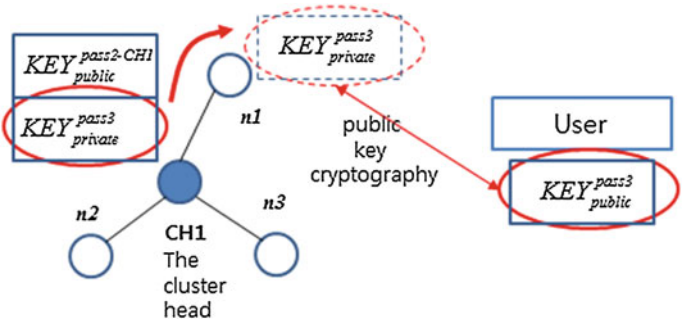


Fig. 10.5 The sensor node authentication by the authenticated user in the pass 3

10.3.4 Pass 2: Gateway Authentication by the Sensor Node

On receiving the message $[E, T_1]$ from the gateway, the sensor node performs the following operations to authenticate the gateway. Let us first check to see if the time interval for completely transmit the message is too long or not. If it is received within a reasonable time, then decode the request message using the public key PU_G. Two parameters, $\mathrm{PW_G}$ and MAC, are extracted from the message. To verify the message, it computes a new MAC^* by using $H(PW_G||T_1)$ and then compares it with the MAC received. If two MACs are equal, then the new MAC is created by using the function $H(PW_G||DATA||T_2)$ as a message authentication code.

To create a secure response message, the tuple $(PW_G, DATA, MAC)$ are encrypted using the private key of the node PR_N. Finally, the encrypted message $\boldsymbol{E}$ and the current timestamp T_2 are sent to the user. Each step in the gateway authentication by the sensor node is described as follows:

- Verify $(T - T_1) < \Delta T$, where ΔT is a pre-defined time interval
- If it does not hold, abort the process; Otherwise, decrypt the E with $PU_G{:}D_{PU_G}(E) \Rightarrow PW_G||MAC$
- Compute $MAC^* = H(PW_G||T_1)$
- Check to see if $MAC^* = ?MAC$
- If it does not hold, reject the request; Otherwise, compute another MAC: $MAC = H(PW_G||DATA||T_2)$
- Compute new E: $E = E_{PR_N}(PW_G||DATA||MAC)$
- Send $[E, T_2]$ to the user.

10.3.5 Pass 3: Sensor Node Authentication by the User

Finally, the user performs the following operations to authenticate the sensor node as well as the data integrity. On receiving the message $[E, T_2]$ from the sensor node, the user performs the following operations to authenticate the sensor node. If it is received within a reasonable time, then decode the request message using the public key PU_N. PW_G, MAC, and $DATA$, are extracted from the message. To verify the message, it computes a new MAC^* by using $H(PW_G||DATA||T_2)$ and then compares it with the MAC received. If two MACs are equal, the user authenticates the sensor node. In addition, the user also authenticates the gateway if the condition $PW_G = H(OTP)$ holds. In summary, the user can authenticate the gateway as well as the sensor node, because he/she can trust the sensor node that has already authenticated by the gateway. In the similar fashion, also the sensor node can authenticate both the user and the gateway. On the other hand, the gateway also can authenticate both the user and the sensor node.

- Verify $(T - T_2) < \Delta T$, where ΔT is a pre-defined time interval;
- If it does not hold, abort the process; Otherwise, decrypt the E with $PU_N{:}\, D_{PU_N}(E) \Rightarrow PW||DATA||MAC$;

- Compute $MAC^* = H(PW_G||DATA||T_2)$;
- Check to see if $MAC^* = ?MAC$;
- If it does not hold, reject it; Otherwise, check to see if the condition $PW_G = H(OTP)$ holds or not.

10.4 Analysis of Proposed Scheme

10.4.1 Security Analysis

1. **Replay attack**

This protocol is robust to the replay attack. Although an attacker obtain a valid login message [ID, OTP], but he cannot login with the same information, because when the valid user successes login, gateway will change its reference. Already used OTP is not possible to be authenticated any more. If the attacker obtain a valid request from the gateway to a sensor node, the replay attack does not work, because of the time interval ΔT. The attacker can forge the timestamp T1, but it still does not work because of the MAC. If he changes the timestamp, then MAC* that computed at sensor node is different with MAC that was sent from the gateway. In the same way, the message from the sensor node to the user can resist the replay attack.

2. **Stolen-verifier attack**

A password/verifier in the gateway is changed when a user successes his/her login request. Also because passwords are consist of hash function which has irreversible feature, if an attacker obtain a password table from the gateway, he/she cannot impersonate the gateway and get the user's initial password or next OTP from the stolen password table.

3. **Guessing attack**

Guessing attack is a critical concern in any password-based system. Because our protocol uses OTP with the one-way hash function, the security is based on the one-way hash property of the hash function. The security of SHA-256 is approximately 2^{128}. Thus it is extremely hard to get the initial password [8].

4. **Impersonation attack**

Suppose an attacker obtain login message [ID, OTP] from the network. However the attacker cannot impersonate valid user, because it is impossible to calculate next OTP, H-1(OTP) to login again. Also, the attacker cannot pretend the gateway as well as sensor node without knowing secret keys, PR_G, PU_G and PR_N.

Table 10.2 Comparison of the number of transmission

Scheme	U	G	SN	Total
Wong et al.'s scheme	1	1	2	4
Das et al.'s scheme	1	1	1	3
Vaidya et al.'s scheme	1	2	2	5
Proposed scheme	1	1	1	3

Table 10.3 Comparison of the authentication at each component

Scheme	U	G	SN
Wong et al.'s scheme		SN	U
Das et al.'s scheme		U	SN
Vaidya et al.'s scheme		U, SN	G
Proposed scheme	G, SN	U, SN	U, G

10.4.2 Performance Analysis

As shown in Table 10.2, proposed scheme is able to obtain data from network with less communication than other schemes. In the WSNs, decreasing the number of transmissions is very important because it may reduce a chance of network congestion and energy consumption of a node.

As shown in Table 10.3, our protocol provides mutual authentication at every component, the user, the gateway and the sensor node. Wong et al. and Das does not provide mutual authentication, and Vaidya et al. only provides mutual authentication between the gateway and sensor node. Mutual authentication can reduce the risk of impersonation attack against each component.

10.5 Conclusion

In this paper, we have proposed a 3-way chaining authentication scheme using S/Key OTP in wireless sensor networks. The proposed protocol resists replay attack, *many* logged in users with the *same* login-id and stolen-verifier attack as well as other treats. In addition, the proposed protocol provides mutual authentication for all components with the small number of transmissions. Also we have shown the efficiency of the proposed protocol in comparisons with the related ones. To reduce the communication overheads a WSN is partitioned into a set of clusters. Particularly for authenticating the gateway and the sensor node with public key cryptography, all nodes in the same cluster can share the common keys. The approach taken here is more robust compared to symmetric key scheme.

References

1. Wong K, Zheng Y, Cao J, Wang S (2006) A dynamic user authentication scheme for wireless sensor networks. In: Proceedings of IEEE international conference on sensor networks, ubiquitous, and trustworthy computing (SUTC'06), pp 244–251
2. Tseng HR, Jan RH, Yang W (2007) An improved dynamic user authentication scheme for wireless sensor networks. IEEE Globecom'07, pp 986–990
3. Vaidya B, Rodrigues J, Park JH (2010) User authentication schemes with pseudonymity for ubiquitous sensor network in NGN. Int J Commun Syst 23(9–10):1201–1220
4. Das ML (2009) Two-factor user authentication in wireless sensor networks. IEEE Trans Wirel Commun 8(3):1086–1090
5. Arikumar KS, Thirumoorthy K (2011) Improved user authentication in wireless sensor networks. In: Proceedings of ICETECT 2011, pp 1010–1015
6. Vaidya B, Makrakis D, Mouftah HT (2010) Improved two-factor user authentication in wireless sensor networks. IEEE 6th international conference of wireless and mobile computing, networking and communications (WiMob), pp 600–606
7. Haller N (1994) The S/KEY one-time password system. In: Proceedings of the ISOC symposium on network and distributed system security, San Diego, CA
8. Agarwal S, Rungta A, Padmavathy R, Shankar M, Rajan N (2012) An improved fast and secure hash algorithm. J Inf Process Syst 8(1):119–132

Chapter 11
Comparative Study of Cyber Security Characteristics for Nuclear Systems

Jinsoo Shin, Hanseong Son and Gyunyoung Heo

Abstract In this work, the cyber security risk model is introduced to perform the cyber security analysis for industrial control system. The model is based on Bayesian network and makes it possible to analyse the vulnerability of a system and the effectiveness of mitigation measure for cyber-attack. To demonstrate the applicability of the model, a comparative study was performed. The study compared the cyber security aspects of two different control systems, which have the same functions but slightly different architectures. The study took advantage of various features of Bayesian network and produced meaningful perspectives. The comparative study results are expected to be applied to select the effective and optimal security architecture of an industrial control system.

Keywords Cyber security risk model · Cyber security evaluation index · Bayesian network · Reactor protection system

11.1 Introduction

Cyber security means the prevention and mitigation about the occurrence probability of cyber terror previously and responding properly if the event of cyber-attack happens. Cyber security has been issued significantly for industrial control

J. Shin · G. Heo
Kyung Hee University, 1732 Deogyeong-daero, Giheung-gu, Gyeonggi-do, Yongin-si 446-701, Republic of Korea
e-mail: bigjoyman@khu.ac.kr

G. Heo
e-mail: gheo@khu.ac.kr

H. Son (✉)
Joongbu University, 201 Daehak-ro, Chubu-Myeon, Chungnam, Geumsan-gun 312-702, Republic of Korea
e-mail: hsson@joongbu.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_11, © Springer Science+Business Media Dordrecht 2014

systems as well as general computing systems [1]. These systems inevitably have their own vulnerabilities and thus have to include appropriate mitigation measures in their architecture. Furthermore, the management aspects of cyber security should be considered to assure the cyber security of the systems. For example, in nuclear industry, regulatory authorities published regulatory guides such as Regulatory Guide 5.71, 1.152 [2–4] for controlling cyber security of nuclear facilities. They deal with the system architecture aspects and the management aspects of cyber security at the same time.

In this work, the cyber security risk model is introduced to perform the cyber security analysis for industrial control system [5, 6]. The model is based on Bayesian network and makes it possible to analyze the vulnerability of a system and the effectiveness of mitigation measure for cyber-attack. To demonstrate the applicability of the model, a comparative study was performed. The study compared the cyber security aspects of two different control systems, which have the same functions but slightly different architectures. The two control systems are research reactor RPS and power plant RPS. The comparison was performed to analyze a few scenarios such as an occurrence of cyber-attack to RPS or subsystem of RPS. The study took advantage of various features of Bayesian network and produced meaningful perspectives. The comparative study results are expected to be applied to determine the effective and optimal security architecture of an industrial control system.

Section 11.2 describes the cyber security risk model. In Sect. 11.3, the analysis results from the previous work [5, 6] are summarized. In addition, the analysis results for RPS of power plants are provided and the comparison of cyber security characteristics is explained. Section 11.4 concludes this article.

11.2 Models for Cyber Security Analysis

The cyber security risk model was developed to perform cyber security analysis for nuclear facilities like power plants and research reactors [5, 6]. Models for cyber security analysis consist of architecture model to consider that architecture of system and activity-quality model to check the management aspects of cyber security. The models are based on Bayesian Network (BN) [7]. The model focused on research reactor RPS architecture as demonstration and performed cyber security risk analysis assuming two threats, which are unexpected reactor trip due to insertion of control rod maliciously and fail to trip timely due to cyber-attack. The model for power plant RPS was also constructed in this work and compared with that for research reactor RPS.

For the development of architecture model, the RPS architectures were studied. The RPS is composed of Bi-stable Processor (BP), Coincidence Processor (CP), Interface and Test Processor (ITP) and Maintenance and Test Processor (MTP) in a single channel and the channel communicate with other channels through Intra-channel [8, 9]. BP makes trip signals for the safety of reactor by comparing

standard trip set points, which are fixed values to determine the critical hazard of reactor, with input data, which are real-signal to inform the state of reactor. The signals are transferred to CP when input data is abnormally matched with trip set points. CP can prevent reactor trip caused by error signal from BP by using the logic of 2-out-of-4 (power plants) or 2-out-of-3 (research reactors) and generate trip signal continually. CP of power plants is a digital system while CP of research reactor is an analog system. Trip information transfer to CP by each set-point parameters in case of power plant RPS but it transfer to CP as a whole in case of research reactor RPS. The number of set-points of power plant RPS is bigger than that of research reactor RPS. ITP tests the state of signals from BP and monitors the status of each RPS channel and handovers these results and values to MTP and PAMS (Post-Accident Monitoring System). MTP provides the display and control needed to support and maintain RPS. In addition, it handovers information to MCR (Main Control Room). In research reactor RPS, data communication direction between BP and CP is one way. Meanwhile, power plant RPS has the two way data communication between BP, CP, MTP and ITP for the bypass functions and the maintenance.

Five vulnerabilities and 6 mitigation measures were modelled in the consideration of this RPS architecture. The 5 vulnerabilities are (1) Denial of Service attack (DoS) occurrences and malware carrying out on systems network during maintenance works (V1), (2) System shut-down by contagion of malware from maintenance works (V2), (3) Data alteration by contagion of malware from maintenance works (V3), (4) DoS occurrences and malware carrying out on other systems by vulnerabilities existing in the system (V4) and (5) Data alteration by using recognized vulnerabilities of standard communication protocols (V5). The 6 mitigation measures are (1) Establishment of managing infection detection systems for external storage media like USB or PC used for PLC maintenance works (M1), (2) Establishment of security system such as firewalls/Intrusion detection system/ intrusion prevention system (M2), (3) Check for running services (M3), (4) Network monitoring (M4), (5) Establishment of device validation policies (M5) and (6) Vulnerability patches (M6). The vulnerabilities and mitigation measures and the RPS architecture are converted to cyber security risk model [5, 6].

For the analysis, cyber security evaluation index (CSEI) was introduced, which means the occurrence probability of each vulnerability node or the completeness of mitigation measures of each node. CSEI of a single node is evaluated by multiplying the numeric values of each degree with the probabilities of happening for each degree. This evaluation is performed according to Eq. (11.1).

$$CSEI = \sum_{s=1}^{5} 10(2s - 1) \times EP \tag{11.1}$$

Here, s is a numeric value of each degree and EP is the probability of happening of each degree. CSEI of each node varies from 10 to 90 points. The point near to 10 means that the occurrence probability of the node is low and the completeness is high.

11.3 Analysis Results and Their Comparison

11.3.1 Analysis Results for RPS of Research Reactor

In [5, 6], the analysis was performed for the following three aims: (1) Vulnerability analysis of MTP; (2) Vulnerability and mitigation measure analysis in case of cyber-attack to RPS; (3) Determination of optimal mitigation measure against vulnerability.

The first analysis aimed to identify and prioritize the vulnerabilities of MTP when a cyber-attack occurs to MTP. One of the analysis results is that V4 came out as the most risk significant vulnerability. This means that it is likely the cyber-attack is done through V4. Another analysis result is that the importance of mitigation measures comes in the following order: M1, M2, M4 and M5. It means that M1 is the most important mitigation measure to the security of MTP against cyber-attacks. The second analysis was to derive the risk information on the vulnerability and mitigation measures of RPS by using the back propagation feature of BN. The back propagated CSEIs showed that CP and Intra-Channel are more risk significant than others. It is consistent with the general intuition that CP and Intra-Channel are critical to illegitimately control the control rod by a cyber-attack to RPS for reactor trip. The vulnerabilities for RPS were prioritized in the following order: V1 and V3 (both are in the same rank), V2, V5, and V4. The important order of mitigation measures is as follows: M3, M2, M1, M6 and M4. In the third analysis, weakness of RPS subsystems against each vulnerability was examined by raising the cyber-attack occurrence probability of subsystems. The results showed that BP and CP are the weakest against V1, V2 and V3, ITP is the weakest against V1 and V3, Intra-Channel is the weakest against V5. Another result of this analysis was that M4 is the optimal mitigation measure. This means M4 reduces the risk of almost all vulnerabilities meaningfully as its completeness increases.

11.3.2 Analysis Results for RPS of Power Plants

The analyses of the cyber security risk for power plant RPS has been performed for the same purposes as those of the analyses for research reactor RPS [5, 6]. The Fig. 11.1 is a simplification of the cyber security risk model for power plant RPS.

The activity-quality node checks how people and/or organization comply with the cyber security regulatory guide [5, 6] and the number of the activity-quality checklists is 27. These checklists have influence on each mitigation measures. The 5 vulnerabilities and 6 mitigation measures have influence on architectural components like BP, CP, ITP, MTP, Intra-channel and other channel BP of RPS of power plant. Cyber-attack propagates through these components, depending upon the safety class and importance of each component. For evaluation of cyber

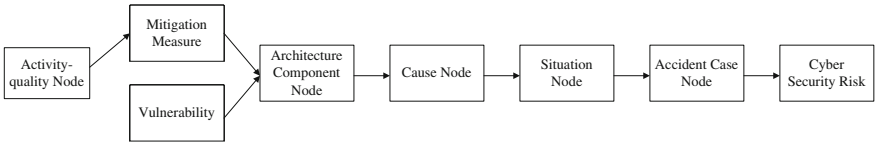


Fig. 11.1 The simplified cyber security risk model for power plant RPS

security risk, cyber-attack scenarios such as 'unexpected power plant trip' and 'fail to trip of power plant' were defined.

The vulnerability analysis of MTP for power plant RPS showed that the significance of vulnerabilities of power plant RPS comes in the following order: V2, V4, V5, V1 and V3. The vulnerability V2 has decreased to 24.85 points by 34.86 points, which is the greatest decrease and lowest CSEI points among vulnerabilities in this analysis, indicating that this vulnerability is the most risk significant. V4 should also be considered as a significant vulnerability because it decreased to 28.81 points by 32.50 points. It means that the cyber-attack to MTP is likely done through V2 or V4. Another result from this analysis is that the importance of mitigation measures comes in the following order: M5, M2, M1, M4, M6 and M3. It means that if MTP was attacked, the completeness of M5 on power plant RPS is most probably weak and thus M5 should be checked with the highest priority for cyber security during maintenance.

The second analysis observed the subsystems of RPS with the assumption that RPS was attacked as a whole. After the back propagation with the model, the CSEIs for subsystems of RPS have become 32.70 points (BP), 32.80 points (CP), 25.02 points (ITP), 32.55 points (MTP), 30.55 points (Intra-Channel) and 36.46 points (BP on the other channel). Being different from the result for the research reactor, the highest CSEI points and the lowest CSEI points are not shown up. It means all subsystems have almost the same security importance for cyber-attacks against power plants. The vulnerabilities for RPS were prioritized in the following order: V3, V2, V4, V1, and V5. It means that V3 affect the cyber security of RPS more significantly than any other vulnerability does, if there is no hard evidence that a specific subsystem is under attack. The CSEIs for mitigation measures have become 37.02 points (M1), 37.79 points (M2), 38.82 points (M3), 37.81 points (M4), 39.55 points (M5) and 37.19 points (M6). This analysis suggests the prioritization of maintenance mitigation measures assuming a cyber-attack to the whole RPS.

The third analysis tried to find out the relation between RPS subsystems and the vulnerabilities. For example, BP and CP are strong against V4 and V5. ITP is the weakest against V3 and MTP is strong against V1 and V5. For power plant RPS architecture, the increase of cyber security risk due to the vulnerability comes in the following order: V3, V2, V4, V1, and V5. These results show that V3 is the most significant to RPS of the vulnerabilities. This analysis also tried to reveal the relation between the mitigation measures and the vulnerabilities. The analysis

results showed that M4 is the best mitigation measure to mitigate V1, V2 and V3. In addition, M2 is the best mitigation measure for V4.

11.3.3 Comparison Between Two Analysis Results

The comparison between the prioritized vulnerability orders for research reactor RPS and power plant RPS when a cyber-attack occurs to MTP showed a difference. This is because a new cyber-attack scenario such as shut-down of CP system through MTP is added due to the use of digital system for CP. The importance orders among mitigation measures were also different according to the vulnerabilities. This means that, when a cyber-attack to RPS occurs, the CSEIs for each subsystems of power plant RPS are less than those of research reactor RPS because power plant RPS has more complex architecture like bidirectional communication and bypass function. The analysis results to find out the relation between RPS subsystems and vulnerabilities and the relation between optimal mitigation measures and vulnerabilities are almost the same for the two facilities. It is because the vulnerability is identified by the function of a subsystem and its mitigation measures are determined according to the type of vulnerability.

11.4 Conclusion

Though Bayesian network has been applied to the risk analysis of cyber-attack, but it has not been applied to architecture of specific system. In this work, cyber security risk model has been developed to reflect the industrial control system architecture with BN. A few analyses for two similar industrial control systems were performed by using the model. An analysis was performed with the assumption that a cyber-attack occurs to a MTP. In this analysis, the important vulnerabilities and mitigation measures could be identified with the CSEI. The vulnerabilities and the mitigation measures were also analysed with a cyber-attack to RPS assumed. If a cyber-attack occurs in a system scale, it is important to have confidence on which component is the key element corresponding to the attack situation. This analysis proved that the developed model could provide this kind of information through the back propagation feature of the BN. In addition, the analysis of the RPS cyber security risk and the optimal mitigation measures regarding vulnerabilities was performed. This analysis infers that the use of the cyber security risk model makes it possible to create simulated penetration test scenarios. Finally, we compared the analysis results between RPSs of two facilities. Though the two systems have similarity in function and design, the differences of important vulnerability and mitigation measure for specific situation, the key elements for cyber-attack and the final cyber-attack risk to each system are found out with the proposed cyber security risk model. The effective and optimal

security architecture of an industrial control system can be selected by using the results from the comparative study.

Since the completeness of a BN model is affected by the accuracy of relationship values of the nodes, a future research will be done to improve the values by incorporating more expert opinions and information.

References

1. Collins S, McCombie S (2012) Stuxnet: the emergence of a new cyber weapon and its implications. J Polic Intell Count Terror 7(1):80–91
2. USNRC (2010) Regulatory guide 5.71, Cyber security programs for nuclear facilities
3. USNRC (2006) Regulatory guide 1.152 revision 2, Criteria for use of computers in safety systems of nuclear power plants
4. USNRC (2011) Regulatory guide 1.152 revision 3, Criteria for use of computers in safety systems of nuclear power plants
5. Shin JS, Son HS, Rahman K, Heo GY (2013) Development of cyber security evaluation model using bayesian networks, Reliability engineering and system safety (submission)
6. Shin JS, Son HS, Heo GY (2013) Cyber security risk analysis model composed with activity-quality and architecture model. International conference on computer, networks and communication engineering, pp 609–612
7. Heckerman D (1996) A tutorial on learning with Bayesian networks
8. Lee DY, Choi JG, Lyou J (2006) A safety assessment methodology for a digital reactor protection system. Int J Control Autom Syst 105–112
9. Park GY, Bae SH, Bang DI, Kim TG, Park JK, Kim YK (2011) Design of instrumentation and control system for research reactors. 11th international conference on control, automation and systems, pp 1728–1731, Oct 2011

Chapter 12
Optical Computing for Digital Signal Process in Incoherent Fiber System Using Positive Realization

Kyungsup Kim, Jaecheol Ryou and Woo-Tak Jung

Abstract We discuss an optical digital signal process in incoherent optical fiber system. Due to the strong dependence on environmental fluctuations, stable photonic filters are difficult to implement under coherent operation. We pay attention to a stable robust technique operating in incoherent optical domain. Most of coefficients of filters working under incoherent optical domain are positive. Finding a realization of an optimal digital filter under positive constraints is known as an open and difficult problem. We propose a novel constructive method to implement optical digital filter using the sparse positive realization with possible lower order in the incoherent optical domain.

Keywords Optical computing · Photonic filter · Optical fiber · Incoherent domain · Positive system · Positive realization · Optical networks

12.1 Introduction

In a digital electronic computer, a coprocessor is an important special purpose processor. A high performance coprocessor is necessary for performing massive computational tasks, such as signal or image processing problems. There has been considerable research interest in developing optical computing coprocessors for general purpose digital electronic computers [3]. Optical or photonic computing

K. Kim (✉) · J. Ryou · W.-T. Jung
Department of Computer Engineering, Chungnam National University,
99 Daehak-ro, Yuseong-gu, Daejeon, Korea
e-mail: sclkim@cnu.ac.kr

J. Ryou
e-mail: jcryou00@gmail.com

W.-T. Jung
e-mail: zpl1nt@gmail.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_12, © Springer Science+Business Media Dordrecht 2014

uses photons and light produced by lasers or diodes for computation. Generally, photonic devises promise to allow a higher bandwidth than the electronic devices used in conventional electronic computers. Our objective of this paper is to develop fiber-optical signal coprocessors for efficient and direct processing of high speed wide band signals in the optical domain. The advantage of the optical amplifier and the development of optical components (variable couplers, modulators, and electro-optic switches) and specific purpose instrumentation fueled the activity toward more flexible structures. Photonic devices can be used to implement flexible filters for microwave and radio frequency signal [6]. The photonic filter using fiber and integrated photonic devices perform the required signal processing tasks of RF signals conveyed by an optical carrier in the optical domain. This approach has several advantages: the low loss of optical delay lines, very high time bandwidth products and short delay in very high-speed sampling frequency [6]. Photonic digital filters are of interest in a wide variety of applications. The potential applications lay in the field of RF based services such as optical networks, mobile telephony (UMTS and 4G), wireless LAN, RADAR systems, satellite communications and signal processing [6, 7].

Specially, we pay special attention to signal processing techniques in coherent optical domain. Incoherent fiber-optic signal processors usually require the coherence time of the optical source to be much shorter than the basic time delay in the system, and can be directly or externally modulated. Although incoherent fiber-optic signal processors are stable and robust, they can only perform positive-valued signal processing operations and thus have limited applications. We address the efforts to overcome the limitations in incoherent signal processing imposed by the positive nature of their coefficients. A positive realization method introduced in control system theory has been applied to implement the optical filter [2, 9, 13]. The existence and minimum problems of positive realization are known as open and difficult problems. We propose a novel constructive method to implement optical digital filter using the sparse positive realization with possible lower dimension in the incoherent optical domain.

The outline of this paper is as follows. In Sect. 12. 2, we review some architectures of optical computing processor for signal processing and describe the fundamental concepts and limitations of optical computing processor. In Sect.12. 3, we propose a theoretical method to construct a sparse positive realization of a given transfer functions for photonic filter in the incoherent optical domain. In Sect.12. 4, we summary and discuss the filter design procedure of the fiber-optic filters in incoherent fiber system.

12.2 Preliminary

The objective of digital optical processor is to replace electrons with photons whenever appropriate in a computing environment. We consider architecture of optical computing processor. Figure 12.1 shows a general scheme of the optical

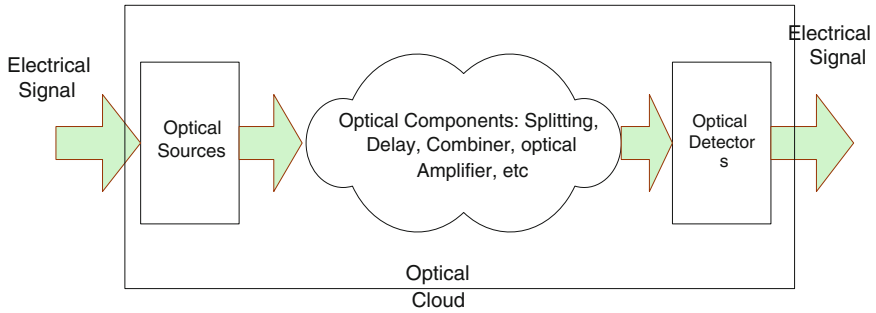


Fig. 12.1 General scheme of the optical processor formed by an optical cloud

processor where the core is formed by an optical component clouds, sources as the input and photonic detectors at the end. To perform the various signal processing operations, we need to employ the elementary optical devices such as time delays, amplifiers, combiners and splitters [6, 13]. Optical splitters and optical combiners are the most common components in the optical communication system. Optical splitters can be realized easily by using optical directional coupler. An essential function in optical networks is combining optical signals. The passive combiner is obtained with fiber couplers. The most convention combiners experience optical energy loss that it has an inherent 3-dB loss. Several combiner designs to solve the optical energy loss was introduced [7, 13]. An optical amplifier is an important component in fiber-optic signal processors because it can compensate for optical losses as well as providing design flexibility resulting in potential applications. Optical amplifiers, in particular erbium-doped fiber amplifiers (EDFAs), have been used to overcome losses as well as to provide greater flexibility in the analysis, synthesis, and construction of incoherent fiber-optic signal processors for various filtering applications.

It is assumed that the optically encoded signals to be processed are modulated onto an optical carrier whose coherence time τ_c is very short compared to the basic time spacing T [6]. Under coherent regime operation, the optical phase of the taps plays a dominant role in overall time and frequency response. Due to the strong dependence on environmental fluctuations, stable photonic filters are difficult to implement under coherent operation. Any slight change in the propagation characteristics of components drastically affects the filter response and its properties. This circumstance makes very serious practical limitation for the system implementation in coherent optical domain. On other hand, since the optical intensity at the photo-detector is detected in incoherent optical operation, optical phase effects can be removed. Incoherent operations are employed in the majority of cases, since it is compact and its performance is very stable and robust.

The main drawback of incoherent approach is that most of filter coefficients are only positive. Photonic filters with only positive coefficients have severe limitations in practical applications [4, 6, 7]. It has been claimed that this synthesis technique can implement not only low-pass filters but also high-pass and band-pass filters.

The limitation of the incoherent (or positive) fiber-optic signal processors may partially be solved by using electronic-optical devises. In order to solve this problem, many methods have been proposed to obtain complex or negative coefficients [4, 6, 7]. A straightforward way to generate negative coefficients is to use differential detector [13]. The negative coefficient was not generated directly in the optical domain; the filter is not all-optical but hybrid. A few techniques have been proposed to implement an all-optical photonic microwave delay-line filter with negative coefficients. We will propose an analytical method for positive realization that can be applied to the optical filter design by combining the optical devices in incoherent domain.

12.3 Constructive Sparse Positive Realization Method

We consider a theoretical method to construct an incoherent photonic filter for given specification. Because the signals are modulated as intension variations on optical carriers whose coherence time is less than the shortest relevant time spacing, only filter taps with nonnegative values have to be implemented. But, the most widely used filters such as Butterworth and Chebyshev have no sign limitation on the their impulse response [2]. We note that an arbitrary filter can be implemented by the difference of two positive filters [1, 13]. Using electronic-optical devises, the difference is implemented in the system end [13]. The following theorem proven in [1, 2] provides an affirmative answers to design a general optical filter under incoherent optical domain.

Theorem 1 *Let $H(z)$ be a strictly proper asymptotically stable transfer function of order n. Then $H(z)$ can be realized as the difference of an N-dimensional positive system and a one-dimensional positive system for an appropriate $N > n$, i.e., $H(z) = H_+^1(z) - H_+^2(z)$ where $H_+^1(z)$ is an N-dimensional positive system and $H_+^2(z) = \frac{R}{z-\gamma}$ for a sufficiently large $R > 0$ and $\sigma(H(z)) < \gamma < 1$.*

The previous theorem states a possibility to realize an arbitrary proper transfer function as the difference of two positive filters, but it does not provide how to construct the positive realization of $H_+^1(z)$ with possibly minimal N. The question has two problems: the positive realization problem and the minimality problem. I is an identity matrix. The positive realization problem is to find, if possible, a triple (A_+, B_+, C_+) with nonnegative entries, such that $H_+^1(z) = C_+(zI - A_+)^{-1}B_+$ holds. The minimality problem is to find the minimal possible value of N (clearly, $N \geq n$). Recently, these problems have extensively been studied by many researchers [9]. Some constructive methods to implement positive realizations with lower bound of the proper transfer functions with possible multiple real and multiple complex poles have been studied in [11, 15]. When a positive realization is made, we can construct oriented weighted graphs and convert the previous graphs into two fiber-optic filter schemes. Therefore, the problem of the sparse

positive realization with possibly lower bound is very important for reducing cost and complexity in view of implementation. For a given proper transfer function $H(z) = C(zI - A)^{-1}B$, the augmented state space realization $(\bar{A}, \bar{B}, \bar{C})$ is defined by

$$\bar{A} = \begin{pmatrix} A & B \\ 0 & 1 \end{pmatrix}, \quad \bar{B} = \mathbf{e}_{n+1,n+1} \\ C = (\, C \quad 0\,). \tag{12.1}$$

A discrete phase-type (DPH) distribution is the distribution of the time until absorption in a discrete-state discrete-time Markov chain with n transient states and one absorbing state [5, 16]. $\mathbf{e}$ is a column vector of ones. For DPH, the Eq. (12.1) satisfies the conditions that A is a sub-stochastic matrix, $(A - I)\mathbf{e} + B = 0$ and $C\mathbf{e} = 1$. For a continuous time case, a positive realization with a proper condition can be transformed into a phase-type representation multiplied by a positive number [8]. Similarly, For discrete-time case, a positive realization can be transformed into a DPH representation multiplied by a positive number. We consider an excitable system as a new class of positive system introduced in [9, 14]. A positive system (A, B, C) is said to be excitable if each state variable can be made positive by applying an appropriate nonnegative input to the system.

Theorem 2 *Assume that (A, B, C) is a positive realization of a transfer function H(z) where (A, B) is excitable, A is an asymptotically stable and a nonnegative matrix. Then there is an equivalent positive realization* (A_+, B_+, C_+) *of H(z) such that* $(A_+ - I)\boldsymbol{e} = -B_+$, A_+ *is a sub-stochastic matrix and* $C_+ \geq 0$.

A very important consequence of the above theorem is that a positive realization is converted in the form of DPH representation. By the page limit, the detailed proof is omitted. The DPH representation is known to be non-unique and non-minimal. The number of parameters of the representation generally is $n^2 + n - 1$, while its Laplace transform has at most algebraic $2n - 1$ free parameters. We need to find a canonical realization with minimal free parameters. For continuous time cases, several candidates for canonical positive realizations such as coxian, mono-cyclic coxian and unicyclic coxian have been introduced in [5, 8, 12, 16]. A triangular PH representation is equivalent to an ordered coxian representation of at most the same order [8]. Every phase-type representation has a mixture of mono-cyclic generalized erlang representation [12]. Similarly, we can remark that every DPH representation has an equivalent positive realization consisting of a mixture of mono-cyclic coxian.

Remark 1 A monocyclic coxian realization $M(\lambda, z)$ is defined by

$$M(\lambda, z) = \begin{bmatrix} \lambda_1 & 0 & \dots & 0 & z\lambda_{1'} \\ \lambda_{2'} & \lambda_2 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & \lambda_{n-1} & 0 \\ 0 & 0 & \dots & \lambda'_n & \lambda_n \end{bmatrix}$$

where $\lambda_i + \lambda_{i'} = 1$ and $0 \leq z < 1$. As $z = 0$, it is a coxian representation. A hypo-monocyclic-coxian representation (Θ, α) is defined by a transient generator having the following special structure

$$\Theta = \begin{bmatrix} M_1 & 0 & 0 & \dots & 0 \\ M_2^* & M_2 & 0 & \dots & 0 \\ 0 & M_3^* & M_3 & \dots & 0 \\ \vdots & & & \ddots & \\ 0 & 0 & & & M_J \end{bmatrix} \tag{12.2}$$

where M_i is a mono-cyclic Coxian for each i and M_j^* is defined by $M_j^* = (-(M_j - I))\mathbf{e}e_n$ and e_n is a column vector whose i-th elements is one and all others are zeros. Every DPH representation can be transformed into mixture of mono-cyclic coxian representation (12.2) (i.e., called a hypo-monocyclic-Coxian) with a proper order. The order of a hypo-mono-coxian representation is usually larger than that of a DPH representation. We need a nonlinear optimization programming to compute Θ, with the (possibly) minimal order, whose eigenvalue set includes all eigenvalues of A counting multiplicities.

A spectral polynomial algorithm introduced in [10] is easily modified to computing a mono-cyclic coxian representation from a given positive realization with real poles. We can extend it to a mono-cyclic coxian representation of the transfer function with complex pairs of poles. We modify a spectral polynomial algorithm for a mono-cyclic coxian representation. The eigenvalues of A are included in the set $\{\lambda_1, \dots, \lambda_N\}$ with order $\lambda_1 \leq \lambda_2 \dots \leq \lambda_N$. For a given column vector $p_1 = B$ of size m, we define

$$\begin{cases} p_k = (-\lambda_{k-1} I + A) p_{k-1} / \lambda_{k'}., & for\, 2 \leq k \leq N; \\ p_{N+1} = ((-\lambda_N I + A) p_N + z\lambda_1 p_1, & n = N. \end{cases} \tag{12.3}$$

We have $AP = PM(\lambda, z)$ where P is an $n \times N$ matrix $P = [p_1 \quad p_2 \quad \cdots \quad p_N]$. If all eigenvalues of A (counting multiplicities) are in λ, we have $p_{N+1} = 0$. We can also generalize the spectral polynomial algorithm for a hypo-moncyclic-coxian Θ without difficulty. we have $AP = \Theta P$ and $CP = \alpha \geq 0$. Finally, we obtain an equivalent sparse positive realization of the form $(\Theta, \Theta\mathbf{e}, \alpha)$ with $2N - 1$ free parameters from a given (A, B, C).

12.4 Filter Design Procedure and Discussions

By using various filter design methods such as Butterworth, Chebyshev, Elliptic and etc., we can obtain a transfer function for a given specifications. The most widely used filters have no sign limitation on system parameters. In incoherent fiber-optic signal processors, they can only perform positive-valued signal processing

operations and thus have limited applications. We propose a novel constructive method to implement optical digital filter. We summary the filter design process of the fiber-optic filters as follows. We assume that a strictly proper function $H(z)$ can be obtain by using one of various filter design methods for a given specification.

By using Theorem 1, the transfer function $H(z)$ can be realized by the difference of an n-dimensional positive system $H_+^1(z)$ with state space realization (A_1, B_1, C_1) and a one-dimensional positive system $H_+^2(z)$ with (A_2, B_2, C_2). Construct sparse positive realizations for the given positive systems.

(a) Compute each state space realization (A_i, B_i, C_i).
(b) Compute the eigenvalues set Ω_i of A_i counting multiplicities
(c) Compute a hypo-monocyclic coxian representation Θ whose eigenvalue set includes Ω_i, with a possible minimal order.

Finally, we obtain an equivalent sparse positive realization $(\Theta, \Theta\mathbf{e}, \alpha)$ such that $\alpha \geq 0$ by using Remark 1.

Construct the oriented weighted graphs with respect to the sparse positive realization $(\Theta, \Theta\mathbf{e}, \alpha)$. Convert the previous graphs into two fiber-optic filter schemes. In order to compute minus operation in Theorem 1, use a few difference detectors made of the combination of electronic devises and optical ones. In Step 3–3 and 3–4, we should solve nonlinear optimization problems to find positive realizations satisfying the positive constraints. In Step 4, every positive linear system can be realized by a combination of delays, positive amplifiers, adders and splitters [9] depending on the positive realization (A_+, B_+, C_+). These operations can be implemented by some fiber-optic directional couplers, delay lines, and optical amplifiers [2, 4, 13].

This paper presented a design method of optical digital signal process in incoherent optical fiber system. To improve the implementation by the positive nature of the coefficients under incoherent optical domain, we proposed a novel constructive method to implement optical digital filter using the sparse positive realization with possible lower order.

Acknowledgments This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the Support industry-university cooperation Specialization program (NIPA-2012-H0803-13-1004) supervised by the NIPA (National IT Industry Promotion Agency).

References

1. Benvenuti L, Farina L, Anderson B (1999) Filtering through combination of positive filters. IEEE Trans Circ Syst I: Fundam Theor Appl 46(12):1431–1440
2. Benvenuti L, Farina L (2001) The design of fiber-optic filters. J Lightwave Technol 19(9):1366–1375
3. Bruce Berra P, Ghafoor A, Guizani M, Marcinkowski S, Mitkas P (1989) Optics and supercomputing. Proc IEEE 77(12):1797–1815
4. Binh LN (2008) Photonic signal processing: techniques and applications. CRC Press, Taylor and Francis, Boca Raton

5. Bobbio A, Horvath A, Scarpa M, Telek M (2003) Acyclic discrete phase type distributions: properties and a parameter estimation algorithm. Perform Eval 54(1):1–32
6. Capmany J, Ortega B, Pastor D, Sales S (2005) Discrete-time optical processing of microwave signals. J Lightwave Technol 23(2):702–723
7. Capmany J, Ortega B, Pastor D (2006) A tutorial on microwave photonic filters. J Lightwave Technol 24(1):201
8. Commault C, Mocanu S (2003) Phase-type distributions and representations: some results and open problems for system theory. Int J Control 76(6):566–580
9. Farina L, Rinaldi S (2000) Positive linear systems: theory and applications. Wiley Interscience, New York
10. He QM, Zhang H, Xue J (2011) Algorithms for coxianization of phase-type generators. INFORMS J Comput 23(1):153–164
11. Kim K (2012) A construction method for positive realizations with an order bound. Syst Control Lett 61(7):759–765
12. Mocanu S, Commault C (1999) Sparse representations of phase-type distributions. Communications in statistics. Stoch Models 15(4):759–778
13. Moslehi B, Goodman J, Tur M, Shaw H (1984) Fiber-optic lattice signal processing. Proc IEEE 72(7):909–930
14. Muratori S, Rinaldi S (1991) Excitability, stability, and sign of equilibria in positive linear systems. Syst Control Lett 16(1):59–63
15. Nagy B, Matolcsi M, Szilvasi M (2007) Order bound for the realization of a combination of positive filters. IEEE Trans Autom Control 52(4):724–729
16. Neuts MF (1994) Matrix-geometric solutions in stochastic models. Dover Publications, New York

Chapter 13
Optimization of Non-fuzzy Neural Networks Based on Crisp Rules in Scatter Partition

Keon-Jun Park, Byun-Gon Kim, Kwan-Woong Kim, Jung-Won Choi and Yong-Kab Kim

Abstract We introduce a design of non-fuzzy neural networks that have crisp rules in scatter partition. To generate the crisp rules and construct the networks, we use hard c-means clustering algorithm. The partitioned local spaces indicate the crisp rules of the proposed networks. The consequence part of the rule is represented by polynomial functions. The coefficients of the polynomial functions are learned using back-propagation algorithm. In order to optimize the parameters of the proposed networks we use particle swarm optimization techniques. The proposed networks are evaluated with the example for nonlinear process.

Keywords Non-fuzzy neural networks (NFNNs) · Crisp rules · Scatter partition · Hard C-means clustering · Particle swarm optimization

K.-J. Park · J.-W. Choi · Y.-K. Kim (✉)
Department of Information and Communication Engineering, Wonkwang University, 344-2, Shinyong-dong, Iksan-si, Chonbuk 570-749, South Korea
e-mail: ykim@wonkwang.ac.kr

K.-J. Park
e-mail: bird75@wonkwang.ac.kr

B.-G. Kim
Department of Electronic Engineering, Kunsan National University, Kunsan, South Korea
e-mail: bgkim@kunsan.ac.kr

K.-W. Kim
Thunder Technology, Director in Digital Signer Processing Team, ChonJu, South Korea
e-mail: kwkim@thunderspeaker.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_13, © Springer Science+Business Media Dordrecht 2014

13.1 Introduction

Fuzzy neural networks (FNNs) come from an intelligent model by combining the human-like reasoning method of fuzzy inference systems and connectionist structure of neural networks [1–3]. FNNs refer to synergize these two hybrid techniques. Generally, FNNs are represented by fuzzy "if-then" rules while the back propagation (BP) is used to optimize the parameters. There are still some difficulties to develop adequate fuzzy rules and membership functions to reflect the essence of the data. In particular, there still remains the problem of the curse of the dimensionality in conventional grid space [3].

In this paper, we introduce a design of non-fuzzy neural networks based on crisp rules by means of hard scatter partition of input space using hard c-means (HCM) clustering algorithm [4]. The premise part of the rules is realized with the aid of the hard partition of input space generated by HCM clustering algorithm. The consequence part of the rule is represented by polynomial functions. And the coefficients of the polynomial functions are learned by BP algorithm. We also optimize the parameters of the proposed networks using particle swarm optimization (PSO) algorithm [5]. The proposed networks are evaluated through the numeric experimentation for nonlinear process [6].

13.2 Design of Non-fuzzy Neural Networks

13.2.1 The Structure of Non-fuzzy Neural Networks

The structure of the non-fuzzy neural networks involves HCM clustering algorithm in the premise part and neural networks present in the consequence part of the rules. The structure of the proposed networks is composed of four-layers as shown in Fig. 13.1.

The proposed networks are implied by the hard scatter partition of input spaces. In this sense, each rule can be viewed as a certain rule of the following format.

$$R^j : \text{If } x_1 \text{ and} \cdots \text{and} x_d \text{ is} H_j \quad \text{Then} y_j = f(x_1, \ldots, x_d). \tag{13.1}$$

As far as inference schemes are concerned, we distinguish these cases:

Type 1 (Simplified Inference): $f = w_{j0}$
Type 2 (Linear Inference): $f = w_{j0} + \sum_{k=1}^{d} w_{jk} x_k$

To be more specific, R^j is the j-th crisp rule, while H_j denotes j-th membership grades using HCM clustering algorithm. w's are consequent parameters of the rule.

The functionality of each layer is described as follows.

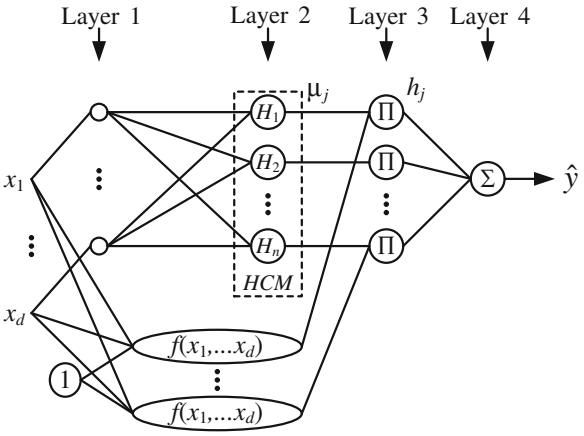


Fig. 13.1 The structure of non-fuzzy neural networks

[Layer 1] The nodes in this layer transfer the inputs.
[Layer 2] The nodes here are used to calculate the membership degrees using HCM clustering algorithm.

$$\mu_j = HCM(\mathbf{x}). \tag{13.2}$$

[Layer 3] The nodes in this layer realize a certain inference process.

$$h_j = \mu_j y_j. \tag{13.3}$$

[Layer 4] The nodes in this layer compute the outputs.

$$\hat{y} = \sum_{j=1}^{n} h_j. \tag{13.4}$$

13.2.2 The Learning Algorithm

The parametric learning of the network is realized by adjusting connections of the neurons and as such it could be realized by running a standard back-propagation (BP) algorithm. The performance index is based on the Euclidean distance. As far as learning is concerned, the connections are adjusted in a standard fashion.

Quite commonly to accelerate convergence, a momentum coefficient α is being added to the learning expression. The complete equations are as follows:

$$\Delta w_{j0} = \eta\big(y_p - \hat{y}_p\big)\mu_j + \alpha\big(w_{j0}(p) - w_{j0}(p-1)\big) \tag{13.5}$$

$$\Delta w_{jk} = \eta\big(y_p - \hat{y}_p\big)\mu_j x_k + \alpha\big(w_{jk}(p) - w_{jk}(p-1)\big). \tag{13.6}$$

13.3 Optimization

Particle swarm optimization (PSO) [5] was proposed by Kennedy, Eberhart to simulate social behavior by representing the movement of a bird flock or fish school. PSO is a computational algorithm that optimizes a given problem by iteratively trying to improve candidate solutions (particles). PSO algorithm optimizes a given problem by having a swarm of particles and moving these particles around in the search space.

Each particle's movement is affected by its local best positions and is also guided toward the global best positions in the search-space over the particle's position and velocity. And these positions are updated as better positions. The swarm move toward the best solutions.

In order to optimize the parameters of the proposed networks, we determined the learning rate and moment coefficient composed of the consequent part of the rules.

13.4 Experimental Studies

In this section, we discuss numerical example to apply to the nonlinear process. This time series data (296 input–output pairs) coming from the gas furnace nonlinear process has been intensively studied [6]. The delayed terms of methane gas flow rate $u(t)$ and carbon dioxide density $y(t)$ are used as six input variables organized in a vector format as $[u(t-3), y(t-1)]$. $y(t)$ is the output variable. The first part of the data set (consisting of 148 pairs) was used for training purposes. The remaining part of the series serves as a testing data set. We consider the MSE as a performance index.

Table 13.1 summarizes the performance index for training and testing data by setting the number of clusters and inference type before and after optimization. Here, PI and E_PI stand for the performance index for the training data set and the testing data set, respectively.

From Table 13.1, we know that the optimized networks are better than before optimization. From Table 13.1a we select the network with seven rules (clusters) with linear inference that exhibits PI = 0.022 and E_PI = 0.332. And from Table 13.1b we select the best network with two rules (clusters) with linear inference that exhibits PI = 0.040 and E_PI = 0.275. From the results, we constructed the non-fuzzy neural network that had a small number of rules because the generation of the rules depends on the clusters.

Figure 13.2 shows hard-partitioned input spaces using HCM clustering algorithm for the selected networks. Figure 13.2a and b depict local spaces of the network with seven rules before optimization and of the network with two rules after optimization, respectively.

Table 13.1 Performance index for the proposed networks

No. of rules	Before optimization		After optimization	
	PI	E_PI	PI	E_PI
(*a*) *Simplified inference*				
2	3.589	3.804	3.570	3.771
3	1.921	2.205	1.672	2.158
4	1.089	1.649	1.069	1.667
5	1.035	1.703	1.034	1.703
6	0.933	1.849	0.901	1.847
7	0.846	1.961	0.828	1.949
8	0.760	1.494	0.714	1.455
9	0.833	1.723	0.804	1.671
10	0.675	1.415	0.624	1.377
(*b*) *Linear inference*				
2	0.025	0.344	**0.040**	**0.275**
3	0.022	0.354	0.031	0.330
4	0.022	0.367	0.024	0.371
5	0.023	0.364	0.029	0.365
6	0.028	0.363	0.030	0.345
7	**0.022**	**0.332**	0.024	0.301
8	0.027	0.357	0.029	0.353
9	0.033	0.362	0.038	0.315
10	0.072	0.310	0.040	0.285

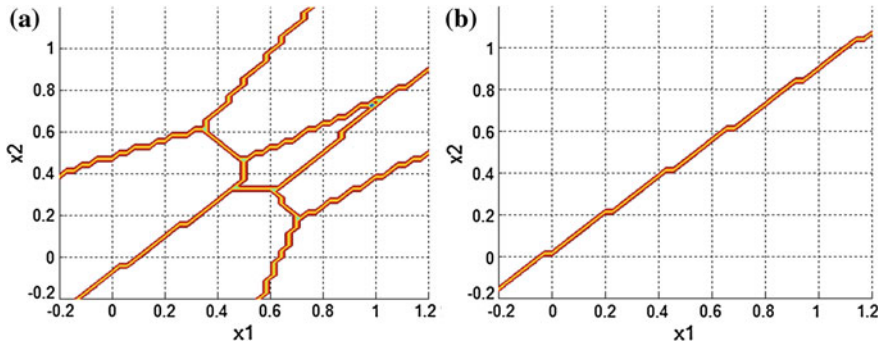


Fig. 13.2 Hard-partitioned input spaces for the selected network. **a** Before optimization. **b** After optimization

Figure 13.3a and b show the optimization process of the objective function and parameters of the network, respectively. Here, gBObj and min(obj) mean the objective functions for the global best position and for the minimization position of the particles, respectively. These figures show that the convergence is quick.

Figure 13.4 illustrates the original and model outputs of training and testing data for the selected network. This figure shows that the model output is

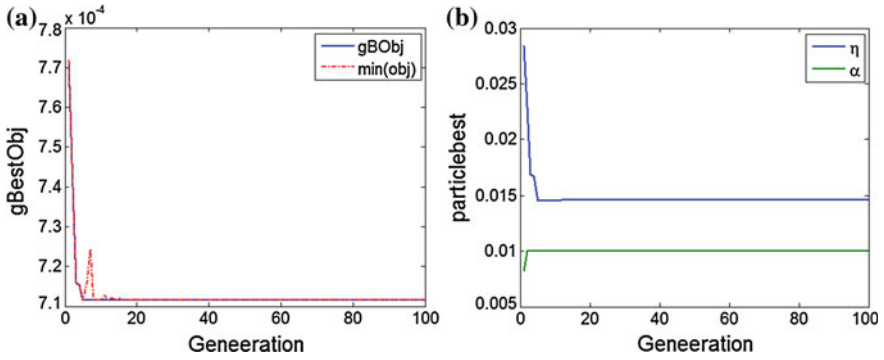


Fig. 13.3 Optimization processes for the selected networks. **a** Objective function, **b** Parameters

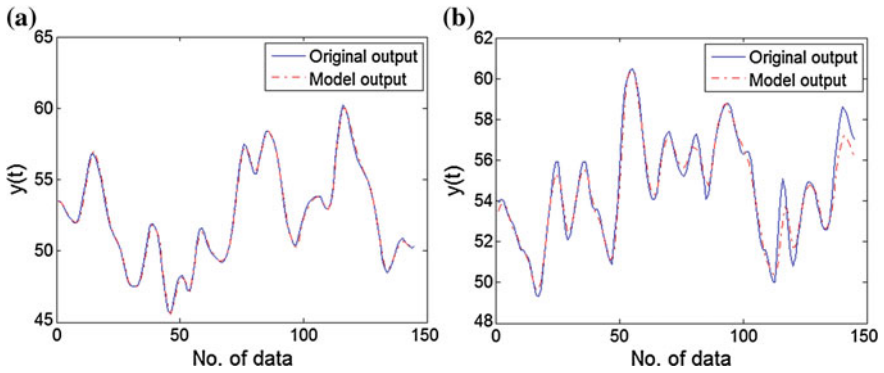


Fig. 13.4 Model outputs. **a** Training data. **b** Testing data

approximately the same for the original output. Finally we conclude that the proposed networks have better result than the traditional fuzzy neural networks.

13.5 Conclusions

In this paper, we introduced the non-fuzzy neural networks based on crisp rules in hard-partition of input space for nonlinear process. The input spaces of the proposed networks were divided as the hard scatter partition using HCM clustering algorithm. And we also optimize the parameters of the proposed networks. By this method, we constructed the good networks that are compact and simple.

From the results, we were able to design the preferred networks. Through the use of a performance index, we were able to achieve a balance between the

approximation and generalization abilities of the resulting networks. Finally, this approach would find potential application in various areas.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2011835).

References

1. Buckley JJ, Hayashi Y (1994) Fuzzy neural networks: a survey. Fuzzy Sets Syst 66:1–13
2. Jang JSR, Mizutani E, Sun CT (1997) Neuro-fuzzy and soft computing, a computational approach to learning and machine intelligence. Prentice Hall, NJ
3. Park KJ, Kim YK, Kim BG, Hoang GC (2013) Design of optimized pattern recognizer by means of fuzzy neural networks based on individual input space. JIIBC 13:181–189
4. Krishnaiah PR, Kanal LN (eds) (1982) Classification, pattern recognition, and reduction of dimensionality, volume 2 of handbook of statistics. North-Holland Amsterdam
5. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of IEEE international conference on neural networks, vol 4, pp 1942–1948
6. Box GEP, Jenkins GM (1976) Time series analysis: forecasting and control, 2nd edn. Holden-Day, San Francisco

Chapter 14
LBG-Based Non-fuzzy Inference System for Nonlinear Process

Keon-Jun Park, Kwan-Woong Kim, Byun-Gon Kim, Jung-Won Choi and Yong-Kab Kim

Abstract We introduce a non-fuzzy inference system based on Linde-Buzo-Gray (LBG) Algorithm to construct model for nonlinear process. In grid partition, the generation of fuzzy rules has the problem that the number of fuzzy rules exponentially increases. To solve this problem, we generate the subspaces using the scatter partition of input space. The rules of non-fuzzy inference systems are generated by partitioning the input space in the scatter form using LBG algorithm. The consequence part of the rules is represented in the form of polynomial functions. The data widely used in nonlinear process is used to evaluate the performance of the proposed model.

Keywords Non-fuzzy inference systems · Linde-buzo-gray (LBG) algorithm · Hard scatter partition · Rule generation · Nonlinear characteristics

14.1 Introduction

The researches on fuzzy modeling to deal with complex, ill-defined, and uncertain systems in many other avenues have been exploited for a long time [1]. Linguistic modeling [2] and fuzzy relation equation-based approach [3] were proposed as

K.-J. Park · J.-W. Choi · Y.-K. Kim (✉)
Department of Information and Communication Engineering, Wonkwang University, 344-2, Shinyong-dong, Iksan-si, Chonbuk 570-749, South Korea
e-mail: ykim@wonkwang.ac.kr; bird75@wonkwang.ac.kr

K.-W. Kim
Thunder Technology, Director in Digital Signer Processing Team, ChonJu, Korea
e-mail: kwkim@thunderspeaker.com

B.-G. Kim
Department of Electronic Engineering, Kunsan National University, Kunsan, Korea
e-mail: bgkim@kunsan.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_14, © Springer Science+Business Media Dordrecht 2014

primordial identification methods for fuzzy models. The general class of Sugeno-Takagi models [4] gave rise to more sophisticated rule-based systems. The designers find it difficult to develop adequate fuzzy rules and membership functions to reflect the essence of the data. Some enhancements to the model have been proposed by many researchers, there still remains the problem of finding "good" parameters of the fuzzy sets and of partitioning spaces in the rules.

In this paper, we introduce a non-fuzzy inference system based on hard scatter partition of input space using Linde-Buzo-Gray (LBG) Algorithm [5] that helps determine the rules of non-fuzzy model. The partitioned subspaces describe the rules. The consequence part of the rules is represented by polynomial functions. The proposed model is evaluated with numerical experimentation for nonlinear process.

14.2 Non-fuzzy Inference System

14.2.1 Premise Identification

The premise part of the non-fuzzy inference system is developed by means of the LBG algorithm [5]. The LBG algorithm is an iterative algorithm which requires an initial codebook. This initial codebook is obtained by the splitting method. An initial codevector is set as the average of the entire training dataset. This codevector is then split into two. The iterative algorithm is run with these two vectors as the initial codebook. The final two codevectors are splitted into four and the process is repeated until the desired number of codevectors is obtained. LBG algorithm is like a K-means clustering algorithm [6] which takes a set of input vectors S as input and generates a representative subset of vectors C as output according to the similarity measure. The algorithm is summarized below.

1. Input training vectors S

$$S = \{\mathbf{x}_i \in \mathbf{R}^d | i = 1, 2, \ldots, n\}. \tag{14.1}$$

2. Initiate a codebook C

$$C = \{\mathbf{c}_i \in \mathbf{R}^d | j = 1, 2, \ldots, k\}. \tag{14.2}$$

3. Set $D_0 = 0$ and let $k = 0$.

4. Classify the n training vectors into K clusters

$$\mathbf{x}_i \in S_q, \quad if \ \left\|\mathbf{x}_i - \mathbf{c}_q\right\|_p \leq \left\|\mathbf{x}_i - \mathbf{c}_j\right\|_p \quad for\ j \neq q. \tag{14.3}$$

5. Update cluster centers $\mathbf{c}_j$

$$\mathbf{c}_j = \frac{1}{|S_j|}\sum_{\mathbf{x}_i \in S_j} \mathbf{x}_i. \tag{14.4}$$

6. Set $k \leftarrow k + 1$ and compute the distortion D

$$D_k = \sum_{j=1}^{K}\sum_{\mathbf{x}_i \in S_j} \left\|\mathbf{x}_i - \mathbf{c}_j\right\|_p. \tag{14.5}$$

7. If $(D_{k-1} - D_k)/\, D_k > \mathrm{Q}$ (a small number), repeat steps 4 $\sim$ 6.

8. Output the codebook C.

14.2.2 Consequence Identification

The identification of the conclusion parts of the rules deals with a selection of their structure that is followed by the determination of the respective parameters of the local functions occurring there.

The conclusion is expressed as follows.

$$R^j:\ \textit{If}\ \ x_1\ \textit{and}\ \cdots\ \textit{and}\ x_d\ \textit{is}\ H_j \quad \textit{Then}\ y_j = f(x_1, \ldots, x_d). \tag{14.6}$$

Type 1 (Simplified Inference):

$$f = a_{j0} \tag{14.7}$$

Type 2 (Linear Inference):

$$f = a_{j0} + \sum\nolimits_{k=1}^{d} a_{jk}x_k \tag{14.8}$$

Type 3 (Quadratic Inference):

$$f = a_{j0} + \sum\nolimits_{k=1}^{d} a_{jk}x_k + \sum\nolimits_{k=1}^{d} a_{j,(d+k)}x_k + \sum\nolimits_{k=1}^{d}\sum\nolimits_{l=k+1}^{d} a_{jz}x_k x_l \tag{14.9}$$

Type 4 (Modified Quadratic Inference):

$$f = a_{j0} + \sum\nolimits_{k=1}^{d} a_{jk}x_k + \sum\nolimits_{k=1}^{d}\sum\nolimits_{l=k+1}^{d} a_{jz}x_k x_l \tag{14.10}$$

where R^j is the j-th rule, x_k represents the input variables, H_j is a membership grade (membership matrix) obtained using LBG algorithm, a's are coefficient of polynomial function, z is the number of combinations of input variables.

The calculations of the numeric output of the model, based on the activation (matching) levels of the rules there, are carried out in the well-known format

$$y^* = \frac{\sum_{j=1}^{n} w_{jp} y_j}{\sum_{j=1}^{n} w_{jp}} = \sum_{j=1}^{n} \hat{w}_{jp} y_j. \tag{14.11}$$

Here, as the normalized value of w_{jp}, we use an abbreviated notation to describe an activation level $\hat{w}_{jp}$, which values is determined by the membership matrix $\mathbf{U} = [u_{ip}]$;

$$w_{jp} = u_{ip}. \tag{14.12}$$

Therefore, the inferred output value y^* can be expressed as

$$y^* = \sum_{j=1}^{n} w_{jp} y_j. \tag{14.13}$$

If the input variables of the premise and parameters are given in consequence parameter identification, the optimal consequence parameters that minimize the assumed performance index can be determined. The minimal value produced by the least-squares method.

14.3 Experimental Studies

We discuss numerical example in order to evaluate the advantages and the effectiveness of the proposed approach. This time series data (296 input–output pairs) coming from the gas furnace nonlinear process has been intensively studied in the previous literature [7]. The delayed terms of methane gas flow rate $u(t)$ and carbon dioxide density $y(t)$ are used as input variables organized in a vector format as $[u(t-3), y(t-1)]$. $y(t)$ is the output variable. The first part of the data set (consisting of 148 pairs) was used for training purposes. The remaining part of the series serves as a testing data set. We consider the MSE as a performance index.

Table 14.1 summarizes the performance index for training and testing data by setting the number of clusters from 2 to 10 using LBG algorithm and inference type. Here, PI and E_PI stand for the performance index for the training data set and the testing data set, respectively. From the Table 14.1, linear inference, quadratic inference and modified quadratic inference shows better performance than simplified inference. We selected the best model with 5 rules (clusters) with quadratic inference that exhibits PI = 0.016 and E_PI = 0.284.

Figure 14.1 shows hard-partitioned input spaces and membership grades generated by LBG algorithm for the selected model. The boundaries of each local space indicate the binary boundary by the membership matrix with a value of 0 or 1.

Table 14.1 Performance evaluation for two input system

No. of clusters	Type	PI	E_PI
2	Type 1	3.459	4.104
	Type 2	0.023	0.335
	Type 3	0.022	0.337
	Type 4	0.022	0.348
3	Type 1	1.698	2.165
	Type 2	0.022	0.349
	Type 3	0.021	0.336
	Type 4	0.022	0.353
4	Type 1	1.015	1.973
	Type 2	0.021	0.343
	Type 3	0.019	0.350
	Type 4	0.021	0.335
5	Type 1	0.943	2.111
	Type 2	0.020	0.293
	Type 3	**0.016**	**0.284**
	Type 4	0.019	0.301
6	Type 1	0.904	2.078
	Type 2	0.020	0.294
	Type 3	0.015	0.284
	Type 4	0.019	0.302
7	Type 1	0.867	1.720
	Type 2	0.019	0.304
	Type 3	0.015	0.293
	Type 4	0.017	0.307
8	Type 1	0.849	1.895
	Type 2	0.017	0.290
	Type 3	0.014	0.287
	Type 4	0.016	0.300
9	Type 1	0.705	1.298
	Type 2	0.016	0.302
	Type 3	0.013	0.298
	Type 4	0.015	0.302
10	Type 1	0.611	1.214
	Type 2	0.016	0.304
	Type 3	0.012	0.302
	Type 4	0.014	0.302

Figure 14.2 shows the space plane of the input and output for selected models according to both the partitioning of the input space and the reasoning methods. The local spaces show independent input and output characteristics not to overlap with each other.

Figure 14.3 depicts the original and model outputs of training and testing data for the selected model. This figure shows that the model output is approximately

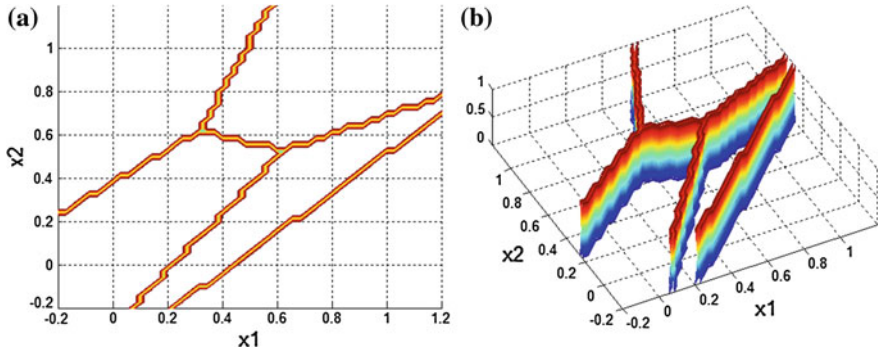


Fig. 14.1 Hard partitioned spaces using LBG algorithm (5 clusters, quadratic inference). **a** Hard partitioned spaces. **b** Membership grades

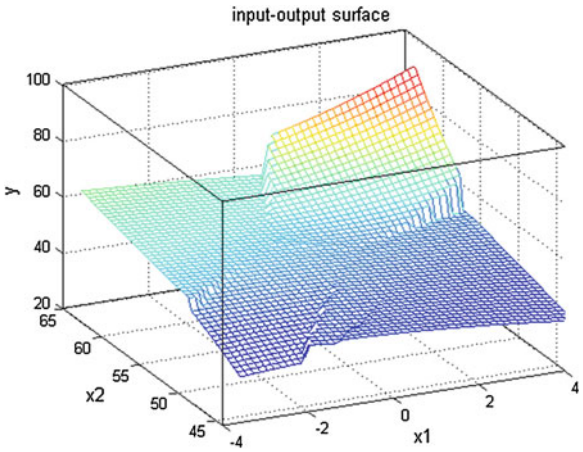


Fig. 14.2 Input–output spaces (5 clusters, quadratic inference)

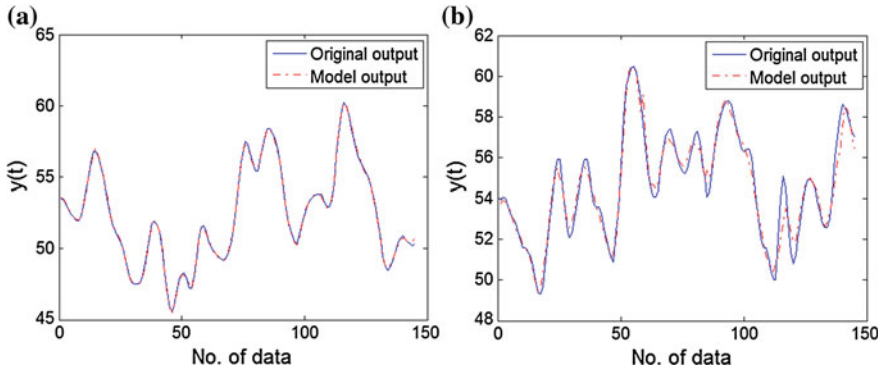


Fig. 14.3 Model outputs (5 clusters, quadratic inference). **a** Training data. **b** Testing data

the same for original output. Finally we conclude that the proposed model have better result than the traditional fuzzy inference systems that exhibits PI = 0.019 and E_PI = 0.305 in a similar environment.

14.4 Conclusions

We introduced a non-fuzzy inference system based on hard scatter partition of input space in this paper. The input spaces of the proposed model were divided as the scatter form using LBG algorithm to generate the rules of the system for nonlinear process. By this method, we could alleviate the problem of the curse of dimensionality and design the non-fuzzy inference systems that are compact and simple. From the results in the previous section, we were able to design preferred model that has faster computations and improved performance. Finally, this approach would find potential application in many fields. In the future, we are going to apply to the actual field data through partitioning strategies to fit the data structure.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2011835).

References

1. Jang JSR, Mizutani E, Sun T (1997) Neuro-Fuzzy and Soft Computing, a computational approach to learning and machine intelligence. Prentice Hall, NJ
2. Tong RM (1978) Synthesis of fuzzy models for industrial processes. Int J Gen Syst 4:143–162
3. Pedrycz W (1983) Numerical and application aspects of fuzzy relational equations. Fuzzy Sets Syst 11:1–18
4. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its applications to modeling and control. IEEE Trans Syst Cybern SMC-15(1):116–132
5. Linde Y, Buzo A, Gray RM (1980) An algorithm for vector quantizer design. IEEE Trans Commun 28:702–710
6. Krishnaiah PR, Kanal LN (eds) (1982) Classification, pattern recognition, and reduction of dimensionality, volume 2 of handbook of statistics. North-Holland Amsterdam
7. Box GEP, Jenkins GM (1976) Time series analysis: forecasting and control, 2nd edn. Holden-Day, San Francisco

Chapter 15
An Efficient Embedded System for the Detection and Recognition of Speed-Limit Signs

Hsing-Lung Chen, Ming-Sze Chen and Shu-Hua Hu

Abstract This paper is aimed at detecting and recognizing speed-limit signs in real-time. The recognition of speed-limit signs gives the necessary reminders and warnings to the drivers who may ignore the speed-limit signs. In the previous literatures, most of proposed schemes for real-time detection and recognition of traffic signs employed the software approach. It not only needs a large amount of data processing, but also requires a high demand of hardware. Therefore, we propose a hardware-software co-design scheme on FPGA, where regular computations are executed in parallel with hardware, resulting in detecting and recognizing speed-limit signs in real-time. The experiments show that our proposed scheme can detect and recognize a speed-limit sign in 62.558 ms. The detection rate is about 99 %. And the recognition rate is 100 %.

15.1 Introduction

To make road vehicles fully autonomous has been studied for many years. Different approaches have been employed: color segmentation, control theory, neural networks, etc. Obstacle detection and avoidance is an open research area. Automatic recognition of traffic signs studies started more recently, and are increasing rapidly.

H.-L. Chen (✉) · M.-S. Chen
Department of Electronic and Computer Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan
e-mail: hlchen@mail.ntust.edu.tw

M.-S. Chen
e-mail: m9702107@mail.ntust.edu.tw

S.-H. Hu
Department of Computer Science and Information Engineering, Jinwen University of Science and Technology, New Taipei, Taiwan
e-mail: shhu@just.edu.tw

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_15, © Springer Science+Business Media Dordrecht 2014

It could be used as an assistant for drivers, alerting them about the presence of some specific signs or some risky situations.

Escalera et al. [1] proposed a method for traffic sign detection and classification. They employ color threshold to segment the image and shape analysis to detect signs. Then, a neural network is applied for classification. Hirose et al. [2] proposed the *Simple Vector Filter* to distinguish between objects and background pixels. The genetic algorithm with search limits is proposed to realize a real-time position recognition. Janssen et al. [3] proposed a scheme for traffic sign recognition by using color, shape and pictogram. Martinovic et al. [4] proposed a method for automated localization of certain traffic signs. And the speed limit is determined with number character partition.

Most of previous approaches [5–9] detect the traffic signs first by locating the border of traffic signs and then extract the information within the traffic signs. They employed complicated chromaticity models which need a lot of computations. This paper employs CIE-rg chromaticity space which requires simple computations, making hardware realization feasible. Therefore, we propose a hardware-software co-design scheme on FPGA, where regular computations are executed in parallel with hardware, resulting in detecting and recognizing speed-limit signs in real-time.

15.2 Detection and Recognition of Speed-Limit Signs

15.2.1 CIE-rg Chromaticity Space

It is obvious that the most efficient method for detecting traffic signs in the picture is to employ color characterization. In this paper, we employ CIE-rg chromaticity space as color characterization because of its simple computations. CIE-rg values can be derived from RGB values by Eq. (15.1). The speed-limit signs have a white background and are surrounded by a red ring. Hence, the detection of speed-limit signs can employ the processing of red-color parts. From Fig. 15.1, it is observed that red colors are almost resided in the triangle enclosed by blue lines, which is satisfied with Eq. (15.2).

$$\begin{cases} r = R/(R+G+B) \\ g = G/(R+G+B) \end{cases} \tag{15.1}$$

$$\begin{cases} r \geq 0.329 \\ g \leq 0.324 \\ g \leq r - 0.341 \end{cases} \tag{15.2}$$

It is obvious that the rg values are ranging from 0 to 1. In order to reduce hardware resources, the required floating-point computations need to be transferred to integer computations. Hence, $r'g'$ values are introduced, which are rg

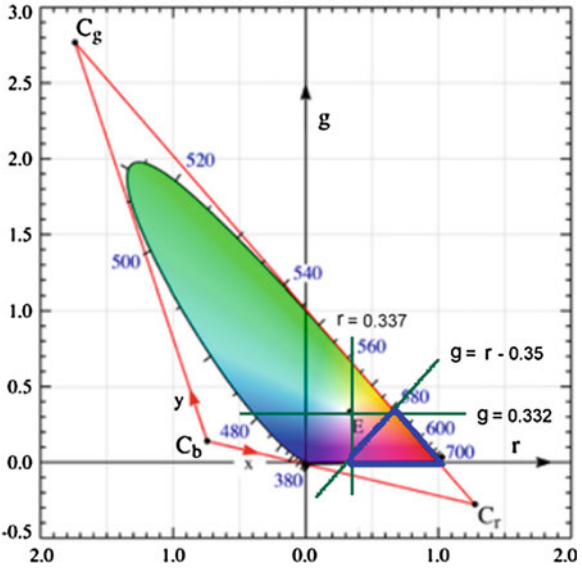


Fig. 15.1 CIE-rg chromaticity space

values multiplied by 1,024 as shown in Eq. (15.3). The pixels with red color can be identified by Eq. (15.4). With Eq. (15.4), the picture with RGB can be transferred to the binary image, where 1 represents red color and 0 non-red color.

$$\begin{cases} r^{'} = R * 1{,}024/(R+G+B) \\ g^{'} = G * 1{,}024/(R+G+B) \end{cases} \tag{15.3}$$

$$\begin{cases} r' \geq 377 \\ g' \leq 332 \\ g' \leq r' - 35 \end{cases} \tag{15.4}$$

15.2.2 Noise Eliminating and Boundary Smoothing

In this paper, the detection of speed-limit signs is conducted by recognizing the boundary of red-colored ring. It is important to eliminate noises and to smooth the boundary of red-colored regions. Here, 5 × 5 template is employed. Let X denote the center of 5 × 5 template and S the number of 1's in the template. By Eq. (15.5), noise eliminating and boundary smoothing can be conducted with different δ_l and δ_u. Note that X remains unchanged if $\delta_l < \mathrm{S} < \delta_u$.

$$\mathrm{X} = \begin{cases} 0, & \text{if} \quad S \leq \delta_l \\ 1, & \text{if} \quad S \geq \delta_u \end{cases} \tag{15.5}$$

Noise eliminating process is conducted gradually. At first, the process are conducted several times with $\delta_l = 6$ and $\delta_u = 19$, resulting in small noises being eliminated. Secondary, the process are conducted several times with $\delta_l = 8$ and $\delta_u = 17$ to remove big noises. Finally, the process are conducted several times with $\delta_l = 10$ and $\delta_u = 15$ to ensure that all noises are removed. Finally, boundary smoothing can be derived with $\delta_l = 11$ and $\delta_u = 14$.

15.2.3 Recognizing the Inner Boundary of the Red-Colored Ring

In this paper, assume that the picture is of size 800 × 400 pixels. The speed-limit signs with the diameter of their inner area larger than 40 pixels are considered. Hence, only rows 35, 70, …, 385 need to be scanned for searching all the speed-limit signs. The scanning begins with the left end of each scan-line. The next scan-line is going to be searched when the right end of the current scan-line is reached, as shown in Fig. 15.2a–c. Let pixel (L_x, L_y) represent the second visited boundary pixel with red color and (S_x, S_y) the third visited boundary pixel with red color. When (S_x, S_y) is reached, then scan the boundary clockwise. The case that (S_x, S_y) is in the outer boundary is considered first, as shown in Fig. 15.2d. (S_x, S_y) will be reached again and the scanning is restarted from $(S_{x\text{-}1}, S_y)$, as shown in Fig. 15.2e. In the other case, (L_x, L_y) will be reached first and the inner boundary is found, as shown in Fig. 15.2f.

15.2.4 Extracting Boundary Information

The vector from the currently visiting pixel to the next visiting pixel is with one of eight degrees 0°, 45°, …, 315°. The vector codes 0, 1, … 7 denote degrees 0°, 45°, …, 315°, respectively. The labeling of adjacent pixels of the currently visiting pixel is shown in Fig. 15.3. As described before, the scanning of the inner boundary is conducted clockwise. Initially, the base pixel is with red color. Pixel(0) must be with red color because of eliminating noises. Then we search the adjacent pixels from pixel(0) clockwise until the first adjacent pixel with non-red color pixel(j) is visited. The vector code ((j + 1) mod 8) is derived and the scanning is moved to pixel ((j + 1) mod 8). In the following, the vector code is derived similarly except for the selection of the starting adjacent pixel. Since the smoothing has been conducted, the difference between the current vector code and the last vector code is at most 1. Supposed that the last vector code is i, the starting adjacent pixel is set to be pixel ((i + 1) mod 8).

The quasi-curvature of the segment is defined as the difference between the sum of vector directions of the first half and that of the second half. For simplicity, the

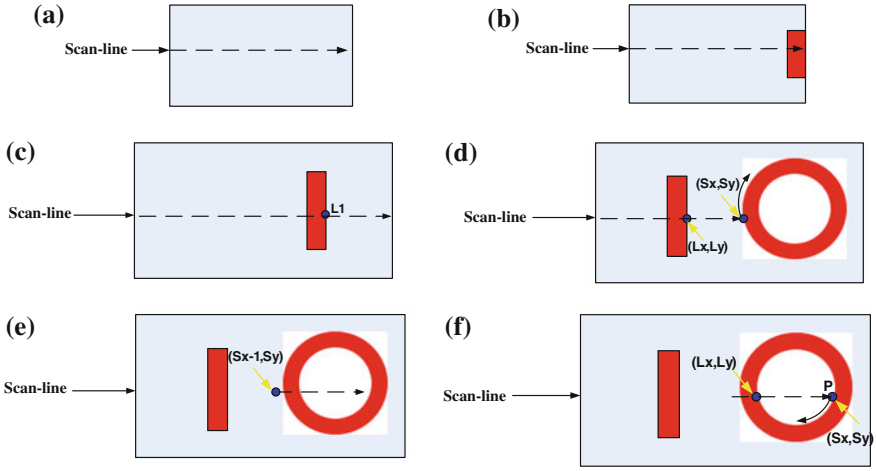


Fig. 15.2 Recognizing the inner boundary of the *red-colored ring*

Fig. 15.3 Labeling adjacent pixels

pixel(3)	pixel(2)	pixel(1)
pixel(4)	Base	pixel(0)
pixel(5)	pixel(6)	pixel(7)

vector code is employed as the vector direction. The length of the segment (2n) will be described in the next section. Supposed that the first half consists of vectors F_1, F_2, …, F_n, and the second half consists of vectors S_1, S_2, …, S_n. Without ambiguity, vector representation is used for vector code. The quasi-curvature of the segment can derived by Eq. (15.6). The difference of two vectors is difficult to calculate correctly. However the difference of two adjacent vectors is simple to calculate correctly. $(F_1 - S_1)$ can be reformulated as n differences of adjacent vectors, i.e., $(F_1 - F_2) + (F_2 - F_3) + \ldots + (F_n - S_1)$. Other subtractions in Eq. (15.6) can be reformulated similarly.

$$\begin{aligned} C &= (F_1 + F_2 + \ldots + F_n) - (S_1 + S_2 + \cdots + S_n) \\ &= (F_1 - S_1) + (F_2 - S_2) + \cdots + (F_n - S_n) \end{aligned} \tag{15.6}$$

The segment with the quasi-curvatures −1, 0, or +1 is defined as the straight segment (π). The segment with the quasi-curvature less than −1 is defined as the clockwise curve segment (ρ). Similarly, the segment with the quasi-curvature greater than +1 is defined as the counter-clockwise curve segment (φ). If the inner boundary consists of ρ or $\rho\pi\rho\pi$, a speed-limit sign is detected.

15.2.5 Recognizing Speed

The speed-limit sign has black numbers with white background. The first step is to derive the histogram of gray values of the inner area. The suitable threshold is determined from the derived histogram. The binary image thus can be generated by the determined threshold. By the method described in Sect. 15.2.2, noise elimination and boundary smoothing are applied on the binary image.

From the observations of numbers, the length of the segment (2n) is determined by the length and width of the character as follows.

$$2n = \min(\text{width} \times 0.6, \text{length} \times 0.4) \tag{15.7}$$

From previous subsection, the boundary can be decomposed into several straight segments, clockwise curve segments and counter-clockwise curve segments. However, some recognized straight segments may not be straight, because their first half and second half are symmetric. The recognized straight segments must be verified by various segment lengths $((\text{sub} = 2 \times \lfloor n/2 \rfloor^i)$, for $\text{sub} > 2$. Each number character can be coded with a sequence of segment types starting from the top. Each number character may have several codes with various patterns. With all of these codes, the search tree is established. In the recognition process, the derived code is matched through the search tree for recognition.

15.3 System Design

In this paper, we use DE2-70 FPGA board supported by Altera. It contains an embedded processor NIOS II and a CMOS sensor. We need to configure data transfer module and processing module through FPGA. Data transfer module is responsive for data transmission between NIOS II and processing module. Processing module conducts the generation of binary red image, noise elimination and smoothing, which require a lot of computations. Other operations are done by software with NOIS II (Fig. 15.4).

15.4 Performance Evaluation

We have collected 104 pictures. In one of pictures, the red ring of the speed-limit sign is partly covered by tree leaves. The speed-limit signs in 103 pictures are detected correctly. One traffic sign covered by tree leaves can not be detected because the red ring is not recognized. Thus, the detection rate is 99 %. The speed-limit in 103 pictures are recognized correctly. Hence, the recognition rate is 100 %.

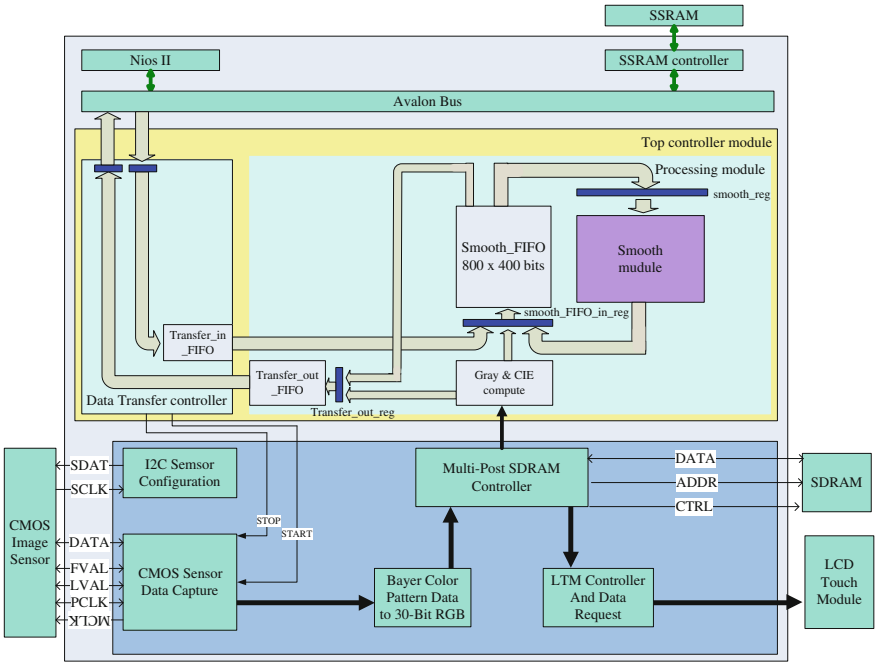


Fig. 15.4 System architecture

In DE2-70, it only needs 62.558 ms totally to recognize a picture, resulting in meeting real-time requirement. Data transfer takes 15.25 ms. Software processing in NOIS II takes lot of times (47.308 ms). Noise elimination and smoothing with hardware costs a little time (0.544 ms), which can be executed in parallel with the data transfer.

15.5 Conclusion

This paper proposes a hardware-software co-design scheme on FPGA, where regular computations are executed in parallel with hardware, resulting in detecting and recognizing speed-limit signs in real-time. The experiments show that our proposed scheme can detect and recognize a speed-limit sign in 62.558 ms. The detection rate is about 99 %. And the recognition rate is 100 %.

The boundary detection algorithm can be enhanced to recognize the border of traffic signs covered by tree leaves. Extending to most of traffic signs is worthy to be investigated in the future.

Acknowledgments This work was support, in part, by National Science Council, Republic of China, under Grant No. NSC 102-2221-E-011-067.

References

1. de la Escalera A, Moreno LE, Salichs MA, Armingol JM (1997) Road traffic sign detection and classification. IEEE Trans Ind Electron 44:848–859
2. Hirose K, Asakura T, Aoyagi Y (2000) Real-time recognition of road traffic sign in moving scene image using new image filter. In: 26th annual conference of the IEEE Ind Electron Soc (IECON 2000) 3, Nagoya, pp 2207–2212
3. Janssen R, Ritter W, Stein F, Ott S (1993), Hybrid approach for traffic sign recognition. In: Intelligent vehicles symposium '93, pp 390–395
4. Martinovic A, Glavas G, Juribasic M, Sutic D, Kalafatic Z (2010), Real-time detection and recognition of traffic signs. In: The 33rd international convention MIPR0 2010 Opatija Croatia, pp 760–765
5. Ach R, Luth N, Techmer A (2008) Real-time detection of traffic signs on a multi-core processor. In: 2008 IEEE intelligent vehicles symposium, Eindhoven, The Netherlands pp 307–312
6. Estable S, Schick J, Stein F, Janssen R, Ott R, Ritter W et al (1994), A real-time traffic sign recognition system. In: Intelligent vehicles symposium '94, pp 213–218
7. Lafuente-Arroyo S, Maldonado-Bascon S, Gil-Jimenez P, Gomez-Moreno H (2008) An intra-image tracking algorithm for traffic sign recognition. In: IEEE international conference on vehicular electronics and safety, 259–264
8. Maldonado-Bascon S, Lafuente-Arroyo S, Siegmann P, Gomez-Moreno H, Acevedo-Rodriguez FJ (2008) Traffic sign recognition system for inventory purposes, In: IEEE intelligent vehicles symposium, pp 590–595
9. Muller M, Braun A, Gerlach J, Rosenstiel W, Nienhuser D, Zollner JM et al (2010) Design of an automotive traffic sign recognition system targeting a multi-core SoC implementation. In: Design, Automation & Test in Europe Conference & Exhibition (DATE), Dresden, pp 532–537

Chapter 16
Future Enterprise Computing Looking into 2020

Yung-Chin Fang, Yuxiang Gao and Cindy Stap

Abstract This study describes today's Enterprise computing architecture from the type of application, operating environment, application requirements, application topologies, resources requirements and storage architectures standpoints. Leveraged historical facts were extrapolated to construct future enterprise computing key components, and apply the applications to future computing components so as to uncover future enterprise computing challenges. A new CPU, converged space, App store, deployment, and resources allocation challenges are addressed in this study, as well as the need for an enterprise computing level pervasive engine and dashboard.

Keywords Enterprise computing · SAN · LAN · Converged infrastructure · Segregated networking · DAS · NAS

16.1 Introduction to Enterprise Computing Complexity

Enterprise computing is comprised of applications and infrastructure. Application is for problem solving; infrastructure is for applications hosting. Enterprise applications fall into eight categories: business processing such as customer relationship management (CRM), IT infrastructure such as a shared file/print server, decision support systems such as a business intelligence and analytic,

Y.-C. Fang (✉) · Y. Gao · C. Stap
Global Solutions Engineering Group, Dell Inc, Austin, USA
e-mail: Yung-Chin_Fang@Dell.com

Y. Gao
e-mail: Shawn_gao@Dell.com

C. Stap
e-mail: Cindy_Stap@Dell.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_16, © Springer Science+Business Media Dordrecht 2014

application development such as an integrated development platform, collaborative such as a mail/message server, web infrastructure such as a store front, industry R&D such as high performance cluster computing (HPCC), and internet computing such as cloud computing. Enterprise computing infrastructure consists of generations of heterogeneous architectures such as physical and virtualization frameworks, both centralized and distributed, dedicated and shared, segregated and converged and hybrid. Enterprise class solution is comprised of a set of application specific design concepts and select HW/SW components. Each application's infrastructure consists of server infrastructure such as monolithic or blade infrastructure, management framework such as in-band and out of band, interconnection technology such as Ethernet and Fibre Channel, application specific topology such as fat tree, application HW/FW support matrix such as switch and storage firmware versions, and application capability implementations such as software level availability. The IT professional's job is to design enterprise computing operating environment along with deploying, monitoring, managing and maintaining the applications and operating environment. Most enterprise class applications have specific requirements such as fabric design and response time, etc. For example: a mail server application may require 4 Ethernet fabrics: public LAN, private LAN (heart beat can reside on the same fabric), a storage area network (SAN) fabric and an optional fabric for backup. A high performance computing cluster may require 56 Gbps Infiniband and non-blocking fabric. Mail servers may require guaranteed storage session response time such as 10 ms for writing and 20 ms for reading. High performance computing applications usually require a nanosecond-level interconnect latency rather than guaranteed storage read/write session time. Different applications can benefit from different storage architecture; for example, direct attached storage (DAS) provides best performance/cost ratio for certain mail servers, local storage and network attached storage (NAS) can satisfy most of the file/print server's needs, SAN storage can meet requirements such as storage latency from Oracle applications, and virtual SAN storage such as Microsoft storage spaces can satisfy the requirements of certain virtual desktop infrastructure (VDI) applications (Fig. 16.1).

IT consists of generations of heterogeneous applications and operating environment from venders. Besides the complexity derived from applications and the operating environment, practices such as business continuity/availability, security, expandability, application/operating environment lifecycle manageability and serviceability, and quality of service (QoS) tuning etc. in addition to considerations like the adoption of government regulations such as the SOX act, corporate policies such as the bring-your-own-device (BOYD), new technologies such as dynamic datacenter, new business models such as globalization, operation models such as just in time (JIT), and new R&D approaches such as Agile process all add complexity to enterprise computing. The complexity of enterprise computing is high. When piling necessary application requirements, general practices and considerations onto the same plate, deploying, monitoring, managing and maintaining a premise enterprise computing environment efficiently becomes a challenge. Cloud computing introduces remote, on demand resources to enterprise

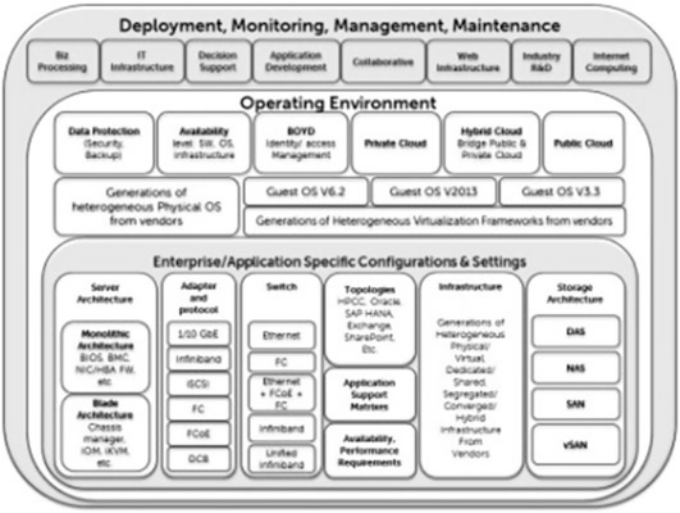


Fig. 16.1 Enterprise architecture

computing; this change extends the lifecycle manageability scope from on premise to the hybrid model and makes the single pane of glass manageability become even more complex and convoluted.

The fundamental components for enterprise computing are: server, network, storage and operating system and hypervisor.

16.2 Enterprise Computing Component Trends

Server Trends. The main components of a server are the CPU, memory, interconnects and local storage. Moore's law and CPU evolution history indicates two trends: higher performance CPU and higher level of System on Chip (SoC) integration. A higher performance CPU trend suggests a higher core count, cache capacity and perhaps clock rate. For instance, top bin Ivy Bridge CPU has 50 % more cores and 50 % more cache when compared with prior generation Sandy Bridge. Tick Tock is a wide known processor development model, advancing manufacturing technology such as introducing 3D tri-gate transistors with 22 nm fab technology. Tock conveys new processor microarchitecture such as video transcoding, encryption and decryption, etc. A conservative extrapolation suggests that by 2020, the mainstream mid bin general purpose CPU will have 40 core and 60 MB of cache (Figs. 16.2, 16.3, 16.4).

The SoC trend implies that more components will be packaged into the SoC; for example, Calxeda EnergyCore packaged a quad core ARM processor, a management engine, 4 MB L2 cache, five 10 Gigabit XAUI, fabric switch, PCIe

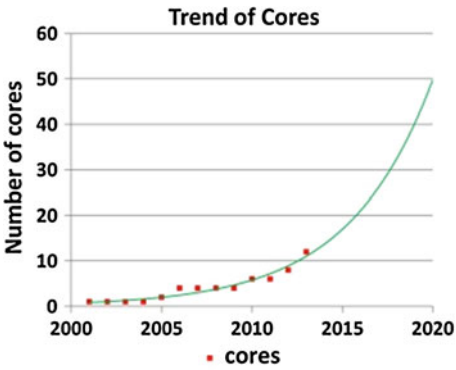


Fig. 16.2 Intel Xeon cores

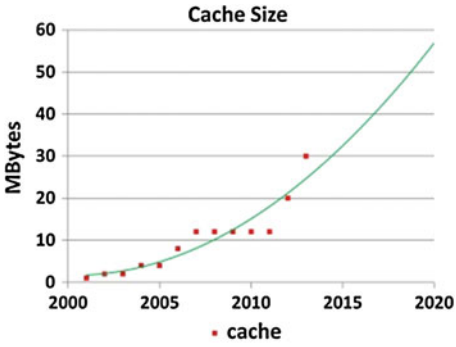


Fig. 16.3 Intel Xeon cache

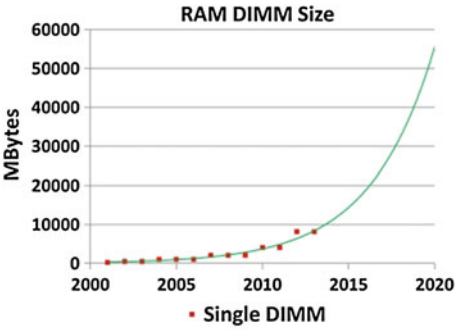


Fig. 16.4 Memory trend

interface, SATA interface, SD/eMMC controller, DDR3 memory controller and UART/I2C interface into a single SoC chip. An EnergyCore based computer consists of 1 SoC, 1 storage connector, 1 storage device and 1 memory DIMM. If SoC fab also follows Moore's law, it will have 8 times as many transistors by 2020. Moreover, it is possible to integrate enough memory and solid state disk (SSD) into the same chip packaging thus SoC based single chip server by 2020 won't be a surprise. However, this trend can have a converged space twist.

Memory is a type of semiconductor, and memory chip capacity growth also follows Moore's law. By 2020, a 64 Gbit memory chip will be mainstream, along with a 1.5-terabyte memory server, if the server architecture remains synonymous.

Interconnect technology has two technology trends; a higher data transfer rate converge and support protocols on the same infrastructure. General purpose network interface data rate has already evolved from 1, 10, 40 to 100 Gbps. IEEE 802.3's 400 Gb/s Ethernet study group is developing a 400 Gbps standard, deliverable is targeting 2007. Certain companies such as Google and Facebook expressed their need for Terabit Ethernet. 10 GbE NIC is widely used in enterprise computing today, so a 40 GbE NIC will be mainstream by 2020. The best practice for an enterprise Gigabit Ethernet based network design is to segregate SAN/LAN traffic, as they have dedicated and particular fabrics, thus, the advantage is that high LAN traffic will not affect SAN fabric performance, though this model does require two separate fabrics. When fabric availability/redundancy is considered, the design becomes two SAN fabrics plus two LAN fabrics associated with deploy, monitor, management and maintenance efforts. In order to reduce the amount of fabric and network lifecycle management effort, the converged network adapter (CNA) was created. CNA is designed to support 10 Gigabit Ethernet, iSCSI, and FCoE frames. Single NIC/cable carries both SAN and LAN traffic. In order to provide guaranteed bandwidth to network partition, data center bridging (DCB) standard was created, ensuring lossless traffic. DCB requires priority-based flow control (PFC, IEEE 802.1Qbb), enhanced transmission selection (ETS, IEEE 802.1Qaz) and data center bridging exchange (DCBX, IEEE 802.1Qaz). CNA and DCB implementation reduced server interconnect to half.

SSD is, likewise, a type of semiconductor and consists of layers of storage chips, each layer having rows of storage chips, with multiple chips in each row. Moore's law can be revised to the 3D Moore's law for describing SSD capacity growth. SSD interface evolved from SATA, SATA express, SAS6, SAS12 to non-volatile memory express (NVMe). Samsung XS175 NVMe 1.6 TB SSD performance is around 3,000 MB throughput and 740 K IOPS as of 2013. Mainstream SSD by 2020 will have 12.8 TB of capacity. Hard disk device (HDD) capacity grows 25–40 % annually. Therefore, 3D Moore's law is applicable to SSD capacity growth and will replace HDD's storage position by 2020. From the RAID technology standpoint, RAID-10 was created to provide aggregate spindle IO performance and availability, when SSD performance is put into consideration, RAID5 or RAID6 will become the mainstream configuration for enterprise computing (Figs. 16.5, 16.6, 16.7).

SSD is connected to PCI Express directly, which is expected to deliver 32 Gigatransfers per second by 2020. In a dual CPU server, each CPU connects to a PCI Express directly, the two SSD providing around 64 Gigatransfers per second. Traditional tiered memory/storage architecture is due to slow storage and cost considerations; when SSD is fast, large and cheap enough, SSD will be able to replace what is currently memory's position to form a converged space. This space serves as both memory and storage for computing. A mainstream mid-end performance server by 2020 can encompass 80 cores, a 120 MB cache, a 25.6 TB

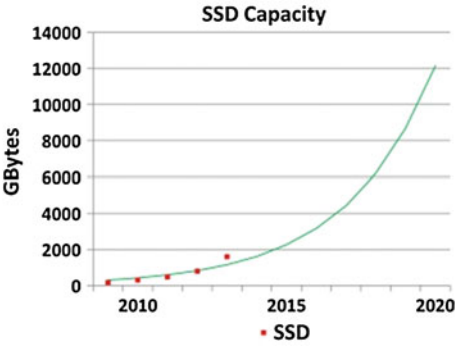


Fig. 16.5 SSD capacity

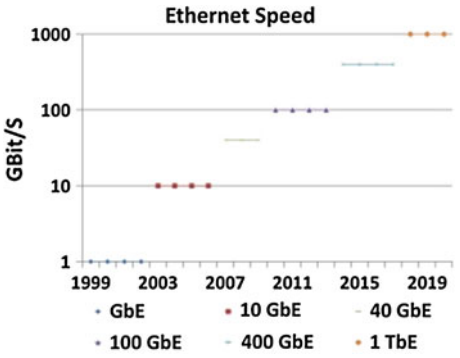


Fig. 16.6 Ethernet speed

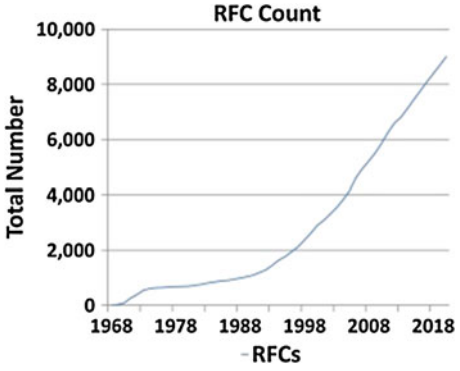


Fig. 16.7 RFC count

of converged space (or 1.5 TB of RAM), dual 40 GbE CNAs, and optional 60 SSD (connected via SATA or SAS interface). A main stream SoC server in 2020 could be a single chip containing said converged space or integrated RAM and SSD.

External Storage Trends. Certain enterprise class infrastructure applications require external storage systems to provide guaranteed performance, capacity and availability. Guaranteed performance and capacity are usually achieved via

dedicated external storage. Common practices are: storage vMotion for virtual machine disk level availability, RAID for disk device level availability, replica for enclosure level availability, and disaster recovery for site level availability. When a server is constituted of space and 60 high-capacity SSDs, part of the demand for external storage will shift to the server. Operating system and hypervisor now support virtual SAN capability, which is to aggregate local disks in servers to form a virtual SAN, this trend also shifts some demands from external storage to virtual SAN/servers. Certain enterprise class applications improved their architecture to support distributed storage, running on two or more servers, each with ample storage, and multiple copies of the same data, which are distributed across servers for example, an active data partition for hot data, a passive data partition for failover, a backup data partition for recovery and a hot spare partition for bad disk. The demand from enterprise applications which require dedicated external block level storage probably will remain the same.

Network Trends. New standards will be defined for large scale infrastructure services as well as software services. An exemplification of this is that, the 4096 vLAN defined under IEEE 802.1Q is simply not enough for cloud service providers such Amazon AWS and MS Azure, thus shortest path bridging is created to support up to 16 million vLAN. Large scale infrastructure software services, for example, virtualization technology are widely used by IaaS/PaaS providers. VXLAN enables virtual machines to create a logical network across several different networks. More user-defined network objects will be defined for special needs such as software defined network. Over 6,800 requests for comments (RFC) are submitted to IETF as of 2013. Extrapolation suggests more than 9,000 RFC by year 2020. Summing up these standards and implementations, just to learn and understand these changes requires a good size team, changes of this magnitude make enterprise network administration become an extremely challenging task.

16.3 Future Enterprise Computing Challenges

More applications on shared servers create new challenges. An enterprise usually runs thousands of applications. The App store is usually implemented to store and deploy applications. Enterprise applications usually support certain automated installation mechanisms such as PXE, iSCSI and FC boot. Enterprise applications usually require multiple fabrics with guaranteed storage performance. Challenges for future App store are: identify the application installation mechanism, deploy said application to select hardware and allocate resources accordingly without affecting other in-production application's performance. For example, an application (say, a virtual machine) supports a ISO image with iSCSI boot, and the infrastructure is DCB based, then the App store needs to have the intelligence to select the iSCSI device with iSCSI TLV feature for guaranteed storage throughput. The other App store challenge is: each application has its specific operating environment requirements; different applications require different OS, Java, driver

versions and BIOS settings, etc. How to deploy applications to a shared environment automatically without causing compatibility issue remains a challenge. For example: R&D applications usually require turning off BIOS virtualization support for extreme performance, it's not feasible to deploy a R&D application to hypervisors. Each application has its specific performance requirement, making guaranteed performance on a shared computing infrastructure quite formidable, not to mention unpredictable runtime payload. Enterprise application usually has multiple roles/processes, which in turn have inter-process and intra-process communication; application role and CPU/core affiliation can affect performance largely, deploying multi-role applications with the best fit CPU/core affiliation for optimal performance is quite demanding. Due to the number of cores, the growing number of hybrid programming based applications, application/CPU/core affiliation become more important than ever. How to satisfy the application/server/CPU/core affiliation for best performance is a challenge. Converged space proposes new operating system architecture. Locating the converged space for multiple applications dynamically for most efficient storage space utilization is also a strenuous affair. Enterprise application usually requires guaranteed storage response time; allocating shared storage space (in local disk, converged space or external storage) to multiple applications whilst maintaining the guaranteed IO throughput/latency is yet another conundrum.

Other challenges include: CIO having to use many management consoles to oversee application, IT health condition and resource utilization to support the business continuity and IT operation growth. The number of consoles is growing rapidly, and CIO requires a configurable pervasive engine and dashboard to support the business need. Besides premise computing, the pervasive engine/dashboard also needs to extend the visual from on premise to the hybrid cloud arena as well as support enterprise computing optimization.

References

1. Hilbert M et al (2011) The world's technological capacity to store, communicate, and compute information. Science 332:60. doi:10.1126/science.1200970
2. NVMe SSD Performance http://www.samsung.com/global/business/semiconductor/news-events/press-releases/detail?newsId=12961
3. Feldman M (2010) Facebook dreams of terabit ethernet. HPCwire, 3 Feb 2010

Chapter 17
Implementation of a Multi-core Prototyping System for a Video-Based Fire Detection Algorithm

Yong-Hun Park, Myeongsu Kang and Jong-Myon Kim

Abstract As the demand of an automatic fire detection system is increasing in recent years, a vision-based fire detection system is appealing. However, vision-based fire detection algorithms require tremendous computational performance, limiting their use in real-time applications. This paper proposes a multi-core prototyping system to support these high computational algorithms. The multi-core architecture including 16 processing elements (PEs) is implemented on a vertex 4 FPGA chip of the HUINS SoC Master 3 board. In addition, the performance of a selected four-stage fire detection algorithm is evaluated using the multi-core prototyping system. Experimental results show that the proposed multi-core prototyping system executing at 50 MHz clock frequency supports seven frames per second for a 240 × 200 resolution video. In addition, the proposed multi-core system achieves about 1.87 × speedup over commercial high-performance TI DSP operating at 720 MHz clock frequency.

Keywords Multi-core architecture · Fire detection algorithm · FPGA · Prototyping system

17.1 Introduction

Fire detection becomes increasingly appealing because of its important application in surveillance systems [1–3]. It is very critical for the personal security and commercial applications. A number of methods have been proposed in previous

Y.-H. Park · M. Kang · J.-M. Kim (✉)
School of Electrical Engineering, University of Ulsan, Ulsan, South korea
e-mail: jmkim07@mail.ulsan.ac.kr; jongmyon.kim@gmail.com

Y.-H. Park
e-mail: ase018@mail.ulsan.ac.kr

M. Kang
e-mail: ilmareboy@mail.ulsan.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_17, © Springer Science+Business Media Dordrecht 2014

researches to detect fire. However, most of these methods require a close proximity to the source of the fire and are based on particle sensors. Therefore, they cannot detect fire in open or large spaces and cannot provide additional information regarding to the process of burning. In order to overcome these drawbacks, video-based fire detection systems are considered as the most suitable candidates to replace the conventional fire detection techniques.

A number of vision based fire detection systems have been introduced with the rapid development in digital-camera technology and contents-based video processing [4]. Most of these algorithms are based on the color pixel recognition and/or motion detection [5–7]. These vision based fire detection algorithms, however, require tremendous computational and I/O requirements, prohibiting their use in real-time applications.

Application-specific integrated circuits (ASICs) can meet high performance and low power consumption for the vision based fire detection system. However, they provide limited, if any, programmability or flexibility needed for varied application requirements. General-purpose microprocessors (GPPs) or digital signal processors (DSPs) offer the necessary programmability and flexibility for various applications. However, neither GPPs nor DSPs will be able to meet the much higher levels of performance required by emerging fire detection applications because of lacking ability to exploit the parallelism available in these applications. Among many computational models available for such image processing based applications, multi-core architecture is a promising candidate [8–11]. Whereas instruction-level or thread-level processors use silicon area for large multi-ported register files, large caches, and deeply pipelined functional units, the multi-core architecture contains simple low cost processing elements (PEs) for the same silicon area.

This paper presents an implementation of a selected fire detection algorithm on the multi-core architecture which consists of 16 processing elements (PE), where the selected fire detection algorithm consists of 4 stages: (1) detecting moving regions using an approximately median method, (2) color segmentation to distinguish between moving regions and fire regions, (3) feature extraction for efficient fire detection, and (4) classifying into fire or non-fire using the generic-based back propagation neural network (GBPNN) with extracted features. In addition, the hardware of the multi-core architecture is implemented and verified on a FPGA prototyping system. The proposed multi-core prototyping system executing at 50 MHz clock frequency supports seven frames per second for a 240 × 240 resolution video and outperforms the commercial TI DSP which is operated at 720 MHz clock frequency in terms of execution time.

The rest of this paper is organized as follows. Section 17.2 presents the reference multi-core architecture and its prototyping system for supporting the selected fire detection algorithm. Section 17.3 presents parallel implementation of the reference fire detection algorithm, and Sect. 17.4 shows synthesis results and performance of the multi-core prototyping system. Section 17.5 concludes this paper.

17.2 A Multi-core Prototyping System

17.2.1 The Reference Multi-core Architecture

Figure 17.1 shows a reference multi-core architecture used in this study. It consists of 2-dimensional processing element (PE) array, local memory, and array control unit (ACU). PEs in the array are interconnected in a mesh network and they execute a set of instructions in a lockstep fashion. Each PE in different PE configurations supports a different size of pixels and local memories to store input images and temporary data produced during processing. In addition, each PE has a reduced instruction set computer (RISC) datapath with the following characteristics:

- Small amount of 32-bit word local memory,
- 16 32-bit three-ported general-purpose registers,
- ALU: computes basic arithmetic and logic operations,
- Barrel shifter: performs multi-bit logic/arithmetic shift operations,
- MACC: multiplies 32-bit values and accumulates into a 64-bit accumulator,
- Sleep: activates or deactivates a PE based on local information, and
- Nearest neighbor communications through a NEWS (north–east–west–south) network and serial I/O unit.

17.2.2 A Multi-core Embedded System

Figure 17.2 shows a multi-core prototyping system that consists of a multi-core architecture, camera controller, LCD controller, and direct memory access unit. The camera controller transfers image data from image memory to a pixel memory of each PE using a memory write operation. The LCD controller reads data from a resultant memory using a memory read operation. The prototyping system is executed as follows.

Step 1: The camera controller transfers pixel data from the CMOS camera to image memory using a memory write operation.
Step 2: DMA accesses 2-dimensional (2D) image data and transfers these 2D image data to each PE.
Step 3: ACU brings both scalar and vector instructions from the program memory and distributes the vector instructions to each PE.
Step 4: The target program is done and the result is stored at a memory.
Step 5: The LCD controller transfers the resultant data to LCD.

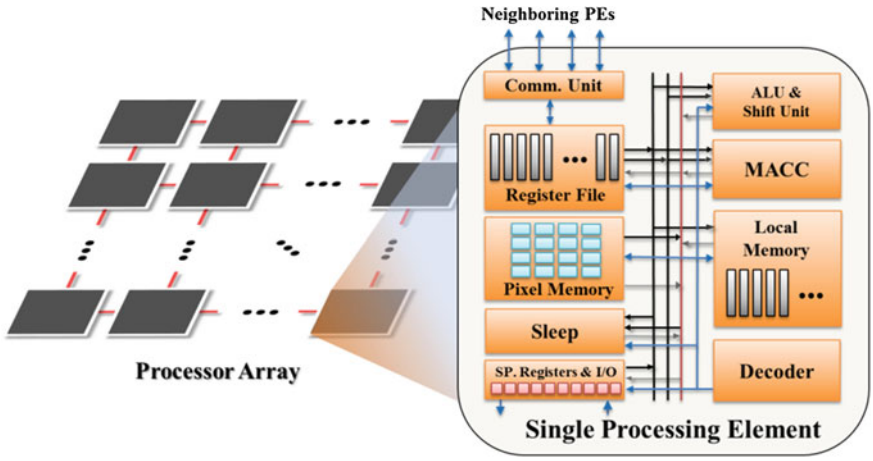


Fig. 17.1 The reference multi-core array and a single processing element

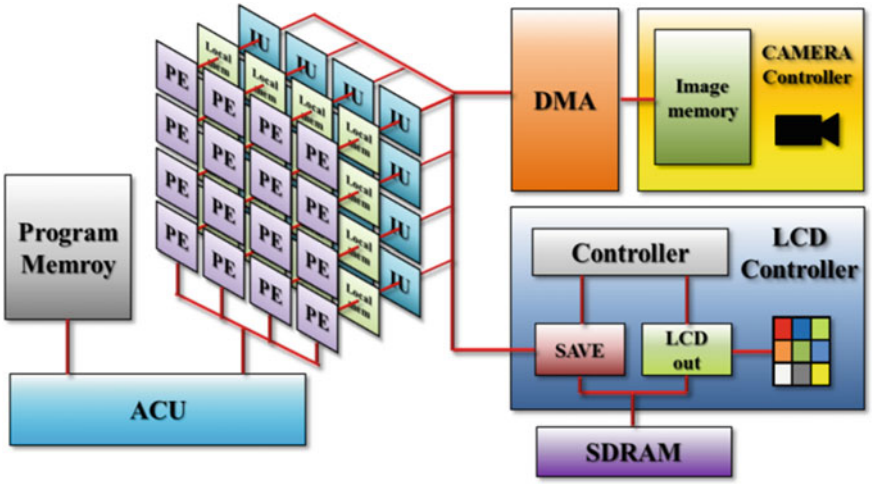


Fig. 17.2 A multi-core embedded system

17.2.3 System Parameters

We implement our multi-core architecture on a vertex 4(xc4vlx60) FPGA chip of the HUINS SoC Master 3 board, as shown in Fig. 17.3. Table 17.1 describes the multi-core system parameter. The reference multi-core architecture includes 16 PEs and 8 and operates at 50 MHz. In addition, local memory size and pixel memory size per PE is 8,192 and 3,000 words, respectively.

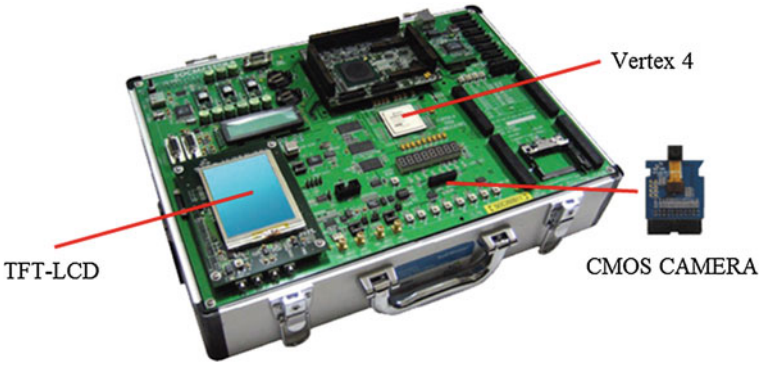


Fig. 17.3 HUINS SoC master 3 board used for the implementation of the multi-core architecture

Table 17.1 System parameter of the reference multi-core architecture

Parameter	Value
System size	16 PEs
Pixel per PE	60 × 50 (3,000)
Image size	240 × 200 pixels
Clock frequency	50 MHz
Interconnection network	Mesh
Int ALU/Int MUL/Barrel shift/Int Macc/Int Comm	1/1/1/1/1/1
Local memory size per PE	8,192 word

17.3 Parallel Implementation of a Fire Detection Algorithm

This paper implements a selected fire detection algorithm on the reference multi-core system. Figure 17.4 illustrates a flow diagram for the four-stage fire detection algorithm: moving region detection, color segmentation, feature extraction, and classification of fire or non-fire using a back propagation neural network (BPNN).

For moving region detection, the approximate median method is performed in each PE. Background information is then obtained by subtracting the current frame from the previous frame, and it is stored in the local memory of each PE. After then, feature extraction is performed only on the moving regions by activating/deactivating PEs to extract features. Figure 17.5 demonstrates the activation/deactivation process.

In the next stage, color segmentation is executed as illustrated in Fig. 17.6. It essentially includes the following three stages as described in Fig. 17.6.

For efficient fire detection, it is required to extract features of fire. Thus, feature extraction is performed by using two-level wavelet decomposition. The extracted

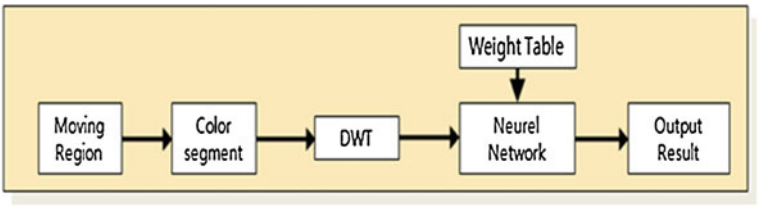


Fig. 17.4 A flow diagram of the four-stage fire detection algorithm

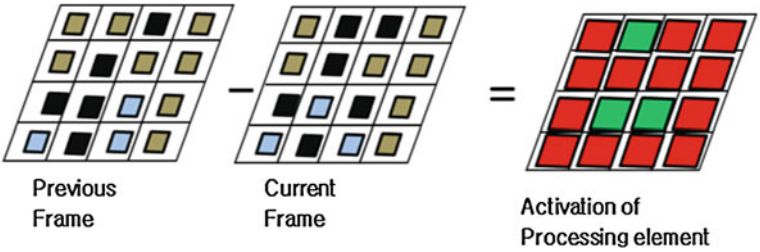


Fig. 17.5 Implementation of moving regions detection on the reference multi-core architecture

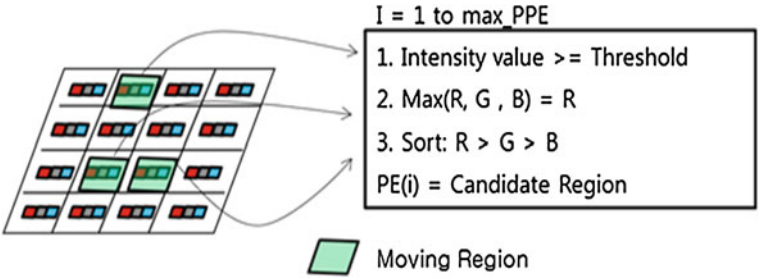


Fig. 17.6 Color segmentation at the candidate moving region

features are then used as an input vector of the BPNN. Figure 17.7 presents the extraction process on the multi-core architecture.

Finally, the BPNN determines fire or non-fire on the multi-core architecture, where each PE is mapped to the 9-14-1 topology of BPNN as shown in Fig. 17.8. We use connection weight values which are previously calculated by the BPNN, which reduces running time to train the BPNN. The log-sigmoid activation function is implemented by using a look-up table.

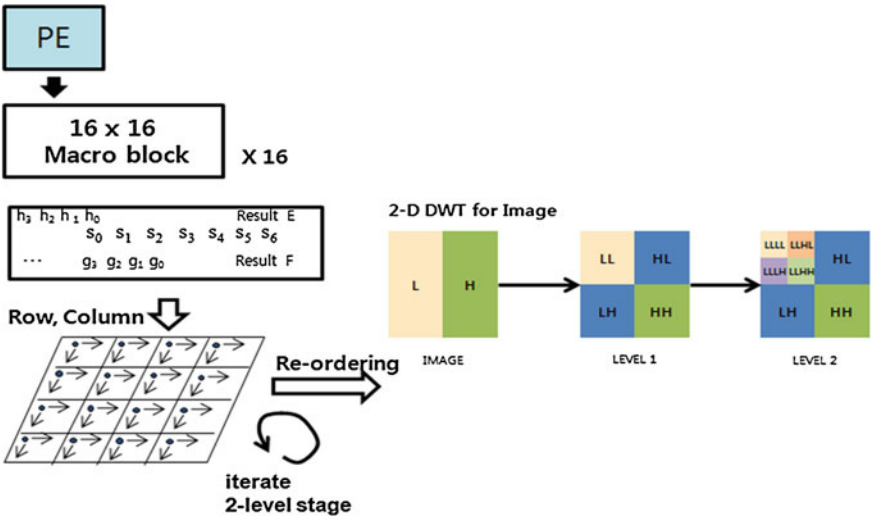


Fig. 17.7 2-level discrete wavelet transform on the multi-core architecture

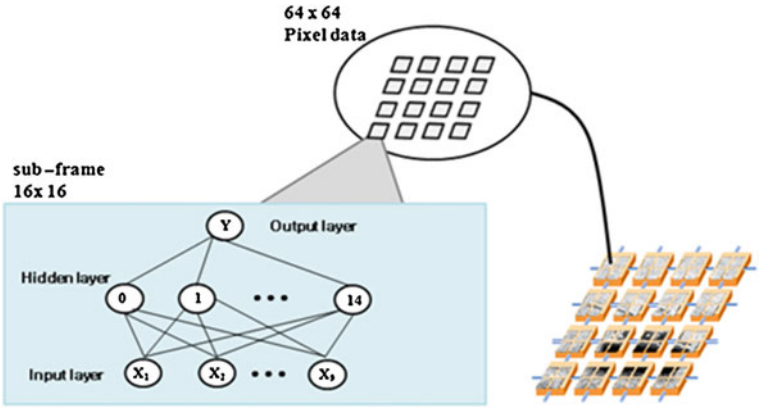


Fig. 17.8 Mapping of BPNN on the many-core architecture

17.4 Synthesis Results and Performance

The proposed multi-core architecture is synthesized using Verilog HDL and implemented on the vertex-4 XC4VLX60 FPGA. Figures 17.9 and 17.10 show a single PE and a 4 × 4 PE array RTL schematics, respectively.

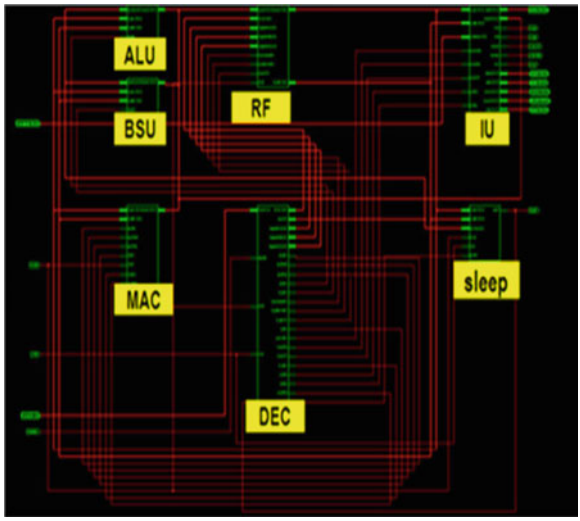


Fig. 17.9 Single PE RTL schematic

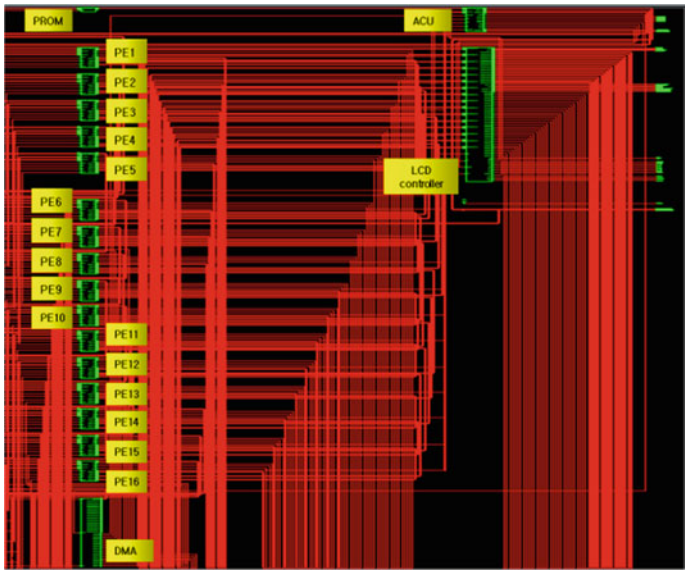


Fig. 17.10 4 × 4 PE array RTL schematic

Table 17.2 describes synthesized results of the multi-core architecture. Each PE requires 634 look-up tables (LUTs), and array control unit (ACU) requires 1,946 slices and 8 K byte RAM. Overall, the multicore architecture requires 12,599 slices and 136 K byte RAM.

Table 17.2 Synthesis results of the multi-core architecture

	LUT	RAM (Kbyte)
ACU	1,021	–
Each PE	681	8
Data exchange interface	1,010	2
Total	11,972	136

Table 17.3 Performance comparison between the multi-core architecture and TI C6146 processor

Parameter	Units	Multi-core architecture	TI C6416
Clock frequency	(MHz)	50	720
Encoding time	(ms)	141.1	266.4

The multi-core prototyping system was evaluated and verified using the selected fire detection algorithm with 240 × 200 resolution videos. It operates at the 50 MHz clock frequency and achieves seven frames per second. In addition, the multi-core system achieves about 1.87× speedup over the commercial high-performance TI C6416 processor operating at 720 MHz clock frequency, as shown in Table 17.3.

17.5 Conclusions

In this paper, we introduced a multi-core prototyping system to support computationally complex fire detection algorithm which includes several techniques: moving region detection, color segmentation, feature extraction, and classification of fire or non-fire using a back propagation neural network. The functionality of the multi-core prototyping system including 16 PEs was verified using the reference fire detection algorithm with 240 × 200 resolution videos, and it achieved seven frames per second at the 50 MHz clock frequency.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2013R1A2A2A05004566), and by the Leading Industry Development for Economic Region (LeadER) grant funded the MKE (The Ministry of Knowledge Economy), Korea in 2013 (No. R0001220).

References

1. Celik T, Demirel H (2009) Fire detection in video sequences using a generic color model. Fire Saf J 44(2):147–158
2. Tao L, Mao Y, Feng P, Haiyang W, Jian D (2013) An efficient fire detection method based on orientation feature. Int J Control Autom Syst 11(5):1038–1045

3. Chunyu Y, Zhibin M, Xi Z (2013) A real-time video fire flame and smoke detection algorithm. In: The 9th Asia-Oceania symposium on fire science and technology, vol 62, pp 891–898
4. Çelik T, Özkaramanlı H, Demirel H (2007) Fire and smoke detection without sensors: image processing based approach. In: 15th European signal processing conference, Poland, pp 1794–1798
5. Chen TH, Wu PH, Chiou YC (2004) An early fire-detection method based on image processing. In: IEEE international conference on image processing, Singapore, vol 3, pp 1707–1710
6. Toreyin BU, Centin AE (2004) Online detection of fire in video. In: IEEE conference on computer vision and pattern recognition, Washington DC, pp 1–5
7. Ko BC, Cheong KH, Nam JY (2009) Fire detection based on vision sensor and support vector machines. Fire Saf J 44(3):322–329
8. Nguyen H, John L (1999) Exploiting SIMD parallelism in DSP and multimedia algorithms using the AltiVec technology. In: Proceedings of international supercomputer conference, New York, pp 11–20
9. Abbo AA, Kleihorst RP, Choudhary V, Sevat L, Wielage P, Mouy S, Vermeulen B, Heijligers M (2008) Xetal-II: a 107 GOPS, 600 mW massively parallel processor for video scene analysis. IEEE J Solid-State Circuits 43(1):192–201
10. Chhugani J, Nguyen AD, Lee VW, Macy W, Hagog M, Chen YK, Baransi A, Kumar S, Dubey P (2008) Efficient implementation of sorting on multi-core SIMD CPU architecture. In: Proceedings of 34th international conference very large data bases, New Zealand, pp 1313–1324
11. Kyo S, Okazaki S, Arai T (2007) An integrated memory array processor for embedded image recognition systems. IEEE Trans Comput 56(5):622–634

Chapter 18
A Novel Spectrum Sensing Scheme Based on Wavelet Denoising with Fuzzy for Cognitive Radio Sensor Networks

Hiep Vu-Van and Insoo Koo

Abstract Ability of cognitive radio (CR) is the most promising solution to solve problems of a common wireless sensor network (WSN), which is assumed to assign a fixed frequency band. Cognitive radio sensor network (CRSN) is the combination of CR's ability into WSN to improve spectrum utilization of each wireless sensor nodes. In order to avoid interference to other users in CRSN, reliable detection of the licensed user signal in the interested spectrum band is a pre-requirement. If we do not know any information about licensed user signal, energy detection will be optimal detection method. However, the energy detector is strongly affected by noise and shadowing of sensing environment. In this paper, we propose a novel spectrum sensing based on wavelet denoising with fuzzy to make cognitive radio users to be high reliable sensor in noisy sensing environment. The simulation results demonstrate the effectiveness of the proposed scheme.

Keywords Cognitive radio · Cognitive radio sensor network · Denoising · Wavelet transform · Fuzzy

18.1 Introduction

Recently, additional bandwidth and higher bit-rates have been required in order to meet the users' demands in wireless communication systems. CR technology [1] has been proposed to improve efficiency of spectrum utilization. The scarcity of

H. Vu-Van · I. Koo (✉)
School of Electrical Engineering, University of Ulsan, San 29, Muger 2-dong, Ulsan 680-749, Republic of Korea
e-mail: iskoo@ulsan.ac.kr

H. Vu-Van
e-mail: vvhiep@gmail.com
URL: http://mcsl.ulsan.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_18, © Springer Science+Business Media Dordrecht 2014

the spectrum band can be relieved by allowing some CR users to opportunistically access the spectrum assigned to a primary user (PU) whenever the channel is free. Otherwise, CR users must vacate their frequency when the presence of a PU is detected. Therefore, high reliable detection of the PU signal is crucial for effectively operating CR networks. The ability of CR can be equipped for the current wireless sensor network, which is assumed to employ a fixed spectrum allocation, to solve their operation challenges such as communication quality, collision problem. This will bring us a new sensor networking paradigm, cognitive radio sensor networks (CRSN). In CRSN, a user which is assigned a frequency band will play a role as a primary user in the band, otherwise it will play a role as a cognitive radio in other frequency band which is assigned to other user.

Energy detection has been widely utilized since it is simple and does not require any prior knowledge of the PU signal. In addition, energy detection is the optimal detection method when only the local noise power is known by the CR users [2]. However, the energy detector is strongly affected by noise and shadowing of sensing environment. Subsequently, noisy sensing environment can cause the significant degradation of CR user's sensing performance. In the paper, a denoising scheme that utilizes both wavelet transform and fuzzy system is proposed to help the energy detector to be more reliable in noisy sensing environment. In the proposed scheme, the fuzzy system is used to improve the thresholding function in order to make wavelet denoising scheme to be more effective.

18.2 System Model

In this paper, we consider the sensing performance of individual CR user which uses energy detector to detect the presence or absence of the PU signal. For normal energy detection, at the ith sensing interval, the received signal energy, $E(i)$, is given as:

$$E(i) = \begin{cases} \sum_{k=k_i}^{k_i+M-1} |n(k)|^2, & H_0 \\ \sum_{k=k_i}^{k_i+M-1} |hs(k) + n(k)|^2, & H_1, \end{cases} \tag{18.1}$$

where M is the number of samples over one sensing interval, k_i is the time slot at which the ith sensing interval begins, $n(k)$ is the additive white Gaussian noise (AWGN), h is the channel gain of the sensing channel between the CR user and the PU and $s(k)$ is the PU signal. H_0 and H_1 represent the absent and present case of PU signal, respectively. When M is relatively large (e.g. $M > 100$), E can be closely approximated as a Gaussian random variable under both hypotheses, such that

$$E \sim \begin{cases} N(M, 2M), & H_0 \\ N(M(\gamma+1), M(2\gamma+1)), & H_1, \end{cases} \tag{18.2}$$

where γ is the SNR of the channel between the PU and the CR user.

18.3 The Proposed Wavelet Denoising Scheme Based on Fuzzy for Spectrum Sensing

18.3.1 Wavelet Denoising

The general wavelet denoising procedure [3, 4] uses soft or hard threshold function to reduce the noise of the signal in the wavelet domain.

The traditional hard threshold function is defined as

$$\widehat{X}_j(t) = \begin{cases} X_j(t), & \left|X_j(t)\right| \geq T_j \\ 0, & \left|X_j(t)\right| < T_j, \end{cases} \tag{18.3}$$

and the traditional soft threshold function is defined as

$$\widehat{X}_j(t) = \begin{cases} sgn\left(X_j(t)\right)\left(\left|X_j(t)\right| - T_j\right), & \left|X_j(t)\right| \geq T_j \\ 0, & \left|X_j(t)\right| < T_j, \end{cases} \tag{18.4}$$

where X_j and $\widehat{X}_j$ are wavelet coefficients of the original signal and the denoised signal respectively, sgn(•) is sign function, and T_j is the threshold value of the jth level wavelet coefficient, which can be calculated as follows:

$$T_j = \frac{\sigma\sqrt{2\ln N}}{\log(j+1)}, \tag{18.5}$$

$$\sigma = \frac{median(|X_1|)}{0.6745}, \tag{18.6}$$

where N is length of original signal and X_1 is the first level wavelet coefficient.

18.3.2 The Proposed Wavelet Denoising Based on Fuzzy Scheme for Spectrum Sensing

In this section, we propose a denoising scheme in order to help energy detector to be robust under the noisy sensing environment. The proposed scheme has two main parts including the denoising part and energy detection part. The denoising part is the combination of wavelet denoising and fuzzy as shown in Fig. 18.1, in which the received signal (noised signal) is denoted as x, D_j is wavelet coefficient of the jth level, μ_j is the output of the fuzzy block and denotes "rate of signal" used for denoising block, $\widehat{D}_j$ is the denoised wavelet coefficients and $\widehat{s}$ is the signal after denoising.

Firstly, each CR user takes samples to get M received signals $x(k)$, $1 \leq k \leq M$, for each sensing interval. The received signal x will be wavelet-transformed to obtain wavelet coefficients D_j. Secondly, the fuzzy block will use D_j as input

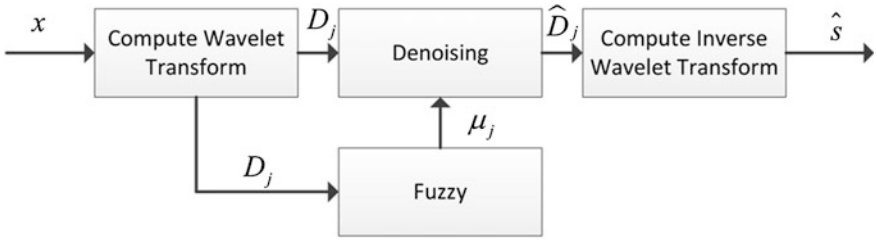


Fig. 18.1 The proposed wavelet denoising scheme

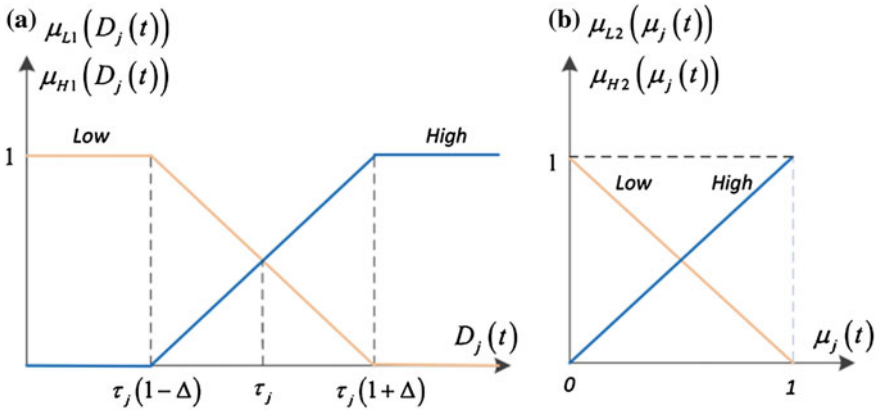


Fig. 18.2 Membership functions of fuzzy sets for **a** wavelet coefficient D_j, and **b** "*rate of signal*" μ_j

information to decide the value of "*rate of signal*" μ_j which ranges from 0 to 1. In the fuzzy block, wavelet coefficients D_j are represented by the two linguistic attributes namely *Low* and *High* whose membership functions are trapezoid shapes as illustrated in Fig. 18.2a in which the value of τ_j is calculated as:

$$\tau_j = \frac{\sigma\sqrt{2\ln M}}{\log(j+1)}, \tag{18.7}$$

$$\sigma = \frac{median(|D_1|)}{0.6745}, \tag{18.8}$$

where D_1 is the first level wavelet coefficient, and Δ can be chosen in range of [0, 1].

The "rate of signal" μ j is represented by the two linguistic attributes Low and High. The membership function are triangle shapes in the range of [0, 1] as illustrated in Fig. 18.2b.

The fuzzy interference rule is proposed as follows:

- Rule 1: IF ($D_j(t)$ *is Low*) THEN ($\mu_j(t)$ *is Low*).
- Rule 2: IF ($D_j(t)$ *is High*) THEN ($\mu_j(t)$ *is High*).

In order to get a low computation cost, the max–min inference method is used in this paper. Subsequently, the truth value of Rule 1 and Rule 2 corresponding with wavelet coefficient can be given as $\mu_{L1}(D_j(t))$ and $\mu_{H1}(D_j(t))$ respectively.

The modified membership function of each rule is derived by taking the minimum of its membership function and the corresponding truth value of the IF clause:

$$\mu_{L1m}(\mu_j(t)) = \min[\mu_{L1}(D_j(t)), \mu_{L2}(\mu_j(t))] \tag{18.9}$$

$$\mu_{H1m}(\mu_j(t)) = \min[\mu_{H1}(D_j(t)), \mu_{H2}(\mu_j(t))] \tag{18.10}$$

The final membership function of μ_f is obtained by taking maximum of two modified membership functions given in Eqs. (18.9) and (18.10):

$$\mu_f(\mu_j(t)) = \max(\mu_{H1m}(\mu_j(t)), \mu_{L1m}(\mu_j(t))) \tag{18.11}$$

The value of $\mu_j(t)$ is obtained by using the center of area defuzzification method as follows:

$$\mu_j(t) = \frac{\int_0^1 \mu_j(t)\mu_f(\mu_j(t))d\mu_j(t)}{\int_0^1 \mu_j(\mu_j(t))d\mu_j(t)} \tag{18.12}$$

After broking the universe of $\mu_j(t)$ into G parts with the same length by $G + 1$ points $\mu(g) = g/G$, for $0 \le g \le G$, the crisp value of $\mu_j(t)$ can be approximately calculated as:

$$\mu_j(t) = \frac{\sum_{g=0}^{G} \frac{g}{G}\mu_f\left(\frac{g}{G}\right)}{\sum_{g=0}^{G} \mu_f\left(\frac{g}{G}\right)} \tag{18.13}$$

Based on the value of "*rate of signal*" $\mu_j(t)$, the improved thresholding function for estimating denoised wavelet coefficient can be given as follows:

$$\widehat{D}_j(t) = \mu_j(t)sign(D_j(t))(|D_j(t)| - \tau_j), \tag{18.14}$$

where τ_j is calculated as Eq. (18.7).

The denoised wavelet coefficient $\widehat{D}_j(t)$ is taken inverse wavelet transform to obtain the denoised signal $\widehat{s}(k)$.

Finally, the denoised signal will be squared to get energy as below:

$$E(i) = \sum_{k=1}^{M} \widehat{s}(k)^2, \tag{18.15}$$

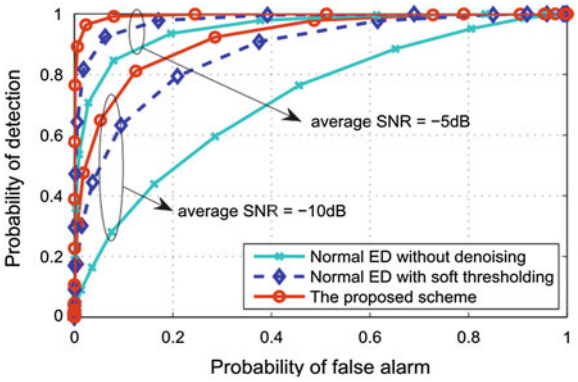


Fig. 18.3 ROC *curves* of spectrum sensing when the signal of PU is constant

where i is the index of sensing interval and M is the number of sensing samples over one sensing interval.

The decision about the presence or absence of the PU signal will be made as follows:

$$\begin{cases} u(i) = 1, & \text{if } E(i) \geq \lambda \\ u(i) = 0, & \text{otherwise,} \end{cases} \tag{16}$$

where λ is the hard threshold for local decision.

18.4 Simulation Results

To evaluate sensing performance of the proposed scheme, we conduct the simulation for the cases of constant and uniform distribution of PU signal when the number of sensing samples per a sensing interval, M is 150 and the value of Δ is chosen as 0.5. As the references, the following detection methods are considered: a normal energy detection method without any denoising scheme (denoted as Normal ED without denoising), an energy detection based on wavelet denoising with traditional soft threshold function (denoted as Normal ED with soft thresholding) and the proposed scheme with wavelet denoising based on fuzzy (denoted as The proposed scheme).

Figure 18.3 illustrates the ROC curves of the proposed scheme and the considered comparison detection schemes in the case of constant PU signal. Figure shows that the proposed scheme can obtain the best sensing performance among these considered schemes. It is mainly due to the better effect of the proposed scheme on denoising. Similar results are obtained in the case of uniform distribution of PU signal as shown in Fig. 18.4. The outstanding performance of the proposed scheme proves that the proposed thresholding function based on fuzzy for wavelet denoising can significantly improve denoising performance compared with the traditional wavelet denoising.

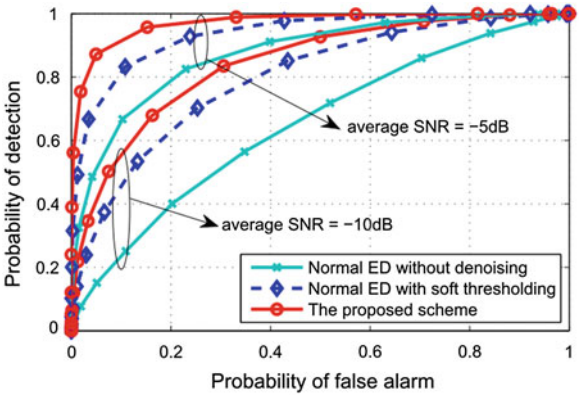


Fig. 18.4 ROC *curves* of spectrum sensing when the signal of PU is uniform distribution

18.5 Conclusion

In this paper, we have proposed a wavelet denoising based on fuzzy for spectrum sensing in CRSN. The proposed scheme provides better denoising than traditional wavelet denoising method in the both cases of constant and uniform distribution of PU signal. Subsequently, the sensing performance of the proposed scheme is significantly improved.

Acknowledgments "This work was supported by the KRF funded by the MEST (NRF-2012R1A1A2038831 and NRF-2013R1A2A2A05004535)".

References

1. Hur Y, Park J, Woo W, Lim K, Lee CH, Kim HS, Laskar J (2006) A wideband analog multi-resolution spectrum sensing (MRSS) technique for cognitive radio (CR) systems. In: Proceedings of IEEE international symposium, circuit and system, Greece, pp 4090–4093
2. Sahai A, Hoven N, Tandra R (2004) Some fundamental limits on cognitive radio. In: Proceedings of allerton conference on communications, control and computing, Monticello
3. Rioul O, Vetterli M (1991) Wavelets and signal processing. IEEE Signal Process Mag 14–38
4. Smith CB, Agaian S, Akopian D (2008) A wavelet-denoising approach using polynomial threshold operators. Signal Process Lett IEEE 15:906–909

Chapter 19
Accelerating a Bellman–Ford Routing Algorithm Using GPU

In-Kyu Jeong, Jia Uddin, Myeongsu Kang, Cheol-Hong Kim and Jong-Myon Kim

Abstract This paper presents a graphics processing unit (GPU)-based implementation of the Bellman–Ford (BF) routing algorithm used in distance-vector routing protocols. In the proposed GPU-based approach, multiple threads concurrently run in numerous streaming processors in the GPU to update the routing information instead of computing the individual vertex distances one-by-one, where an individual vertex distance is considered as a single thread. This paper compares the performance and energy consumption of the GPU-based approach with those of the equivalent central processing unit (CPU) implementation for varying the number of vertices. Experiment results show that the proposed approach outperforms the equivalent sequential CPU implementation in terms of execution time by exploiting massive parallelism inherent in the BF routing algorithm.

Keywords Bellman–Ford algorithm · Internet protocol routing · Graphics processing unit · Routing algorithm

I.-K. Jeong · J. Uddin · M. Kang · J.-M. Kim (✉)
School of Electrical Engineering, University of Ulsan, Ulsan, South Korea
e-mail: jongmyon.kim@gmail.com

I.-K. Jeong
e-mail: jeonginkeyu@gmail.com

J. Uddin
e-mail: engrjiauddin@gmail.com

M. Kang
e-mail: ilmareboy@gmail.com

C.-H. Kim
School of Electronics and Computer Engineering, Chonnam National University, Gwangju, South Korea
e-mail: cheolhong@gmail.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_19, © Springer Science+Business Media Dordrecht 2014

19.1 Introduction

The conventional hardware routers were usually used for a specific application with a precise throughput capacity. Those routers are unable to meet the dynamic programmability for high data rate applications including online gaming, internet TV and broadband video streaming [1, 2]. Thus, software routers have been used to support the dynamic programmability, but the shortcoming of the software routers is that they only provide limited throughput performance up to 3 Gbps range. Nowadays, 4 G-capable Internet protocol (IP) routing requires the throughout ranged from 40 Gbps to 90 Tbps. Consequently, a graphics processing unit (GPU)-based software paradigm is highly desirable.

As a packet processing technique, the Bellman–Ford (BF) algorithm is widely utilized, which computes the shortest paths from a single source vertex to all of the other vertices in a weighted diagraph. However, the time-consuming property of the BF algorithm has limited its use for real-time applications. To deal with this drawback, this paper implements the BF algorithm using GPU and improves its speedup by exploiting massive parallelism inherent in the BF algorithm. To show the potential of the improved performance of the proposed GPU-based approach, this paper compares the proposed approach with that of the CPU-based implementation in terms of speedup, power consumption, and energy consumption.

The rest of this paper is organized as follows. Section 19.2 includes the background study, where an overview of the GPU architecture and the BF algorithm is presented. Section 19.3 presents a parallel implementation of the BF algorithm using GPU. Section 19.4 analyzes the experimental results, and finally Sect. 19.5 concludes the paper.

19.2 Background Information

19.2.1 Compute Unified Device Architecture

Compute Unified Device Architecture (CUDA) is a parallel computing platform, where multiple threads concurrently execute the same set of instructions on different chunks of data. The multiple threads which execute instructions on the data sets can be arranged in either one-dimensional (1D) or multidimensional (2D, 3D) blocks. Figure 19.1 shows a streaming processor (SP) on GPU architecture [3], where different threads process on the blocks B_1, B_2, …, B_n.

Comparing to the task parallelism in multi-core CPU, GPU utilizes a data parallel programming scheme with its streaming multiprocessors (SMs), where thousands of threads run concurrently on disparate data chunks. The special

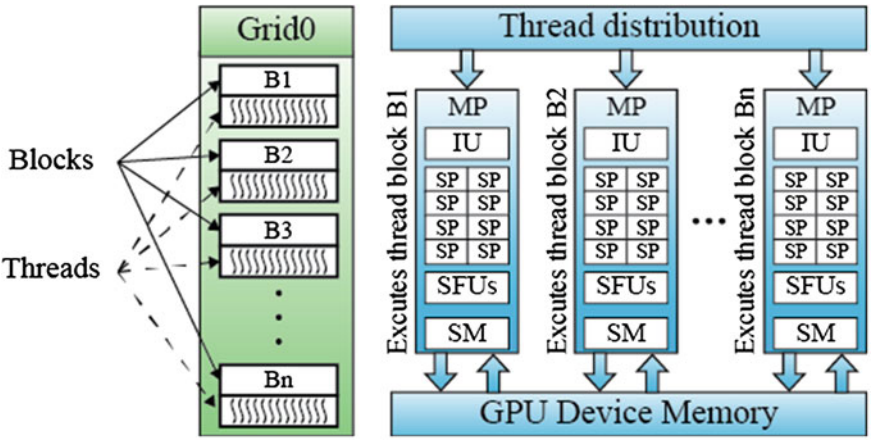


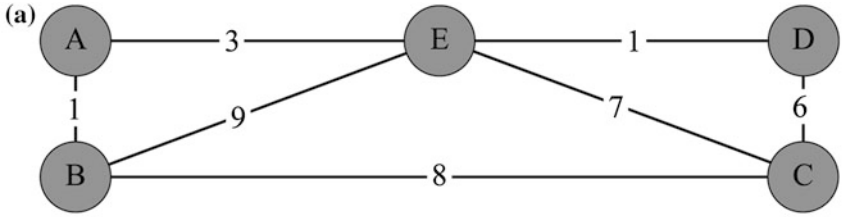
Fig. 19.1 Multi-threading arrangement on GPU

function unit (SFU) executes one instruction per thread per clock cycle. Each thread has a unique identifier (i.e., *threadIdx*) within a block to select specific data. The thread has a relatively large number of registers and a private area of memory (local memory) which is used for register file spilling. Threads within a block can cooperate by using the fast shared memory.

19.2.2 A Bellman Ford Distance Vector Routing Algorithm

Given routers A, B, C, D, and E, let $\exists\ i$ which is a label for the routers A, B, C, D, and E. Furthermore, let $D(i, j)$ be the distance for a known route from i to a router location j. Let $d(i, j)$ be the distance from router i to neighbor j, where $d(i, j)$ is set to infinity if $i = j$ or i and j are not immediate neighbors. The BF equation that can find the best route between router A and E in Fig. 19.2a is as follows. $D(i, j) = \min(d(i, k) + D(k, j))\ \forall i \neq j$. The overall BF algorithm [4] is shown in Fig. 19.2b.

The BF algorithm solves the BF equation by calculating the $D(i, j)$ node, for example, node A and E in Fig. 19.2a by assuming that node i needs $d(i, k)$'s and $D(k, j)$ distances from the neighbor nodes. The associated problem with this is that the $D(k, j)$ from neighbors which are more than one hop to the node is not known by the considered node. Thus, the algorithm resolves this by finding the distance with the least weight link from i to j using single hop links $D(i, j)$ for each node i until the node E is reached from node A.



(b)
Bellman-Ford Algorithm

```
1:  Procedure Bellam Ford(list vertices, list edges, vertex source)
2:  STEP 1: initialize graph
3:  for each vertex v in set of vertices : if v is source then v.distance = 0, else v.distance = infinity
4:  set v.predecessor = null
5:  Step 2: relax edges repeatedly
6:  for i from 1 to size (vertices)-1:
7:  for each edge uv in edges: uv is the edge from u to v
8:  u = uv.source
9:  v = uv.destination
10: if u.distance + uv.weight <  v.distance:  v.distance =u.distance + uv.weight
11: v.predecessor = u
12: Step 3: Check for negative-weight cycles for each edge uv in edges:
13: u = uv.source
14: v = uv.destinationt
15: if u.distance + uv.weight <  v.distance: error, then output a negative-weight cycle message.
```

Fig. 19.2 Connected routers with different link weights

19.3 The Proposed GPU-Based Bellman Ford Algorithm

This paper implements the BF algorithm on a GeForce GTX 580 [5]. In addition, we utilize a CUDART version 4.2 to run CUDA C program with the *nvcc* compiler. Detailed parameters of the host (CPU) and the device (GPU) are given in Tables 19.1 and 19.2.

One of the most important tasks in the implementation of the BF algorithm involves the calculation of the relaxed and negatively weighted edge cycles, where the recurring loop problem is addressed. Concisely, ***step* 2** in Fig. 19.2b guarantees the shortest distance if the graph does not contain a negative weight cycle. In ***step* 3**, there is an attempt to terminate the iteration if a negatively weighted edge cycle is encountered. The iteration is terminated if the graph contains a negative weight cycle for any edge u–v by the following condition: $dist[v] > dist[u] +$ weight of edge u–v.

The difference between the instruction execution on the CPU and GPU is shown in Fig. 19.3a and b, respectively. As shown in Fig. 19.3, a number of threads concurrently execute the necessary instructions for calculating the IP route itself, which results in achieving massive parallelism as the GPU-based approach is performed. The entire graph with all the node and edges can be relaxed in a parallel operation by the GPU. After the initialization operation, relaxing edges and checking the negative cycle operations can be fully accomplished by the GPU in a parallel fashion. In contrast, the CPU-based implementation is executed for each node until the entire graph is traversed. Since the execution pattern using the

Table 19.1 Hardware specification of a CPU

Parameters	Values
Processor	Intel Xeon × 5690
Number of core	6
Thread	12
Clock frequency	3.46 GHz

Table 19.2 Hardware specifications of a GPU

Parameters	Values
Device	GeForce GT × 580
Total number of CUDA cores	512
GPU clock speed	1.54 GHz
Warp size	32
Maximum number of threads per block	1,024
Maximum sizes of each dimension of a block	1,024 × 1,024 × 1,024
Maximum sizes of each dimension of a grid	65,535 × 65,535 × 65,535

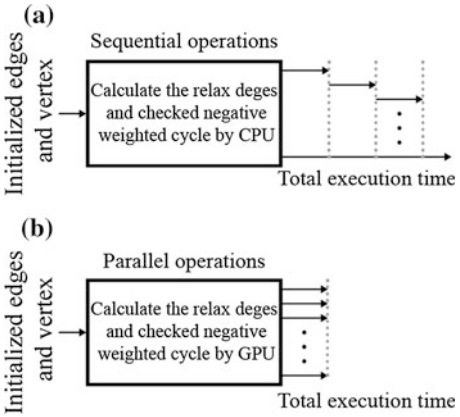


Fig. 19.3 The Bellman–Ford algorithm **a** using CPU and **b** using GPU

CPU is serial, it takes much larger execution time than the GPU-based approach with parallel execution pattern. The total execution time in the CPU depends on the vertexes and edges initialization and the completion of node-by-node and edge-by-edge BF execution. Figure 19.4 shows a flow chart of the BF implementation using a CPU and GPU heterogeneous system.

The CPU in this case is just used to input the data and output the result from the GPU, while the GPU calculates the relax edges and checks for a negative weight cycle. A detailed implementation of the GPU-based BF algorithm is shown in Fig. 19.5, where edges *s*-*r*1, *s*-*r*2, *s*-*r*3 are considered as *thread*1, *thread*2, *thread*3, respectively. As the groups of vertexes are launched concurrently into the GPU, the output of the different blocks by using the GPU is calculated much faster than that by using the CPU.

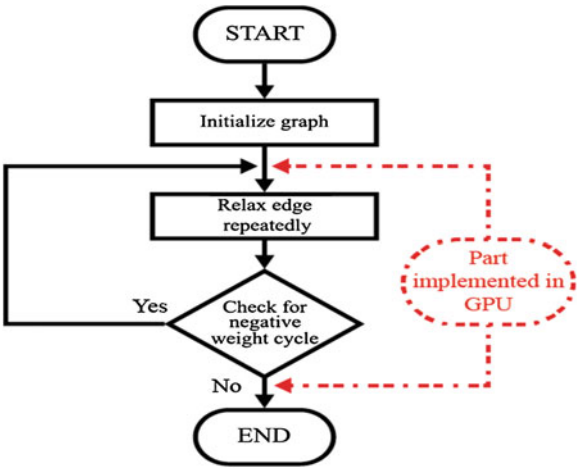


Fig. 19.4 A CPU/GPU implementation of the Bellman–Ford algorithm

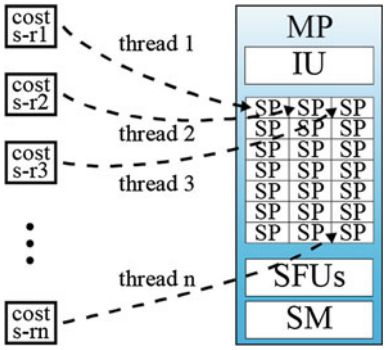


Fig. 19.5 A parallel implementation of the Bellman–Ford algorithm on GPU

In the GPU-based BF implementation, we allocate the memory on the global memory of GPU, and then transfer the detail scenario information from CPU to GPU. After executing the parallel threads in different streaming processors of GPU, the updated scenario information by executing the BF algorithm returns to the CPU. A main CUDA C function with different parts of the program which runs on the GPU is presented in Fig. 19.6, where a kernel function *dvr* is declared and it runs on the GPU as a global function.

19.4 Performance Evaluation

This paper evaluates the performance of the proposed GPU-based approach for the BF algorithm in terms of execution time. In addition, this paper compares the performance of the proposed approach with that of the equivalent CPU implementation. Table 19.3 shows execution times for both approaches with varying

```
main(){
//variable declaration
//allocate the memory in GPU
cudaMalloc((void**)&dev_rt,n*sizeof(node));
cudaMalloc((void**)&dev_dmat,n*n*sizeof(int));
//transfer the scenario information to GPU
cudaMemcpy(dev_dmat,dmat, n*n*sizeof(int)),
cudaMemcpyHostToDevice);
cudaMemcpy(dev_rt, rt, n*n*sizeof(node),
cudaMemcpyHostToDevice);
//declared block and grid size based on the total threads
//kernel function dvr
dvr<<<dimGrid, dimBlock>>>(dev_dmat, dev_rt);
//transfer results from GPU to CPU
cudaMemcpy(dmat, dev_dmat, n*n*sizeof(int),
cudaMemcpyDeviceToHost);
cudaMemcpy(rt, dev_rt, n*n*sizeof(node),
cudaMemcpyDeviceToHost);
//cuda free}
```

Fig. 19.6 Main function of CUDA C code

Table 19.3 Performance comparison of the GPU and the equivalent CPU-based implementation

V	T_c	T_g	$T_{h2d\&d2h}$	T_e	S_p
300	220	15.11514	1.47392	13.64122	14.6×
500	1,100	38.519966	1.529536	36.99043	28.6×
1,000	21,000	147.8969	4.59984	143.297	141.9×
2,000	139,660	577.192444	11.85725	565.3352	241.9×

V number of vertices, T_c CPU execution time in millisecond, T_g GPU execution time in millisecond, $T_{h2d\&d2h}$ transfer time to copy host to device and vice versa in millisecond, T_e Kernel execution time in millisecond, S_p Speedup

numbers of vertices. As presented in Table 19.3, the GPU-based approach increases speedup over the CPU-based approach by exploiting data parallelism inherent in the BF algorithm.

Especially, the GPU-based approach achieves higher speedups over the CPU when the number of vertices is greater than or equal to 500, as shown in Fig. 19.7. One of main reasons is that the percentage of data transfer time from CPU to GPU and vice versa at $V \geq 500$ over the corresponding total execution time of the GPU-based implementation is much smaller than that of the data transfer time at $V < 500$, as shown in Table 19.3, where the data transfer time is one of performance bottlenecks for GPU computing.

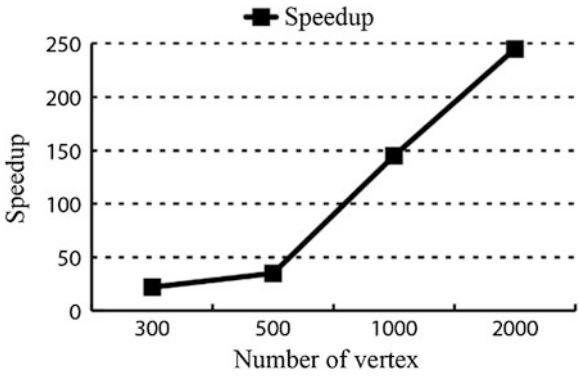


Fig. 19.7 Speedups of GPU over CPU for varying the number of vertices

19.5 Conclusion

This paper accelerated the BF algorithm using a GPU and provided a solution of high-speed software routers. In addition, this paper evaluated the performance of the proposed GPU-based approach with varying the number of vertices in terms of execution time. Moreover, we compared the performance of the GPU-based approach with the equivalent sequential CPU implementation. Experimental results showed that the GPU-based approach outperformed the CPU implementation in terms of execution time for all the cases varying the number of vertices.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2013R1A2A2A05004566).

References

1. Shi L, Chen H, Sun J (2012) vCUDA: GPU accelerated high performance computing in virtual machines. IEEE Trans Comput 61(6):1–11
2. Han S, Jang K, Park KS, Moon S (2010) PacketShader: a GPU-accelerated software router. In: ACM special interest group on data communication (SIGCOMM) 2010 conference. New Delhi, pp 195–206
3. Chikkagoudar S, Wnag K, Li M (2011) GENIE: a software package for gene–gene interaction analysis in genetic association studies using multiple GPU or CPU cores. BMC Res Notes 4(158):1–7
4. Bellman–Ford Algorithm. http://www.geeksforgeeks.org/dynamic-programming-set-23-bellman-ford-algorithm/
5. GeForce GTX 580. http://www.geforce.com/hardware/desktop-gpus/geforce-gtx-580

Chapter 20
Early Fire Detection Using Multi-Stage Pattern Recognition Techniques in Video Sequences

Dongkoo Shon, Myeongsu Kang, Junsang Seo and Jong-Myon Kim

Abstract This paper proposes an efficient early fire detection approach using multi-stage pattern recognition techniques, including background subtraction for movement-containing region detection, statistical rule-based color segmentation in *YCbCr* color space, a single-level spatial wavelet decomposition for observing flicker of fire, and a support vector machine to identify between fire of non-fire. This paper evaluates the proposed approach in terms of percentage of true positive and false negative. Experimental results indicate that average fire detection and false non-fire detection rates are 99.67 and 3.69 %, respectively.

Keywords Background subtraction · Wavelet decomposition · Support vector machine · Surveillance system

20.1 Introduction

Early fire detection has been an increasingly important issue since it is closely related to personal security and property. In spite of the fact that sensor-based fire detection systems first came into the spotlight by detecting either heat or smoke for early identification of whether or not a fire is occurring, these systems have the

D. Shon · M. Kang · J. Seo · J.-M. Kim (✉)
School of Electrical Engineering, University of Ulsan, Ulsan, South Korea
e-mail: jongmyon.kim@gmail.com

D. Shon
e-mail: dongkoo88@gmail.com

M. Kang
e-mail: ilmareboy@gmail.com

J. Seo
e-mail: siberiaj00@gmail.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_20, © Springer Science+Business Media Dordrecht 2014

drawback that sensors should be densely distributed in a wide area for a highly precise fire detection system [1]. Recent advances in video processing technologies have led to a wave of research on computer vision-based fire detection systems whose advantages are summarized as follows [2]: (1) As the speed of light transmission is much faster than that of the heat and smoke, computer vision-based fire detection is appropriate for early detection of fire, and (2) In general, images have more scene information such as color and texture, which enables diverse approaches to fire detection.

Several video-based fire detection algorithms have been introduced by using color pixel recognition, motion detection, or both [3–6]. For example, Toyeyin et al. [6] proposed a flame detection algorithm not only detecting fire and flame colored moving regions in video but also analyzing the motion of such regions in wavelet domain for flicker estimation. This algorithm is with considerable success but it lacks robustness. To improve the performance of early fire detection in video sequences, this paper proposes an efficient fire detection approach using multi-stage pattern recognition techniques.

The rest of this paper is organized as follows. Section 20.2 presents the proposed fire detection approach using background subtraction, color segmentation, wavelet decomposition, and support vector machine. Section 20.3 illustrates experimental results and Sect. 20.4 finally concludes this paper.

20.2 The Proposed Fire Detection Method

20.2.1 Movement-Containing Region Detection Based on Background Subtraction

Since the boundaries of a fire tend to continuously fluctuate, MCRD has been widely used as the first step of fire detection, which selects candidate regions of fire. As mentioned earlier, background subtraction is utilized for MCRD in this study. A pixel positioned at (i, j) is assumed to be moving if the following condition is satisfied:

$$|I_n(i,j) - B_n(i,j)| > Th, \tag{20.1}$$

where $I_n(i, j)$ represents the intensity value of the pixel at location (i, j) in the nth gray-level input video frame, $B_n(i, j)$ is the background intensity value at the same pixel position, and Th is a threshold value (which was experimentally set to 3 in this study). The background intensity value is iteratively updated using (20.2):

$$B_{n+1}(i,j) = \begin{cases} B_n(i,j) + 1 & \text{if } I_n(i,j) > B_n(i,j) \\ B_n(i,j) - 1 & \text{if } I_n(i,j) < B_n(i,j) \\ B_n(i,j) & \text{if } I_n(i,j) = B_n(i,j) \end{cases}, \tag{20.2}$$

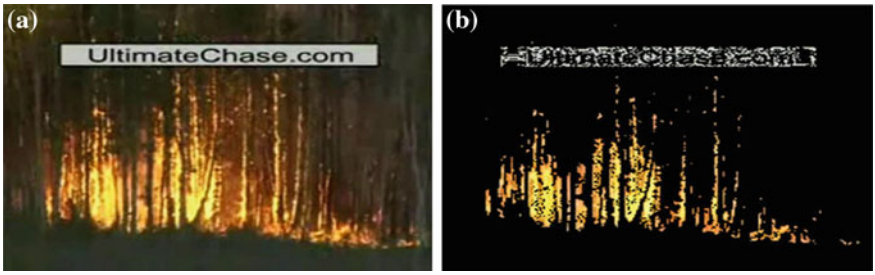


Fig. 20.1 Example of MCRD. **a** An original fire-containing frame and **b** a fire-containing frame after MCRD

where $B_{n+1}(i, j)$ is the estimated background intensity value of the pixel at location (i, j) and $B_n(i, j)$ is the previously estimated background intensity value at the same pixel position. Initially, the background intensity value $B_1(i, j)$ is set to the intensity value of the first video frame, $I_1(i, j)$. Figure 20.1 illustrates an example of MCRD.

20.2.2 Color Segmentation

A number of moving objects (e.g., people, vehicles, animals, and so on) besides fire can be still included after MCRD. Thus, this paper uses further information such as color variations. A set of rules has been developed over the past few decades to classify fire pixels by utilizing raw red–green–blue (*RGB*) information in color video sequences. However, the *RGB* color space has the disadvantage of illumination dependence. The chrominance can be used to model the color of fire rather than its intensity, which gives a more robust representation for fire pixels. Thus, recently, many researchers have used color spaces, such as *YCbCr*, *YUV*, and *CIE Lab*, in which the chrominance components (*Cb*, *Cr*, *U*, *V*, *a*, *b*) and luminance component (*Y*) of an image can be processed independently. In this study, we use the *YCbCr* color space for detecting fire pixels. The conversion from *RGB* to *YCbCr* color space is performed as follows:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.2568 & 0.5041 & 0.0979 \\ -0.1482 & -0.2910 & 0.4392 \\ 0.4392 & -0.3678 & -0.0714 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix}, \tag{20.3}$$

where Y is the luminance and Cb and Cr are the chrominance components for blue-difference and red-difference, respectively. To model fire pixels, the defined rules for the *RGB* color space, i.e., $R > G > B$ and $R > R_{mean}$, can be translated into the *YCbCr* space such as $Y > Cb$ and $Cr > Cb$. In addition, since the fire-containing regions are generally the brightest regions in the observed scene, the mean values

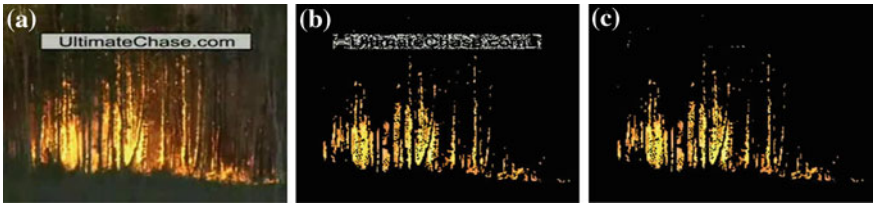


Fig. 20.2 Color segmentation result from a fire-containing image after MCRD. **a** An original fire-containing image, **b** MCRD result, and **c** color segmentation result

of the three channels include important information, which can be expressed as follows:

$$F_{candidate}(i,j) = \begin{cases} 1, & \text{if } Y(i,j) > Y_{mean}, Cb(i,j) < Cb_{mean}, \\ & Cr(i,j) > Cr_{mean} \\ 0, & Otherwise \end{cases}, \tag{20.4}$$

where $F_{candidate}(i, j)$ indicates that any pixel at the spatial location (i, j) which satisfies the condition given in (20.4) is labeled as a fire pixel. Likewise, the mean values of the three channels in the *YCbCr* color space for an $M \times N$ image can be defined as follows:

$$\begin{aligned} Y_{mean} &= \frac{1}{M \times N}\sum_{i}^{M}\sum_{j}^{N} Y(i,j), \quad Cb_{mean} = \frac{1}{M \times N}\sum_{i}^{M}\sum_{j}^{N} Cb(i,j), \\ Cr_{mean} &= \frac{1}{M \times N}\sum_{i}^{M}\sum_{j}^{N} Cr(i,j), \end{aligned} \tag{20.5}$$

where $Y(i, j)$, $Cb(i, j)$, and $Cr(i, j)$ are the luminance, chrominance-blue, and chrominance-red values at the spatial location (i, j) after MCRD, respectively. Figure 20.2 shows the color segmentation result from a fire-containing image after MCRD, and its result is good enough to detect more refined candidate regions compared to the candidate regions of fire right after MCRD.

20.2.3 Color Variations Using Spatial Wavelet Analysis

Due to the nature of turbulent fire flicker, there are generally more color variations in genuine fire-containing regions whereas there are few color variations in the candidate regions of fire, which may still include fire-colored objects after color segmentation. Thus, this paper captures color variations in pixel values by performing spatial wavelet analysis. Since high-frequency information such as edges and texture around the fire is not sensitive to lighting change and more prominent signatures can discriminate irregular fire from the regular movement of

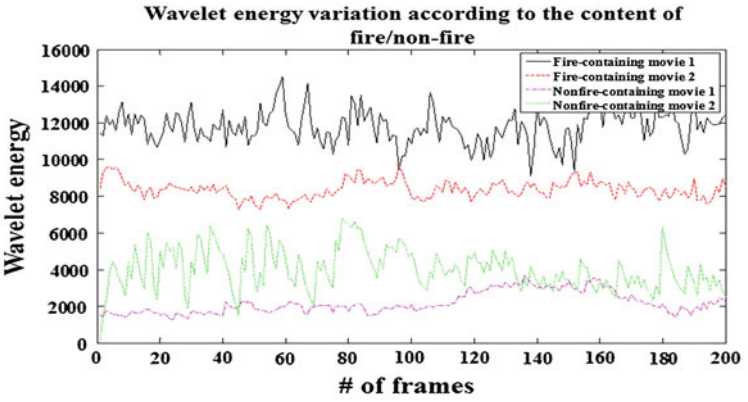


Fig. 20.3 Normalized wavelet energies with fire/non-fire-containing movies

fire-colored objects, the wavelet energy of high frequency sub-images provides a good representation of turbulent fire flicker, which is calculated as follows:

$$E(n) = \frac{1}{floor\left(M/2\right) \times floor\left(N/2\right)} \left\{ |HL_n|^2 + |LH_n|^2 + |HH_n|^2 \right\}, \tag{20.6}$$

where $E(n)$ is the normalized wavelet energy of the n*th* video frame, and HL_n, LH_n, and HH_n contain the horizontal, vertical, and diagonal high frequency of the nth $floor(M/2) \times floor(N/2)$ sub-images obtained by a single-level wavelet transform, respectively, and Fig. 20.3 depicts the normalized wavelet energies with two fire-containing videos and two non-fire containing videos, each video consists of 200 frames. $E(n)$ is then used as the input of a classifier for detecting fire in a video clip. Furthermore, we use a Daubechies 4-tap filter in this study, which avoids bad localizations.

20.2.4 Classification

To classify candidate pixels as fire or non-fire pixels, this paper employs a support vector machine (SVM), which offers high classification accuracy with limited training data and does not require heuristic parameters for detecting fire pixels. The SVM is a non-probabilistic binary classifier and its main goal is to find an optimal hyper-plane that correctly separates the largest fraction of data points while maximizing the distance between two classes on the hyper-plane. The SVM classification function is defined as:

$$f(x) = sign\left(\sum_{i=0}^{l-1} w_i \cdot k(x, x_i) + b\right), \tag{20.7}$$

where w_i are weights for outputs of each kernel, $k()$ is a kernel function, b is a bias term, l is the number of support vectors of x_i, and $sign()$ determines the class membership of x (i.e., +1 class and −1 class). The classification function determined by support vectors is then used to measure how much a pixel belonging to the fire class (e.g., +1 class) is different from the non-fire class (e.g., −1 class). In this study, we use a one-dimension feature vector including fire signatures in order to identify fire in the video clip. However, since two classes (e.g., fire or non-fire) are not linearly separable with the non-linear feature vector, it is necessary to find an optimal hyper-plane that can split the non-linear feature vector by mapping it to a high-dimensional feature space. To deal with this problem, we use the radial basis function (RBF) kernel as follows:

$$k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right) \text{ for } \sigma > 0, \tag{20.8}$$

where x and y are input feature vectors, and σ is a parameter that determines the width of the effective basis function, which affects the classification accuracy. In this study, we experimentally set the standard deviation (σ) to 0.1 yielding high classification performance. The input test value x and the support vectors x_i obtained from a training data set are non-linearly mapped features using the RBF kernel. A candidate fire pixel is finally classified as either a real fire pixel if the result is 1 or a non-fire pixel if the result is −1 by using (20.7). To train the SVM, we build a training dataset that includes 200 wavelet energies from training fire pixels and 200 wavelet energies from fire-colored moving pixels, respectively.

20.3 Experimental Results

We implement the selected fire detection algorithm in MATLAB 2012b on an Intel Quad-Core 3.4 GHz PC platform. Furthermore, five videos are used for evaluating the accuracy of the fire detection algorithm, including 2,642 samples with dimensions of 256 × 256 (1,301 samples containing fire and 1,341 samples containing non-fire), as illustrated in Fig. 20.4.

Table 20.1 presents the accuracy of the fire detection algorithm in terms of true positives (TP) and false negatives (FN). TP is the number of all frames that correctly detect a real fire as a fire and the percentage of TP (PTP) is the overall fire detection rate. Moreover, the FN is the number of all frames that detect a real fire as a non-fire and the percentage of FN (PTN) is the overall false non-fire detection rate. As shown in Table 20.1, the results indicate that average fire detection and false non-fire detection rates are 99.67 and 3.69 %, respectively, which are good

Fig. 20.4 Examples of test videos used in this study

Table 20.1 Result of the selected fire detection algorithm

Movies (# of frames)									
Movie 1 (500)		Movie 2 (599)		Movie 3 (199)		Movie 4 (946)		Movie 5 (393)	
TP	PTP	TP	PTP	TP	PTP	FN	PFN	FN	PFN
500	100.00	598	99.83	198	99.50	0	0.00	29	7.38

enough for fire detection since they consistently increase the accuracy of fire detection while decreasing the error of false fire detection in all videos.

20.4 Conclusion

This paper proposed an efficient fire detection approach using multi-stage pattern recognition techniques: background subtraction for MCRD, rule-based CS in *YCbCr* color space, wavelet decomposition for describing the behavior of fire, and a SVM for identifying between fire and non-fire. Experimental results showed that the proposed method achieves a low false alarm rate, and high reliability in test videos. These results demonstrate that the proposed method is a promising candidate for use in automatic fire-alarm systems.

Acknowledgments This work was supported by the Leading Industry Development for Economic Region (LeadER) grant funded the MOTIE (Ministry of Trade, Industry and Energy), Korea in 2013 (No. R0001220) and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2013R1A2A2A05004566).

References

1. Celik T, Demirel H (2009) Fire detection in video sequences using a generic color model. Fire Saf J 44(2):147–158
2. Qiu T, Yan Y, Lu G (2012) An autoadaptive edge-detection algorithm for flame and fire image processing. IEEE Trans Instrum Meas 61(5):1486–1493

3. Chen TH, Wu PH, Chiou YC (2004) An early fire-detection method based on image processing. In: IEEE international conference on image processing. Singapore, pp 1707–1710
4. Toreyin BU, Centin AE (2007) Online detection of fire in video. In: IEEE international conference on computer vision and pattern recognition. Minneapolis, pp 1–5
5. Ko BC, Cheong KH, Nam JY (2009) Fire detection based on vision sensor and support vector machine. Fire Saf J 44(3):322–329
6. Toreyin BU, Dedeoglu Y, Gudukbay U, Cetin AE (2006) Computer vision based method for real-time fire and flame detection. Pattern Recogn Lett 27(1):49–58

Chapter 21
Energy Efficient Backoff-Free Fragment Retransmission Fragmentation Scheme for IEEE 802.11 WLAN

Prosper Mafole, Yuki Manabe, Teruaki Kitasuka and Masayoshi Aritsugi

Abstract This paper investigates the energy efficiency of MAC layer fragmentation. Despite the usefulness of fragmentation in improving transmission reliability in poor quality wireless channels, it comes with overhead costs that lead to its energy inefficiency. We propose backoff-free fragment retransmission (BFFR). BFFR is a fragmentation scheme that addresses some of classical fragmentation overheads. It aims at improving classical fragmentation energy efficiency without adversely affecting its throughput. We implement BFFR in NS3 and performance evaluation showed that BFFR outperforms classical fragmentation scheme in both throughput and energy efficiency.

Keywords Fragmentation · IEEE 802.11 MAC · DCF · Energy efficiency · Energy consumption · Backoff-free · Throughput · Fragment retransmission

21.1 Introduction

In recent years, IEEE 802.11 wireless local area networks (WLANs) have been widely deployed in public and private areas due to their low cost and improved throughput [1, 2]. Apart from faster and higher capacity demands [1], energy

P. Mafole (✉) · Y. Manabe · T. Kitasuka · M. Aritsugi
Department of Computer Science and Electrical Engineering, Graduate School of Science and Technology, Kumamoto University, Kumamoto, Japan
e-mail: mafole@dbms.cs.kumamoto-u.ac.jp

Y. Manabe
e-mail: y-manabe@cs.kumamoto-u.ac.jp

T. Kitasuka
e-mail: kitasuka@cs.kumamoto-u.ac.jp

M. Aritsugi
e-mail: aritsugi@cs.kumamoto-u.ac.jp

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_21, © Springer Science+Business Media Dordrecht 2014

efficiency demands have been increasing drastically. On one hand, this is due to the fact that more portable and mobile devices are equipped with WLAN interfaces and thus enabling users to access many emerging mobile broadband Internet applications and services [2]. And on the other, these devices rely on battery power and they communicate over a wireless radio link whose quality variation, in time and space, is random. The compulsory and default medium access control (MAC) protocol employed by devices communicating over IEEE 802.11 WLANs is another reason for the increased energy consumption. These observations motivate the current study of IEEE 802.11 MAC protocol to investigate its impact on the energy efficiency of WLAN interfaces. We specifically focus on MAC layer fragmentation in continuous active mode (CAM) of operation.

Fragmentation is a process of splitting a large frame into a number of smaller ones which are independently transmitted. It has been studied on its own [3, 4] or in combination with rate adaptation [1, 5, 6] for improving throughput. Tourrilhes [4] proposed a simple dynamic fragmentation scheme to deal with interference and thus improve throughput. The behavior of TCP in 802.11 based networks focusing on the effect of fragmentation and frame retransmissions done at MAC layer on the end-to-end TCP performance is examined in [3]. Authors in [5] proposed a dynamic fragmentation scheme whereby the transmission duration of all fragments, but the last one, is the same regardless the physical layer data rate. Different fragmentation sizes at different transmission rates are selected based on the channel condition. In this scheme, a new fragment is generated only when the rate for the next fragment transmission is decided. Throughput is the main performance metric. None of these works investigated the energy efficiency of MAC layer fragmentation.

While there have been a number of research papers pertaining to IEEE 802.11 MAC protocol energy consumption and efficiency, for instance [7–13], few have considered fragmentation from an energy efficiency perspective. Ramos et al. [10] used the received signal strength to estimate bit error rate (BER) based on which the fragmentation threshold for the next transmission is decided. As summarized in [2] techniques for energy efficient MAC protocols for IEEE WLAN in continuous active mode (CAM) can be put in three categories. The first consists of strategies that conserve energy during contentions. Strategies in the second category reduce energy consumption during transmission and retransmission of packets. The third one consists of techniques that reduce energy consumption by eliminating contentions, inter frame spaces (IFSs) and acknowledgements.

Fragmentation is used to improve the reliability of data transfer across a poor quality wireless channel [1, 3, 5, 12, 14]. We aim at exploiting this feature. However, it comes with extra overheads in terms of IFSs and acknowledgements for each fragment transmitted. During contentions the IEEE 802.11 station (STA) will consume energy due to overhearing neighbor transmissions or sensing the channel [9]. When there is a transmission failure there is more overhead due to channel contention and hence more energy consumption. To reduce energy consumption we selected the concept of eliminating contentions, IFSs and acknowledgements. In this paper we propose a novel, energy efficient fragmentation

scheme that reduces the overhead due to contention. We employ fragmentation to reduce the number of retransmissions in poor quality channel. In cases of fragment transmission failure we try to retransmit the fragment immediately without contending for the channel.

The rest of the paper is organized as follows: Sect. 21.2 is on the classical IEEE 802.11 MAC protocol while the proposed scheme is described in Sects. 21.3 and 21.4 presents its performance evaluation. Section 21.5 concludes our paper.

21.2 Fragmentation in IEEE 802.11 MAC Protocol

The compulsory medium access control protocol in IEEE 802.11-based wireless networks is distributed coordination function (DCF). We focus on the basic access scheme. Before the STA sends a packet, it has to sense the shared medium. If the medium is found free it sends a packet and waits for an acknowledgement. Else if the medium is found busy the STA backs-off for a random period of time, the backoff counter. The random backoff counter picked by the contending STAs is meant to reduce collisions during medium access. In addition to backoff, it is a requirement that transmission happens only in the beginning of a slot.

The STA will also backoff in cases whereby an acknowledgment (ACK) is not received within a duration called extended inter frame space (EIFS). When the backoff counter is decremented to zero and the medium is sensed to be free for a DCF inter frame space (DIFS) the STA will retransmit the failed transmission or send a new packet accordingly. Due to space limitations we refer readers interested in the detailed working of DCF to IEEE 802.11 specification documents or [2] as an example of relevant literature.

Fragmentation is a DCF enhancement feature that can be enabled by specifying a fragmentation threshold [2]. When the size of a MAC service data unit (MSDU) arriving from the network layer is larger than that of the fragmentation threshold it will be fragmented. The new fragments, MAC protocol data units (MPDU), are then transmitted independently and separately acknowledged in a fragment burst. During the burst, transmission will continue until all fragments of a packet are sent or an ACK is not received. The first fragment and its corresponding ACK act as virtual channel reservation scheme by using the duration field in their MAC headers [5]. This field defines the duration of the next transmission. Thus, overhearing neighbor STA update their network allocation vectors (NAV). For this reason collisions within a fragment burst are rare [3, 5]. Figure 21.1 shows overheads in a three fragments frame transmission.

Apart from collisions, a missed ACK is caused by a fragment being received in error or a corrupted ACK. In the former case the STA will wait for EIFS, contend for the channel, retransmit the failed fragment and continue with the remaining fragments. In this paper, we are interested in the fragment received in error. Figure 21.2 shows extra overhead incurred when a STA has to retransmit such a fragment. The crossed box means a fragment was sent but received in error and

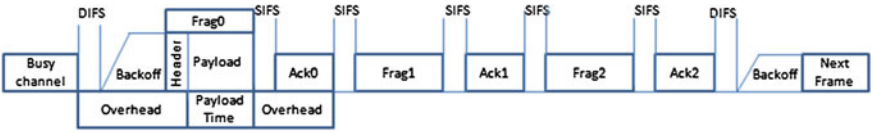


Fig. 21.1 Classical fragmentation overheads (successful frame transmission)

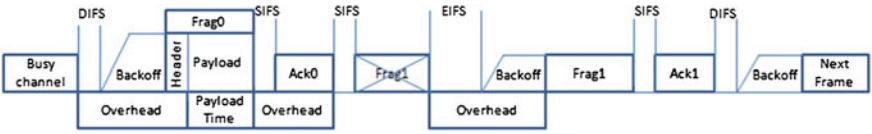


Fig. 21.2 Classical fragmentation contention overhead (successful fragment retransmission)

therefore an ACK was not received at the transmitting STA. It is worth noting that, the frame in Fig. 21.2 has only two fragments, Frag0 and Frag1, whereby Frag1 is received in error. Therefore, the transmitting STA has to wait for EIFS, contend for channel (the extra backoff in Fig. 21.2) and finally retransmit Frag1. In this specific example, it is retransmitted successfully.

The retransmission attempt will continue for a prefixed maximum number of retries (MaxSLRC), each being preceded by channel contention. When all fragments of a packet are successfully sent or the MaxSLRC is exceeded (whereby the packet will be discarded at the transmitting STA) the STA will contend for the channel before sending the next packet. We call this classical fragmentation scheme (CF). The extra contention overhead in Fig. 21.2 is incurred for every fragment that is received in error.

21.3 Backoff-Free Fragment Retransmission Fragmentation

We will look at the fragment burst in detail. During channel induced errors, the receiver silently discards the erroneous fragments [13]. This means the receiving STA is aware of fragments received with errors but lets the transmitting STA wait for EIFS after which it contends for the medium as shown in Fig. 21.2. Recall virtual channel reservation by the previous fragment and its ACK. Implicitly, a fragment can be retransmitted within EIFS without incurring contention overhead and at a lower probability of collision. We, therefore, task the receiver to inform the sender of error fragments received so that they can be retransmitted, at least once, within EIFS. The retransmission is done without contending for the channel and will continue until ACK is received or MaxSLRC is exceeded. We call this scheme backoff-free fragment retransmission (BFFR). Our scheme relies on the

fact that the receiving station decides whether a fragment is received in error or not as long as it can be decoded.

In CAM the STA keeps sensing the physical channel and thus the wireless interface radio is consuming energy during backoff, DIFS, SIFS and EIFS periods even though it is not necessarily receiving packets destined to it. These durations are necessary to prevent multiple stations from owning the medium immediately after the completion of the preceding transmission [3]. We argue that, during a fragment burst, backoff overhead can be reduced. BFFR is meant to save energy consumed during backoff by retransmitting a fragment received in error without contending for the channel.

$$\text{DIFS} = \text{SIFS} + 2 * \text{SlotTime} \tag{21.1}$$

$$\text{EIFS} = \text{SIFS} + \text{ACKT} \times \text{Time} + \text{DIFS} \tag{21.2}$$

The parameters SIFS, DIFS, EIFS and slot durations in Eqs. (21.1) and (21.2) are defined and fixed per physical layer (PHY) in DCF as stipulated in IEEE802.11-2007 specification document. SlotTime refers to the slot duration and ACKT $\times$ Time refers to ACK transmission duration. The SIFS is used by the STA that has seized the medium and needs to keep it for the duration of fragments exchange. Being the shortest IFS between transmissions within the fragment burst, it gives the transmitting STA priority over other STAs that have to wait for the medium to be idle for a longer DIFS. Since transmission is required to happen only in the beginning of a slot, this further minimizes the chance of collisions within the fragment burst.

In BFFR, when a STA receives an error fragment it will notify the sending station after SIFS. This notification is sent over ACKT $\times$ Time. We assume the notification is always received error free. Even if it does not, EIFS will expire and BFFR will roll back to CF behavior. When the notification is received, a fragment is retransmitted after a SIFS. This is shown in Fig. 21.3. In contrast to Fig. 21.2, there is no overhead due to channel contention. The EIFS is used for error notification and fragment's first retransmission attempt.

To summarize, BFFR is only triggered when the first fragment of a fragment burst is successfully acknowledged and any of the following fragments is received in error. In other transmission failure scenarios, BFFR behaves exactly as CF scheme. In the Figs. 21.1, 21.2 and 21.3 the Frag0 and Ack0 refer to fragments and their corresponding ACK messages. The figures are not drawn to scale. We modify NS3 MAC model to implement this behavior.

21.4 Performance Evaluation

To demonstrate the energy efficiency of our approach we simulate CF and BFFR in a fading channel at varying sizes of fragmentation thresholds and at different number of STAs. We used NS3.14.1. The fading channel is modeled by chaining Nakagami to a three log distance propagation model. The STAs in our simulations

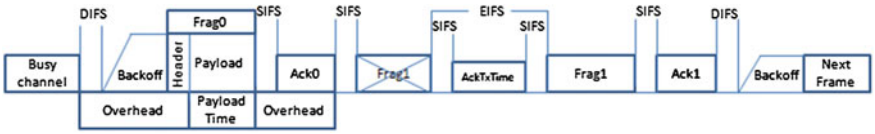


Fig. 21.3 BFFR scheme: no contention overhead (successful fragment retransmission)

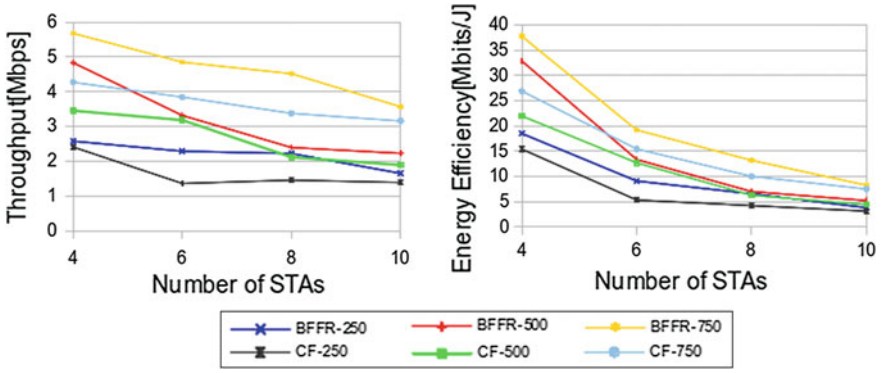


Fig. 21.4 BFFR throughput and energy efficiency in a fading channel

are randomly placed in a circle centered at the access point (AP). The STA to AP distance varies randomly from 10 to 50 m.

The PHY is set to IEEE 802.11 g with SIFS, DIFS, slot duration and minimum contention window values set to 10 μs, 50 μs, 20 μs and 15 slots respectively.We set the MaxSLRC to the default value of 7. The PHY data transmission rate is 54 Mbps while control information is sent at 24 Mbps.We use packets of 1,500 B and generate traffic at a constant bit rate (CBR) of 10 Mbps. To model energy consumption, we use the R-V battery model with its default parameter configurations and the provided device energy model presented in [15].

To compare CF and BFFR schemes we define two performance metrics, namely, throughput and energy efficiency. We define and measure throughput as the total number of useful data bits successfully received per unit time. Energy efficiency is defined in a similar way to [8, 10, 12] as the number of useful bits of information successfully received per one Joule of energy consumed. We measure energy efficiency as in Eq. (21.3). Where T is simulation time, in seconds, and E_t is total energy consumed in achieving the network throughput over T. E_t is measured in Joules.

$$\text{Energy Efficiency [Mbits/J]} = (\text{Throughput} * \text{T})/E_t \tag{21.3}$$

As it can be seen in Fig. 21.4, at a given fragmentation threshold, our BFFR scheme outperforms the CF in both throughput and energy efficiency. In cases where the retransmission attempt succeeds, BFFR saves the time that would

otherwise be wasted during backoff. The saved time is spent in squeezing bits into the channel thus improving overall network throughput.

For the individual transmitting STA, the energy spent in overhearing while in the backoff stage is saved. As pointed out in Sect. 21.3, during backoff the STA is continuously monitoring the medium. It is thus in reception or idle mode and consumes energy accordingly. The BFFR scheme saves this energy and thus the observed improved energy efficiency as compared to CF.

21.5 Conclusion

We have presented a simple, yet, effective scheme for enhancing MAC layer fragmentation in terms of throughput and energy efficiency. The proposed scheme can be used in other MAC enhancement schemes to further make MAC layer energy efficient while maintaining the achievable throughput performance. In our future work we aim at further improving this scheme together with addressing other fragmentation overheads mentioned in Sects. 21.1, 21.2 and in Fig. 21.1 by considering dynamic fragmentation and rate adaptation, among other techniques.

Immediate further investigation includes studying the impact of MaxSLRC on energy efficiency. In the extended version of this paper we will present delay and fairness performance of BFFR as compared to CF.

References

1. He X, Frank YL, Jiaru L (2008) Link adaptation with combined optimal frame size and rate selection in error-prone 802.11n networks. In: IEEE international symposium on wireless communication systems 2008 (ISWCS'2008), pp 733–737
2. Tsao SL, Chung-Huei H (2011) A survey of energy efficient MAC protocols for IEEE 802.11 WLAN. Comput Commun 34(1):54–67
3. Filali F (2005) Link-layer fragmentation and retransmission impact on TCP Performance in 802.11-based networks. In: IFIP mobile and wireless communications networks conference (MWCN), Marrakech, Marocco
4. Tourrilhes J (2001) Fragment adaptive reduction: coping with various interferers in radio unlicensed bands. In: IEEE international conference on communications, 2001 (ICC 2001), vol 1
5. Kim B-S, Yuguang F, Tan FW, Younggoo K (2005) Throughput enhancement through dynamic fragmentation in wireless LANs. IEEE Trans Veh Technol 54(4):1415–1425
6. Chang Y, Lee CP, Kwon B, Copeland JA (2007) Dynamic optimal fragmentation with rate adaptation for goodput enhancement in WLANs. In: ICWN, pp 12–15
7. Toorisaka W, Hasegawa G, Masayuki M (2012) Power consumption analysis of data transmission in IEEE 802.11 multi-hop networks. In: ICNS 2012, The eighth international conference on networking and services, pp 75–80
8. Serrano P, Andres G, Matthias H, Albert B (2010) On the energy efficiency of ieee 802.11 WLANs. In: IEEE, European wireless conference 2010 (EW'2010), pp 932–939

9. Zhao J, Zihua G, Wenwu Z (2003) Power efficiency in IEEE 802.11 a WLAN with cross-layer adaptation. In: IEEE international conference on communications, 2003 (ICC'03), vol 3. pp 2030–2034
10. Ramos N, Debashis P, Sujit D (2003) Energy-efficient link adaptations in IEEE 802.11b wireless LAN. In: International conference on wireless and optical communications
11. Rantala E, Arto K, Seppo G, Pasi S (2009) Modeling energy efficiency in wireless internet communication. In: ACM workshop on networking, systems, and applications for mobile handhelds, pp 67–68
12. Lettieri P, Curt S, Mani S (1999) Adaptive link layer strategies for energy efficient wireless networking. Wireless Netw 5(5):339–355
13. Vazifehdan JR, Venkatesha P, Martin J, Ignas N (2012) An analytical energy consumption model for packet transfer over wireless links. IEEE Commun Lett 16(1):30–33
14. Sidelnikov A, Jeonggyun Y, Sunghyun C (2006) Fragmentation/aggregation scheme for throughput enhancement of IEEE 802.11n WLAN. In: IEEE, The 3rd Asia Pacific wireless communications symposium 2006 (APWCS'2006), Daejeon, Korea
15. Wu H, Sidharth N, Radha P (2011) An energy framework for the network simulator 3 (NS-3). In: ICST, The 4th international conference on simulation tools and techniques 2011 (ICST'2011), pp 222–230

Chapter 22
Fault Classification of an Induction Motor Using Texture Features of Vibration Signals

Won-Chul Jang, Myeongsu Kang and Jong-Myon Kim

Abstract This paper proposes an approach for a two-dimensional (2D) representation of vibration signals for the highly reliable fault classification of an induction motor. The resulting 2D data has texture characteristics showing repetitive patterns and this paper extracts these texture patterns by utilizing gray level co-occurrence matrix. Moreover, this paper employs a distance evaluation technique in order to avoid the unnecessary fault signatures in the extracted texture features. To identify multiple faults of an induction motor, this paper finally utilizes support vector regression as a classifier using the extracted fault signatures. Experimental results indicate that the proposed approach achieves high classification accuracy.

Keywords Fault classification · Gray level co-occurrence matrix · Induction motor · Support vector regression · Texture patterns

22.1 Introduction

Induction motor having simple structure and reasonable durability is a vital component in many industrial applications. Although the induction motor is rugged, low cost, and reliable, it is often faced with severe failures [1], which may result in unscheduled downtime of electrical systems and significant economic

W.-C. Jang · M. Kang · J.-M. Kim (✉)
School of Electrical Engineering, University of Ulsan, Ulsan, South Korea
e-mail: jongmyon.kim@gmail.com

W.-C. Jang
e-mail: nasha0339@gmail.com

M. Kang
e-mail: ilmareboy@gmail.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_22, © Springer Science+Business Media Dordrecht 2014

losses. Thus, the reliable state monitoring system has become an increasingly important component in modern industrial processes. Fault classification of the induction motor is mainly performed by analyzing current signals [2] and vibration signals [3]. Current signals have been widely used, particularly to identify abnormalities of the rotor bars in the induction motor. However, it is mostly appropriate to absorb the states of electrical components. Thus, this paper utilizes vibration signals which are generally proper to well describe overall status of the induction motor.

To identify induction motor faults, it is necessary to extract features that describe the fault characteristics based on analyzing the vibration signals in the time [4], frequency [5], and time-frequency domains [6, 7]. In general, both time-domain features using statistical values and frequency-domain features have difficulty representing the characteristics of the induction motor faults due to the non-stationary nature of the vibration signals. Consequently, time-frequency features using wavelet decomposition (WD) and empirical mode decomposition (EMD) have been widely utilized in recent literatures [2, 7, 8]. Despite the fact that WD and EMD are effective time-frequency decomposition tools which are widely applied in fault diagnosis of the rotating machinery, they have a common shortcoming of node/IMF selection. As we decompose the signal into multiple nodes/IMFs by exploiting either WD or EMD, it is necessary to select effective nodes/IMFs well representing characteristic frequencies of induction motor faults in order to extract good faulty signatures. However, there is no general consensus which nodes/IMFs are the most effective for identifying induction motor faults. In this paper, we extract texture features inherent in two-dimensional (2D) data converted from one-dimensional (1D) vibration signals and identify multiple defects in the induction motor by utilizing these features as inputs of a classifier.

The paper is organized as follows. Section 22.2 explains the proposed fault classification approach of the induction motor, and Sect. 22.3 shows experimental results. Section 22.4 finally concludes this paper.

22.2 Proposed Fault Classification of an Induction Motor

The proposed fault classification system for identifying multiple faults in the induction motor includes data acquisition, data conversion, texture feature extraction, and classification.

22.2.1 Data Acquisition

The experimental setup used in this study consists of a belt, shaft, fan, and pulley, as well as a 4-pole three phase induction motor, which is 0.5 kW, 220 V, and 3,560 revolutions per minute. Four induction motors were utilized to generate the

data needed under full-load or steady-state conditions. One of the motors is healthy and is considered to be a baseline for classifying faults, including broken rotor bars, a bowed shaft, a bearing outer race fault, a rotor imbalance, and adjustable eccentricity (e.g., angular misalignment and parallel misalignment). In particular, two induction motor faults (e.g., angular misalignment and rotor imbalance) are created in the normal induction motor. Consequently, seven different vibration signals, including normal one, were employed for the fault classification of the induction motor: angular misalignment fault (AMF), broken rotor bar fault (BRF), bowed shaft fault (BSF), bearing fault (BF), normal (NO), parallel misalignment fault (PMF), and rotor unbalance fault (RUF). Likewise, vibration signals were acquired from accelerometers located in the axial, horizontal, and vertical directions of the motor bearing housing, and the acquired vibration signals were sampled at 8,000 Hz. In this study, 105 one-second-long vibration signals from the axial accelerometer for each condition were utilized.

22.2.2 Data Conversion

As mentioned in Sect. 22.1, this paper converts 1D vibration signals to 2D data to deal with problems in current literatures for extracting signatures of the induction motor defects. Figure 22.1 depicts data conversion process. The amplitude of each sample of the vibration signal is first normalized in the range from 0 to 255. The coordinate of the corresponding pixel for the ith sample of the vibration signal is pixel (j, k), where $j = floor(i/M)$, $k = modulo(i/N)$. M is the column length of a $M \times N$ data, and N is the row length of a $M \times N$ data, respectively. The normalized amplitude of each sample becomes the intensity of the corresponding pixel as illustrated in Fig. 22.1.

Since the vibration signals are sampled at 8,000 Hz and we use one-second long vibration signals (i.e., 8,000 samples) in this study, the resulting 2D data is 89×89. Figure 22.2 shows examples of 2D representation of the induction motor faults.

22.2.3 Feature Extraction Using GLCM

As shown in Fig. 22.2, the resulting 2D data has texture characteristics showing repetitive patterns. This paper extracts these texture patterns by utilizing gray level co-occurrence matrix (GLCM) that is a 2D histogram of pair of pixels which are separated by a fixed spatial relationship. Using GLCM, we extract the following nine features: contrast, correlation, energy, entropy, homogeneity, autocorrelation, dissimilarity, cluster shade, and cluster prominence whose mathematical forms of GLCM parameters are in [9]. Feature parameters have different sensitivity contributions for classification. In other words, too many input parameters for a

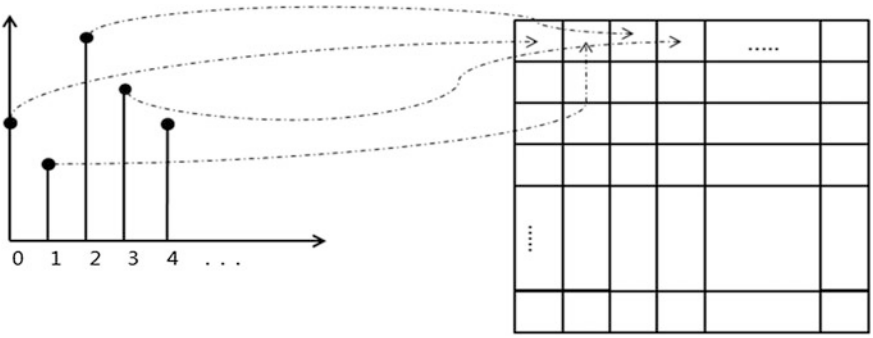


Fig. 22.1 Data conversion process of vibration signals

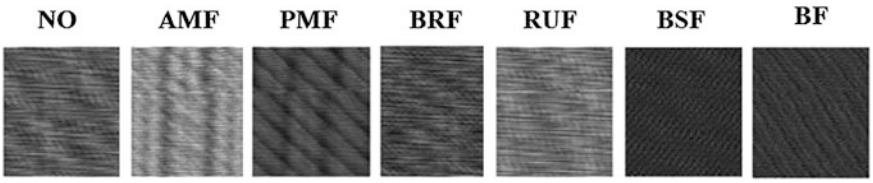


Fig. 22.2 2D representation of the induction motor faults used in this study

classifier can greatly decrease its identification accuracy. Thus, it is necessary to select efficient signatures of the induction motor faults by utilizing a distance evaluation technique (DET) [10] which is to evaluate all features in the feature vector. Sensitive features usually exhibit a small degree of variance for samples belonging to the same class and a relatively large degree for those belonging to different classes.

22.2.4 Fault Classification Using SVR

Support vector regression (SVR) was developed based on SVM [11]. Given a dataset $\{x_i, y_i\}_{i=1}^{N}$, where x_i is an input feature vector, y_i is the target value, and N is the total number of training samples. The purpose of SVR is to obtain a function which can predict the output y_i within the error limit. The function is given as follows.

$$f(x) = \sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right)(x_i, x) + b, \tag{22.1}$$

where α_i and α_i^* are the Lagrange multipliers, x_i is a support vector, x is the input feature vector, and b is bias. Likewise, the linear regression function in (22.1) is

not sufficient to process the non-linear problem, and thus a kernel function should be employed to map the input vector into high dimensional feature space. The regression function is finally derived as follows.

$$f(x) = \sum_{i=1}^{n} \left(\alpha_i - \alpha_i^*\right) K(x_i, x) + b, \tag{22.2}$$

where $K(x_i, x) = exp\left(\frac{-\|x_i - x\|^2}{2\sigma^2}\right)$ is the radial basis function (RBF) kernel, which offers high classification performance compared to other function kernels.

22.3 Experimental Results

To simulate the proposed fault classification approach, it is first necessary to build a training dataset to train SVR and select efficient fault signatures. Among 105 one-second long vibration signals for each induction motor fault, randomly selected 25 one-second long vibration signals are utilized as a training dataset in this study. The remaining signals are used as a test dataset to evaluate the proposed approach. This paper employs the DET to select efficient fault signatures of the induction motor and Fig. 22.3 illustrates the sensitivities of the nine GLCM parameters from the training dataset.

To select efficient fault signatures of the induction motor, a certain predefined threshold is utilized. If the threshold increases, the fewer sensitive signatures are selected. Moreover, different fault pattern can be well distinguished as the sufficient number of sensitive features is used. However, if the number of sensitive features decreases, the features used for representing multiple induction motor faults will weaken the linearity in high dimensional feature space mapped by the RBF kernel. To balance this problem, this paper sets the threshold value to 0.68, achieving high classification accuracy. In addition, this paper utilizes the following GLCM parameters as fault signatures: energy, entropy, and autocorrelation. Figure 22.4a shows the extracted fault signatures of the induction motor, and most of fault signatures are good for identifying multiple induction motor faults. However, the extracted features are similar between BSF and BF, which may result in low classification accuracy. As mentioned earlier, the RBF kernel is utilized to avoid performance degradation due to the nonlinear property of similar features. Figure 22.4b illustrates a classification result of the induction motor faults. Despite the fact that some misclassification occurs, the proposed approach achieves high classification accuracy.

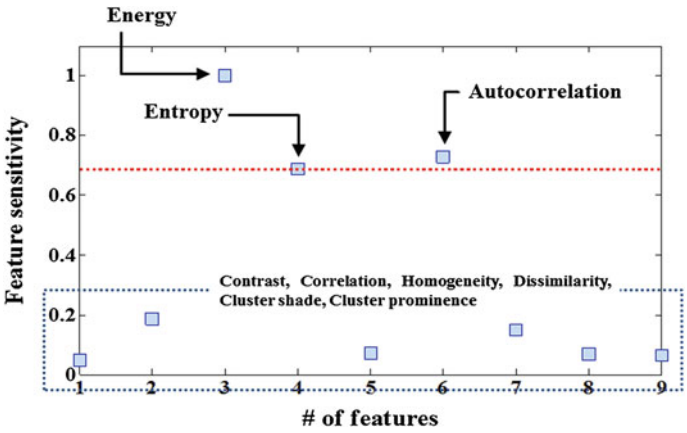


Fig. 22.3 Sensitivities of GLCM parameters from the training dataset

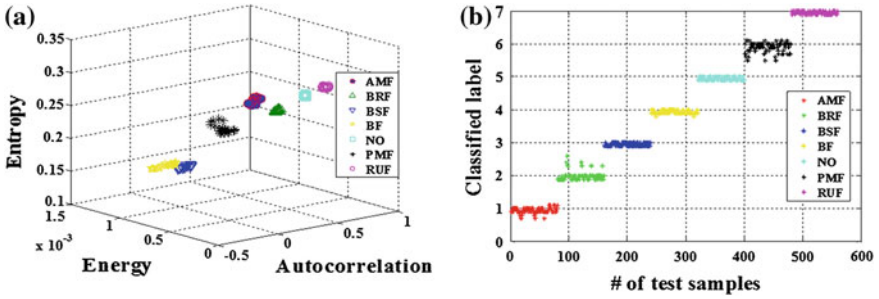


Fig. 22.4 (**a**) Selected fault signatures from the training dataset and (**b**) classification result of the proposed fault classification system (AMF: label 1, BRF: label 2, BSF: label 3, BF: label 4, NO: label 5, PMF: label 6, and RUF: label 7)

22.4 Conclusion

This paper proposed the fault classification system for identifying multiple induction motor faults. In addition, this paper utilized texture features using GLCM and 2D representation of vibration signals. To avoid the unnecessary fault signatures, this paper employed the DET and selected the three effective features, including energy, entropy, and autocorrelation. Finally, this paper utilized SVR to identify induction motor faults. Experimental results indicated that the proposed approach achieves high classification accuracy.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2012R1A1A2043644), and the Leading Industry Development for Economic Region (LeadER) grant funded the MOTIE (Ministry of Trade, Industry and Energy), Korea in 2013 (No. R0001220).

References

1. Wang H, Chen P (2009) Fault diagnosis method based on kurtosis wave and information divergence for rolling element bearings. WSEAS Trans Syst 8(10):1155–1165
2. Rodriguez-Donate C, Romero-Troncoso RJ, Cabal-Yepez E, Garcia-Perez A, Osornio-Rois RA (2010) Wavelet-based general methodology for multiple fault detection on induction motors at the startup vibration transient. J Vib Control 17(9):1299–1309
3. Zarei J (2012) Induction motors bearing fault detection using pattern recognition techniques. Expert Syst Appl 39(1):68–73
4. Bellini A, Filippetti F, Tassoni C, Capolino G-A (2008) Advances in diagnostic techniques for induction machines. IEEE Trans Ind Electron 55(1):4109–4126
5. Kia SH, Henao H, Capolino G–A (2011) Some digital signal processing techniques for induction machines diagnosis. In: IEEE international symposium on diagnostics for electric machines, power electronics and drives (SDEMPED), Bologna, pp 322–329
6. Konar P, Chattopadhyay P (2011) Bearing fault detection of induction motor using wavelet and support vector machines. Appl Soft Comput 11(6):4203–4211
7. Gaeid KS, Pind HW (2011) Wavelet fault diagnosis and tolerant of induction motor: a review. Int J Phys Sci 6(3):358–376
8. Jinde Z, Junsheng C, Yu Y (2013) Generalized empirical mode decomposition and its applications to rolling element bearing fault diagnosis. Mech Syst Signal Process 40(1):136–153
9. Soh L-K, Tsatsoulis C (1999) Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices. IEEE Trans Geosci Remote Sens 37(2):780–795
10. Lei Y, He Z, Zi Y (2009) Application of an intelligent classification method to mechanical fault diagnosis. Expert Syst Appl 36(6):9941–9948
11. Smola AJ, Scholkopf B (2004) A tutorial on support vector regression. Stat Comput 14(3):199–222

Chapter 23
An Efficient Audit Scheme for Cloud Storage

Ohmin Kwon, Dongyoung Koo, Yongjoo Shin and Hyunsoo Yoon

Abstract As Internet technology has grown rapidly, a cloud storage service becomes more very popular and widespread recently. Since there is no guarantee for integrity of data in the cloud storage, audit mechanisms for outsourced data were proposed. However, the conventional audit process requires a lot of computation overheads due to the complex operations used in tag generation and verification. In this paper, we propose a light-weight audit scheme for the cloud storage with efficient cryptographic operations. While keeping high efficiency, our proposed scheme guarantees the consistency of data in the cloud storage. To evaluate the performance of our scheme, we measure the computation costs and the time required for audit process. As a result, our proposed scheme can provide correctness of auditing for outsourced data by an efficient manner.

Keywords Audit · Data integrity · Cloud storage · Efficiency

O. Kwon (✉) · D. Koo · Y. Shin · H. Yoon
Department of Computer Science, Korea Advanced Institute of Science and Technology (KAIST), Yuseong, Daejeon 305-701, Republic of Korea
e-mail: omkwon@nslab.kaist.ac.kr

D. Koo
e-mail: dykoo@nslab.kaist.ac.kr

Y. Shin
e-mail: yj_shin@nslab.kaist.ac.kr

H. Yoon
e-mail: hyoon@nslab.kaist.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_23, © Springer Science+Business Media Dordrecht 2014

23.1 Introduction

Cloud computing has developed rapidly and it is recognized as a promising paradigm to create new various computing environments. Cloud Service Providers (CSPs) provide computing resources to network-connected users in remote locations. A cloud storage service supported by a CSP makes users to store private data in the cloud storage and access them anywhere through network. The management of data in the cloud storage is delegated to the third-party CSP. Thus, users have no way to guarantee the integrity of data in the cloud storage. It is because the CSP may not notify loss and modification of data caused by external attacks, software/hardware failures, inside threats, and so on. In order to address these issues, audit mechanisms were proposed for reliability of the cloud storage.

Many studies have been conducted for verifying integrity of data in cloud storages [1–6]. Some approaches make audit process to be executed by users [1, 2], while others do by a third-party auditor (TPA) [3–6]. Provable Data Possession (PDP) which verifies the integrity of outsourced data in untrusted storage is first introduced by Ateniese et al. [1]. It audits for data blocks randomly selected from outsourced entire data. This sample auditing method is utilized in other works [3–6]. For the consistency of outsourced data, it is required to audit them periodically or on demand. However, due to the complex operations for a tag generation and verification, there are a plenty of computational overheads for users and TPA.

To reduce such computational overheads, we propose a light-weight audit mechanism by employing simple operations. The tag generation and verification of the proposed scheme make the audit process more efficient by reducing the number of the complex operations which require more computation overheads. We measure computation costs and the time required for audit process. The results indicate that the proposed scheme requires less computation costs and shorter time compared with conventional approach. Therefore, the proposed scheme can improve the service quality of cloud storages and is suitable for resource-limited environment.

The rest of this paper is structured as follows. In Sect. 23.2, the proposed scheme is described. Sections 23.3 and 23.4 provide the security analysis and the performance evaluation, respectively. Lastly, we make a conclusion in Sect. 23.5.

23.2 The Proposed Scheme

In this section, we explain our proposed scheme in a per-procedure manner. When the CSP initiates a storage service, it generates public parameters and distributes them to users and a TPA. Before a user utilizes the storage service, he generates secret keys and public keys which used in tag generation and verification respectively. Then, he generates tags using the secret keys and uploads data and tags. After the upload, he can delegate audit process to the TPA by forwarding public keys. Then, the TPA can audit outsourced data by using Challenge-Response protocol (Fig. 23.1).

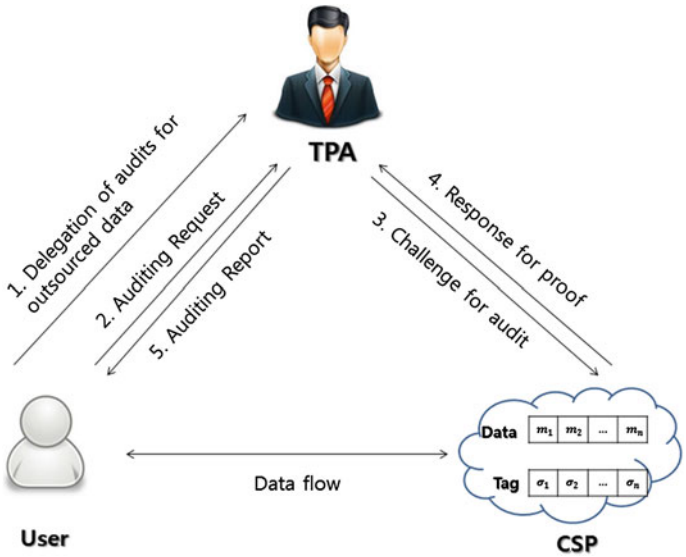


Fig. 23.1 There are three entities in our system model: user, CSP, and TPA

23.2.1 Bilinear Map

Let G_1, G_2 be multiplicative cyclic groups of prime order p, and g be a generator of G_1. Then a bilinear map e satisfies the following properties:

1. Bilinearity: for all $u, v \in G_1$ and $a, b \in Z_p$, $e(u^a, v^b) = e(u, v)^{ab}$.
2. Non-degeneracy: $e(g, g) \neq 1$.
3. Computability: there exists an efficiently computable algorithm e satisfying the above properties such that e: $G_1 \times G_1 \rightarrow G_2$.

23.2.2 Definition

The proposed scheme Π is composed of following six algorithms such as $\Pi = \{Setup, KeyGen, TagGen, ChalGen, RespGen, Verify\}$

$Setup(\gamma) \rightarrow param$: On security parameter γ, the CSP generates public parameters *param* and publicizes them.

$KeyGen(param) \rightarrow (sk, pk)$: A user takes param as input and creates secret keys $sk = \{sk_1, sk_2\}$ used in tag generation. Public keys $pk = \{pk_1, pk_2\}$ are generated from sk, and are used in verification.

$TagGen(m_i, id_i, sk) \rightarrow \sigma_i$: A user generates a tag σ_i of data block m_i using sk and id_i which is an identifier of m_i.

ChalGen$(n) \rightarrow \{l, r_l\}_{l \in L}$: The TPA chooses a random subset L of {1,…,n} and generates random numbers $(r_l)_{l \in L}$ for a challenge. The variable n represents the number of data blocks.

RespGen$\left(M, \sigma, \{l, r_l\}_{l \in L}\right) \rightarrow (\alpha, \beta)$: The CSP generates a response (α, β) for a challenge from the TPA.

Verify$\left(\{l, r_l\}_{l \in L}, (\alpha, \beta), pk, \{id_l\}_{l \in L}\right) \rightarrow$ *True* or *e*: It verifies that a response from the CSP is consistent with a challenge. If they are consistent, it outputs *True*. Otherwise, it outputs *False*.

23.2.3 Construction

Let G_1, G_2 be multiplicative cyclic groups of prime order p, g be a generator of G_1, $H : \{0,1\}^* \rightarrow Z_p$ be a collision-resistant hash function, and $e : G_1 \times G_1 \rightarrow G_2$ be a bilinear map. We denote the uploaded data $M = \left\{m_{ij}\right\}_{i \in [1,n], j \in [1,s]}$ such that $m_{ij} \in Z_p$ and tags $\sigma = \{\sigma_i\}_{i \in [1,n]}$.

Setup: When the CSP initiates a storage service, the CSP chooses multiplicative cyclic groups G_1 and G_2 of primer order p with g for a bilinear map e by running *Setup* algorithm. Then it delivers $param = (G_1, G_2, p, g, e)$ to users and the TPA.

Data Upload: A user runs *KeyGen* algorithm for generating keys. It first chooses $s + 1$ random values $a \in Z_p^*$ and $sk_2 = \{sk_{2,j} = b_j \in Z_p\}_{j \in [1,s]}$, and computes $sk_1 = g^{\frac{1}{a}}$, $pk_1 = g^a$, and $pk_2 = \{pk_{2,j} = g^{b_j}\}_{j \in [1,s]}$.

Then, the user generates tags such as

$$\sigma_i = (sk_1)^{H(id_i) + \sum_{j \in [1,s]} sk_{2,j} \cdot m_{ij}} \tag{23.1}$$

for $1 \leq i \leq n$ by running *TagGen* algorithm. The user uploads (M, σ) to the cloud storage and deletes them in his local storage.

Delegation of Audit: The user delegates audit process for outsourced data to the TPA by forwarding public keys pk.

Audit for Outsourced Data: The TPA chooses subset L of {1,2,…,n} and generates a set of random numbers $\left\{r_l \in Z_p^*\right\}_{l \in L}$ by running *ChalGen* algorithm. An element in L represents an index of a challenged data block. The challenge $chal = (l, r_l)_{l \in L}$ is transmitted to the CSP and the CSP runs *RespGen* for generating a response $(\alpha = \left\{\alpha_j\right\}_{j \in [1,s]}, \beta)$ as follows:

$$\alpha_j = \sum_{l \in L} r_l \cdot m_{lj} \tag{23.2}$$

$$\beta = \prod_{l \in L} \sigma_l^{r_l} \tag{23.3}$$

The CSP transmits $(\alpha = \{\alpha_j\}_{j\in[1,s]}, \beta)$ to the TPA. And then the TPA verifies it by running *Verify* algorithm. It outputs *True* if

$$e\left(g^{\sum_{l\in L} r_l \cdot H(id_l)} \cdot \prod_{j\in[1,s]} pk_{2,j}^{\alpha_j}, g\right) = e(\beta, pk_1) \tag{23.4}$$

Otherwise, it outputs *False*.

23.3 Security Analysis

In this section, we provide the correctness of our audit scheme Π. Correctness of Π is achieved by the bilinearity property of bilinear map. The left hand side (LHS) of Eq. (23.4) expands as follows:

$$\begin{aligned}
\text{LHS} &= e\left(g^{\sum_{l\in L} r_l \cdot H(id_l) + \sum_{j\in[1,s]} sk_{2,j}\cdot\alpha_j}, g\right) \\
&= e\left(g^{\sum_{l\in L} r_l \cdot H(id_l) + \sum_{j\in[1,s]} (sk_{2,j}\cdot \sum_{l\in L} r_l \cdot m_{lj})}, g\right) \\
&= e\left(g^{\sum_{l\in L} \left(r_l \cdot H(id_l) + \sum_{j\in[1,s]} (sk_{2,j}\cdot r_l \cdot m_{lj})\right)}, g\right)
\end{aligned}$$

The right hand side (RHS) of Eq. (23.4) expands as follows:

$$\begin{aligned}
\text{RHS} &= e\left(\prod_{l\in L} \sigma_l^{r_l}, pk_1\right) \\
&= e\left(\prod_{l\in L} \left((sk_1)^{H(id_l) + \sum_{j\in[1,s]} sk_{2,j}\cdot m_{lj}}\right)^{r_l}, g^a\right) \\
&= e\left(\prod_{l\in L} \left(g^{H(id_l) + \sum_{j\in[1,s]} sk_{2,j}\cdot m_{lj}}\right)^{\frac{r_l}{a}}, g^a\right) \\
&= e\left(g^{\sum_{l\in L} \left(r_l \cdot H(id_l) + \sum_{j\in[1,s]} \left(sk_{2,j}\cdot r_l \cdot m_{lj}\right)\right)}, g\right)
\end{aligned}$$

Since the LHS and the RHS of Eq. (23.4) are same, the proof completes.

Table 23.1 Notation

Notation	Description	Notation	Description	Notation	Description
S	The number of sectors in a data block	L	The number of challenged data bocks	H_{Z_p}	Hash from $\{0, 1\}^*$ to Z_p
H_G	Hash from $\{0, 1\}^*$ to G	$Pair$	Paring operation	Exp_G	Exponentiation in G
Mul_G	Multiplication in G	Mul_{Z_p}	Multiplication in Z_p	Add_{Z_p}	Addition in Z_p

Table 23.2 Computation cost

	Our scheme	CPOR
Tag generation	$\mathrm{Exp_G} + \mathrm{S} \cdot \mathrm{Mul_{Z_p}} + \mathrm{S} \cdot \mathrm{Add_{Z_p}} + \mathrm{H_{Z_p}}$	$(\mathrm{S}+1) \cdot \mathrm{Exp_G} + \mathrm{S} \cdot \mathrm{Mul_G} + \mathrm{H_G}$
Verification	$2 \cdot \mathrm{Pair} + (\mathrm{S}+1) \cdot \mathrm{Exp_G} + \mathrm{S} \cdot \mathrm{Mul_G}$ $+\mathrm{L} \cdot \mathrm{Mul_{Z_p}} + (\mathrm{L}-1) \cdot \mathrm{Add_{Z_p}} + \mathrm{L} \cdot \mathrm{H_{Z_p}}$	$2 \cdot \mathrm{Pair} + (\mathrm{L}+\mathrm{S}) \cdot \mathrm{Exp_G}$ $+(\mathrm{L}+\mathrm{S}-1) \cdot \mathrm{Mul_G} + \mathrm{L} \cdot \mathrm{H_G}$

23.4 Performance Evaluation

In this section, we provide a comparison between our scheme and CPOR [3] which is a base scheme for recent works [4–6]. And experimental results demonstrate the efficiency of our scheme (Table 23.1).

23.4.1 Computation Cost

Computation costs for tag generation and verification are described in Table 23.2. Our scheme requires one Exp_G operation in tag generation, while CPOR requires $(S+1) \cdot Exp_G$. When the TPA verifies a response from the CSP, the number of Exp_G and Mul_G operations are just dominated by the number of sectors in our scheme. However, they are linearly proportional to both the number of sectors and the number of challenged data blocks in CPOR.

23.4.2 Experimental Results

We use Pairing Based Cryptography (PBC) library to simulate our scheme. The experiment environment is Ubuntu 12.04 with Intel Core i3 3.10 GHz Processor and 4 GB Memory. We assume that $|p| = 160$ bits.

Figure 23.2 shows the performance of tag generation. Tag generation time in our scheme is 3.12 ms per block when $S = 1$, and 3.3 ms per block when $S = 100$.

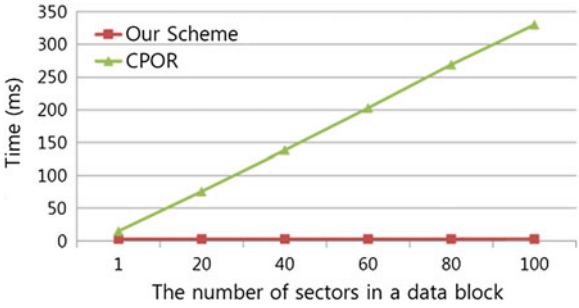


Fig. 23.2 Impact of S on tag generation time

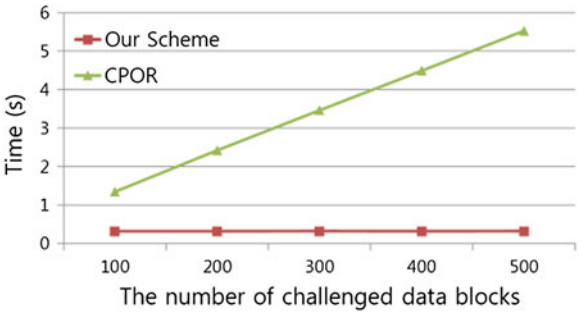


Fig. 23.3 Impact of L on verification time, when $S = 100$

CPOR requires 15 ms per block when $S = 1$, and 330 ms per block when $S = 100$. As depicted in Fig. 23.3, the verification time is dependent on L in CPOR. However, our scheme requires averagely 0.32 s regardless of L.

23.5 Conclusion

In this paper, we propose an efficient audit scheme for a cloud storage service. The proposed scheme is able to verify integrity of data in the cloud storage by employing light-weight operations. The security analysis and performance evaluations show that the proposed scheme is enough to address the important requirement of auditing for the data in the cloud storage.

Acknowledgments This work was supported by the IT R&D program of MKE/KEIT [10041244, SmartTV 2.0 Software Platform] and the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MEST) (No. 2011-0016584).

References

1. Ateniese G, Burns R, Curtmola R, Herring J, Kissner L, Peterson Z, Song D (2007) Provable data possession at untrusted stores. In: Proceedings of the 14th ACM conference on computer and communications security, CCS '07, ACM, New York, USA, pp 598–609
2. Juels A, Kaliski Jr, BS (2007) Pors: proofs of retrievability for large files. In: Proceedings of the 14th ACM conference on computer and communications security, CCS '07, ACM, New York, USA, pp 584–597
3. Shacham H, Waters B (2008) Compact proofs of retrievability. In: Proceedings of the 14th international conference on the theory and application of cryptology and information security: advances in cryptology. ASIACRYPT '08, Springer, Berlin, pp 90–107
4. Wang C, Chow S, Wang Q, Ren K, Lou W (2013) Privacy-preserving public auditing for secure cloud storage. IEEE Trans Comput 62(2):362–375
5. Yang K, Jia X (2013) An efficient and secure dynamic auditing protocol for data storage in cloud computing. IEEE Trans Parallel Distrib Syst 24(9):1717–1726
6. Wang Q, Wang C, Ren K, Lou W, Li J (2011) Enabling public auditability and data dynamics for storage security in cloud computing. IEEE Trans Parallel Distrib Syst 22(5):847–859

Chapter 24
High-Performance Cluster Estimation Using Many-Core Models

Junsang Seo, Myeongsu Kang, Cheol-Hong Kim and Jong-Myon Kim

Abstract This paper presents implementation of high-performance subtractive clustering on a *single instruction multiple data* (SIMD) many-core processor. Since there is no general consensus that which grain size of the many-core processor provides the maximum performance, this paper explores the effects of varying the number of processing elements (PEs) and different amount of memories by introducing image data-per-processing element (IDPE) variation as a design variable, which is the amount of image data directly mapped to each PE. In this study, five PE configurations (IDPEs = 16, 64, 256, 1,024, and 4,096) are used for evaluating the system performance in terms of execution time and system utilization. In addition, this paper compares the proposed approach with a CPU-based implementation to show the potential for the improved performance of the proposed approach. Experimental results show that the proposed approach achieves 16.73× speedup at PEs = 4,096 over the CPU-based implementation.

Keywords Cluster estimation · Fuzzy c-means · Image segmentation · Many-core architecture

J. Seo · M. Kang · J.-M. Kim (✉)
School of Electrical Engineering, University of Ulsan, Ulsan, South Korea
e-mail: jmkim07@mail.ulsan.ac.kr

J. Seo
e-mail: redcircle@mail.ulsan.ac.kr

M. Kang
e-mail: ilmareboy@mail.ulsan.ac.kr

C.-H. Kim
School of Electronics and Computer Engineering, Chonnam National University, Gwangju, South Korea
e-mail: cheolhong@gmail.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_24, © Springer Science+Business Media Dordrecht 2014

24.1 Introduction

Fuzzy c-means is one of the most well-known clustering algorithms and is widely utilized for computer aided diagnosis and therapy [1]. However, the clustering quality of FCM strongly depends on both the number of clusters and an initial value of cluster centroids. To address this problem, *Chiu* proposed a cluster estimation method based on density value of data sets, which is also called a subtractive clustering algorithm [2]. While computational complexity of the subtractive clustering algorithm only becomes a simple ratio relation with dataset and it is applicable to majority data clustering situations, it still demands tremendous computational cost which limits its use in real applications.

Among many computational models available for dealing with this issue, *single instruction multiple data* (SIMD) many-core architectures including graphics processing units (GPUs) are promising candidates. Currently, GPUs have been adopted in many image/video processing applications due to their high computational throughput capability [3]. However, inter-streaming multiprocessors (SMs) communication on GPUs is inefficiently achieved by implementing barrier synchronization via the host. The lack of such inter-SM communication largely limits data-parallel or task-parallel applications [4, 5]. Thus, this paper introduces high-performance many-core architecture and implements subtractive clustering on the many-core architecture. While it is evident that the overall performance improves by increasing the number of PEs [6], there is no general consensus which grain size of processors and memories on the many-core architecture provides the most efficient operation of subtractive clustering with regard to execution time. This paper introduces image data-per-processing element (IDPE) as a design variable, where IDPE is the amount of image data directly mapped to each PE. In addition, this paper quantitatively evaluates the effects of varying IDPE ratio using architectural and workload simulation.

The rest of this paper is organized as follows. Section 24.2 introduces background information regarding subtractive clustering and the reference many-core architecture, and Sect. 24.3 explains parallel implementation of subtractive clustering. Section 24.4 shows experimental results. Finally, Sect. 24.5 concludes this paper.

24.2 Background Information

24.2.1 Subtractive Clustering

The subtractive clustering method calculates a measure of the likelihood that each data point would define the cluster center based on the density of surrounding data points [2]. The density measurement P_i at a data point x_i is expressed by

$$P_i = \sum_{j=1}^{N} e^{-\alpha \left\| x_i - x_j \right\|^2}, \tag{24.1}$$

where N is the total number of data points, $\alpha = 4/r_a^2$, the radius r_a defines the neighborhood radius for each cluster center. After the potential of every data point is computed, the data point with the highest potential is selected as the first cluster center. Suppose that the location of the first cluster center is c_1 and its potential value is P_1^*. Then, the next cluster center is calculated as follows:

$$P_i = P_i - P_1^* \cdot e^{-\beta \left\| x_i - c_1 \right\|^2}, \tag{24.2}$$

where $\beta = 4/r_b^2$ and r_b is a positive constant, respectively. According to [2], the constant r_b should be effectively selected to avoid obtaining closely spaced cluster centers and it is set to $1.5r_a$. As the potential of all data points is revised by utilizing (24.2), the second cluster center is selected as the data point yielding the highest remaining potential values. The selection of next cluster centers is iteratively carried out until the stopping condition is achieved. If the potential of the kth data point is greater than a certain threshold as given in (24.3), the kth data point is selected as the next cluster center c_k.

$$P_k^* > \overline{\varepsilon} P_1^*, \tag{24.3}$$

where $\overline{\varepsilon}$ is a threshold for the potential above which accepts the data point as a cluster center. If the potential of the kth data point is less than the lower threshold as described in (24.4), the algorithm will be terminated.

$$P_k^* < \underline{\varepsilon} P_1^*, \tag{24.4}$$

where $\underline{\varepsilon}$ specifies a threshold for the potential below which rejects the data point. To complete Chiu's subtractive clustering, several parameters are needed to be set. Table 24.1 presents the predefined parameters [2].

The acceptance of a data point which does not satisfy conditions in (24.3) and (24.4) depends on the relative distance between x_k^* and all other cluster centers, which is defined as follows:

$$\frac{d_{min}}{r_a} + \frac{P_k^*}{P_1^*} \geq 1, \tag{24.5}$$

where d_{min} is the minimum distance between x_k^* and all previous cluster centers.

24.2.2 The Reference Many-Core Architecture

Figure 24.1 shows the reference many-core processor which consists of a two-dimensional PE array, local memory, and array control unit (ACU).

Table 24.1 Summary of parameters for completing Chiu's subtractive clustering

Parameters	Cluster radii $[r_a, r_b]$	Thresholds $[\overline{\varepsilon}\ \underline{\varepsilon}]$
Values	[0.3, 0.45]	[0.5 0.15]

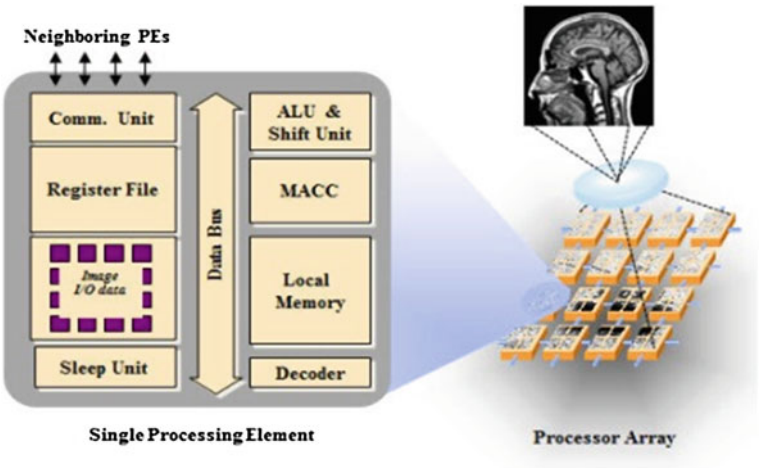


Fig. 24.1 Reference many-core processor array

PEs in the array are interconnected in a mesh network and simultaneously execute a set of instructions. Each PE supports different amounts local memories to store input images and temporary data produced during processing. Each PE has a reduced instruction set computer (RISC) datapath with the following minimum characteristics:

- Arithmetic logic unit (ALU) performs basic arithmetic and logic operations,
- Multiply-accumulator unit (MACC) multiplies 32-bit values and accumulates them into a 64-bit accumulator,
- Barrel shifter performs multi-bit logic/arithmetic shift operations,
- SLEEP unit activates/deactivates PEs based on local information,
- Communication unit allows PEs to communicate with their four nearest neighbors (north, east, west, and south),
- A small amount of 32-bit word local memory and 16 32-bit three-ported general purpose registers are included.

24.3 Parallel Implementation of Subtractive Clustering

The subtractive clustering on the reference many-core architecture can be performed by using the following five steps and more details of each step will be given below.

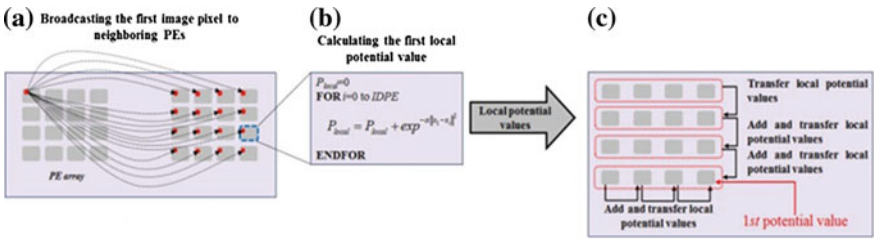


Fig. 24.2 An example of the process computing the first potential value

- **Step 1**: To perform subtractive clustering on the reference many-core processor, both parameters including cluster radii and image data are distributed to corresponding PEs, and then we normalize the image data in the range from 0 to 1.
- **Step 2**: To find the first cluster center of a $M \times N$ input image, it is first necessary to compute $M \times N$ potential values using (1) and select the location of an image pixel yielding the highest potential value as the first cluster center. Figure 24.2 illustrates an example of the procedure for computing the first potential value at the first pixel value on the many-core architecture.

To calculate the first potential value using (1), each PE should share the first pixel value by utilizing communication instructions (e.g., XFER NORTH, XFER EAST, XFER WEST, and XFER SOUTH) as depicted in Fig. 24.2a. Then, each PE simultaneously computes a local potential value by calculating exponential values and adding them up, as shown in Fig. 24.2b, by using the shift, branch and arithmetic instructions (e.g., LSH, BGE, MUL, ADD, and SUB). This process is continuously repeated until the $M \times N$ th local potential value is obtained. After computing the local potential value on a single PE, PEs need to communicate with their neighboring PEs to broadcast and add their local potential values as presented in Fig. 24.2c. Then, the PE located at bottommost right corner in the PE array stores a bigger potential value by comparing between the previous potential value and the current potential value. Thus, the PE obtains the highest potential value, where activating/deactivating PEs are necessary during transferring and summing local potential values, which calls sleep instructions (e.g., SGT, SLT, and WAKEUPI).

- **Step 3**: This step determines whether the subtractive clustering process of acquiring new cluster center is necessary or not by utilizing Eqs. (24.3)–(24.5). If either the condition in (4) is satisfied or the conditions in (3) and (5) are not satisfied, the subtractive clustering algorithm is terminated. However, if these termination conditions are not satisfied, it is necessary to compute the next cluster center for clustering.
- **Step 4**: The next cluster center is computed by using (2), which requires broadcasting the first cluster center c_1 and its potential value P_1^* to other PEs in the PE array. Similar to the process in step 2, PEs simultaneously compute their own local potential values using (2) and transfer them to the neighboring PEs.

Finally, the PE located bottommost right has the maximum potential value among $M \times N$ potential values. The step (3) then decides if the location of the image pixel yielding the maximum potential value is considered as the next cluster center.

24.4 Experimental Results

A methodology infrastructure used in this study is divided into three levels: application, architecture, and technology. At the application level, an instruction-level many-core simulator is utilized to profile execution statistics such as cycle count and PE utilization for five different PE configurations by retargeting and optimizing the subtractive clustering algorithm for each configuration. At architectural level, the architectural models of functional units for the many-core processor are produced. The many-core processor's hardware components are then implemented using the register-transfer level (RTL) Verilog. At the technology level, the hardware of the many-core architecture is synthesized in TSMC 28 nm using the Synopsys design compiler. Finally, our design space analysis tool collects and combines all parameters obtained from the application and technology levels in order to determine execution time and system utilization for each PE configuration.

24.4.1 Evaluation Metrics

Table 24.2 presents evaluation metrics including system utilization and execution time used in this study. Where N_{issued} is the total number of issued instructions, $N_{executed}$ is the total number of executed instructions on each PE, and f_{clk} is the clock frequency. Using the evaluation metrics, an optimal design space can be determined among five PE configurations, which provides the most efficient operation for the subtractive clustering algorithm in terms of the execution time. Table 24.3 shows system parameters for different PE configurations using in this study.

To evaluate the performance of the subtractive clustering algorithm on each PE configuration, we use two magnetic resonance (MR) images including different numbers of cluster centers, as shown in Fig. 24.3.

24.4.2 System Utilization

To achieve high performance in the many-core array system, it is critical to maintain high system utilization. Figure 24.4 illustrates system utilization versus

Table 24.2 Summary of evaluation metrics

System utilization	$U = \left(N_{executed}/N_{issued}\right) \times 100\,[\%]$	Execution time	$t_{exec} = \frac{N_{issued}}{f_{clk}}$

Table 24.3 System parameters for different five PE configurations

Parameters	Values				
Clock frequency			1 GHz		
VLSI technology			TSMC 28 nm		
PEs	16	64	256	1,024	4,096
IDPE	4,096	1,024	256	64	16
Memory/PE [KB]	131.072	32.768	8.192	2.048	0.512

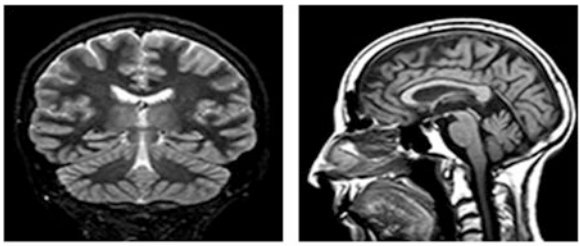


Image 1 (# of clusters = 3) **Image 2 (# of clusters = 5)**

Fig. 24.3 256 × 256 MR images with different numbers of cluster centers

the IDPE, where system utilization is calculated by the average number of active PEs to complete the subtractive clustering algorithm with varying numbers of PEs.

As depicted in Fig. 24.4, system utilization for the subtractive clustering algorithm is higher than 90 % due to the data-parallel nature of the algorithm. However, as the IDPE decreases (or the number of PEs increases), a higher percentage of communication and sleep instructions are required to broadcast image pixels and local potential values to the neighboring PEs, resulting in degradation of the system utilization. An interesting observation in Fig. 24.4 is that the shape of curves is very similar, even if the number of cluster centers of two MR images is different. This is because arithmetic instructions (e.g., MUL and ADD) are highly utilized for computing potential values for every image pixel regardless of the IDPE.

24.4.3 Execution Time

Figure 24.5 shows execution time of the subtractive clustering algorithm for different IDPE values. As expected, execution time of the subtractive clustering algorithm monotonically decreases as the IPDE decreases. In addition, the proposed many-core approach achieves about 16.73× speedup at PEs = 4,096

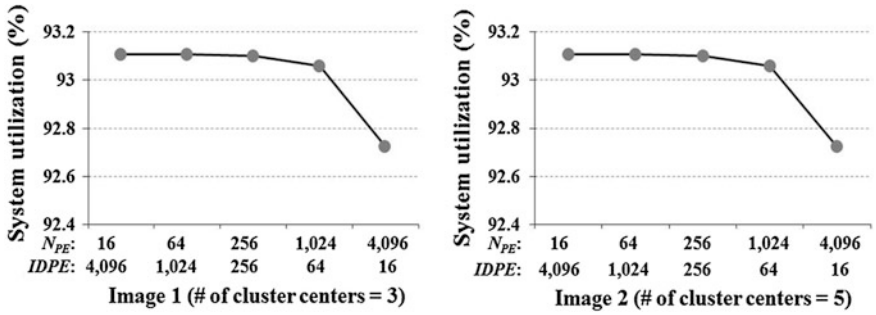


Fig. 24.4 System utilization with varying IDPE values for MR images with different number of cluster centers

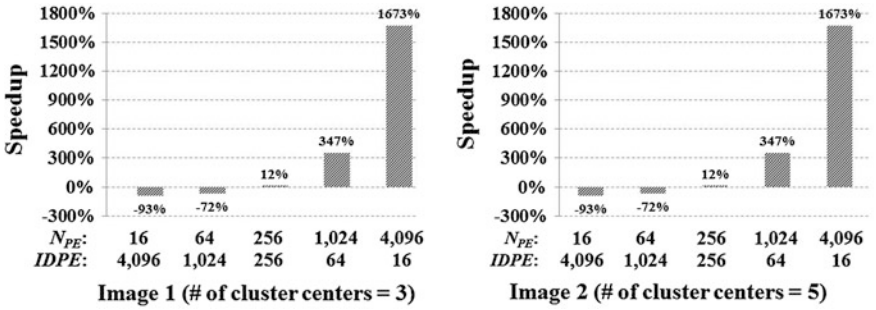


Fig. 24.5 Percentage of speedup with different PE configurations for MR images

compared to an equivalent sequential CPU implementation. For the CPU implementation, the subtractive clustering algorithm was implemented by using MATLAB b2012 on the Quad core CPU.

24.5 Conclusion

This paper presented high-performance subtractive clustering using many-core models. To find an optimal many-core model, this paper introduced image data-per-processing element (IDPE) as a design variable and quantitatively evaluated the effects of varying the IDPE ratio using architectural and workload simulation. Experimental results showed that the most efficient operation for the subtractive clustering algorithm is achieved at the number of PEs = 4,096 (or IDPE = 16) in terms of execution time.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2013R1A2A2A05004566).

References

1. Qing Y, Dongxu Z, Feng T (2010) An initialization method for fuzzy C-means algorithm using subtractive clustering. In: International conference on intelligent networks and intelligent systems, Shenyang, pp 393–396
2. Chiu SL (1994) Fuzzy model identification based on cluster estimation. J Intell Fuzzy Syst 2:267–278
3. Park IK, Singhal N, Lee MH, Cho S, Kim CW (2011) Design and performance evaluation of image processing algorithms on GPUs. IEEE Trans Parallel Distrib Syst 22(1):91–104
4. Xiao S, Feng W–C (2010) Inter-block GPU communication via fast barrier synchronization. In: IEEE international symposium on parallel and distributed processing, Paris, pp 1–12
5. Feng W–C, Xiao S (2010) To GPU synchronize or not GPU synchronize? In: IEEE international symposium on circuits and systems, Paris, pp 3801–3804
6. Woo D-H, Lee H-S (2008) Extending Amdahl's law for energy-efficient computing in the many-core era. IEEE Comput 41(12):24–31

Chapter 25
Dynamic Processing of Floating Continuous Query for Cold Chain Management

Bonghee Hong, Kyungmin Lee and Yongxu Piao

Abstract When some products move to other storage in cold chain, the related continuous queries should be also redefined for different storages. To avoid redefining the existing continuous queries, we introduce a new concept of floating continuous query (FCQ) for solving the problem of redefinition of the transformed continuous queries. Dynamic processing of FCQ not only simplifies query management but also improves query processing performance.

Keywords Continuous query · Sensor · RFID · Cold chain management

25.1 Introduction

Both RFID and sensors can be used for monitoring of storage condition of product in cold chain management [1, 2]. RFID tags are used to keep track of products for product storage location. Figure 25.1 illustrates the movement of products among different warehouses in which temperature and humidity should be continuously monitored.

Continuous query [3] is an efficient way to detect alert event. Each continuous query defines threshold on storage condition for a product. We say user-defined continuous query as original continuous query (OCQ). For example, "temperature

B. Hong · K. Lee (✉) · Y. Piao
Department of Electrical and Computer Engineering, Pusan National University, Busan, South Korea
e-mail: ykm4223@pusan.ac.kr

B. Hong
e-mail: bhhong@pusan.ac.kr

Y. Piao
e-mail: ywpark@pusan.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_25, © Springer Science+Business Media Dordrecht 2014

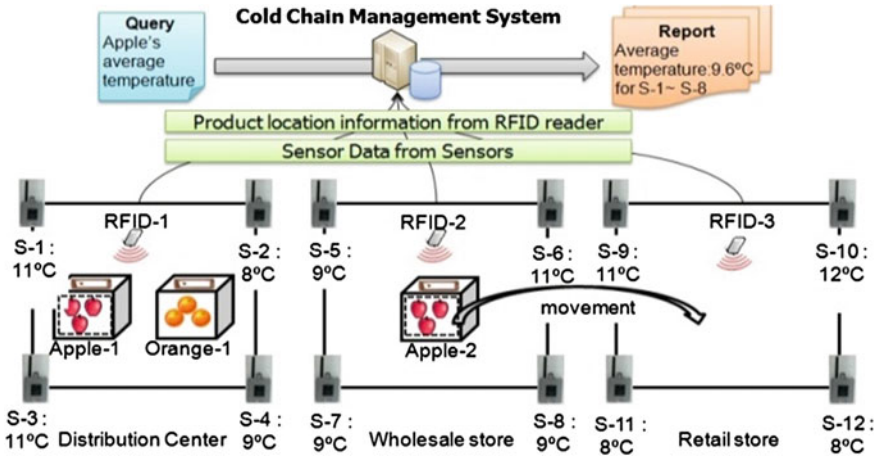


Fig. 25.1 RFID tag and sensor tag application in cold chain management

of orange greater than 10 °C" is OCQ on orange's temperature. Its SQL-like representation is as follows;

Select orange's temperature
From S sensor tag data stream
Where orange's temperature > 10 °C

It is obvious that OCQ cannot be directly executed. The reason is that OCQ has no information about sensors for checking storage condition of orange. Naive approach to execution of OCQ is that OCQ should be transformed into sensor based continuous query. Figure 25.2 presents an example of transformation of OCQ.

Transformed continuous query (TCQ) would be based on product location and sensor location information. Its SQL-like representation is as follows;

Select average(S.temperature)
From S sensor tag data stream
Where S.ID in (s1–s4) and average(S.temperature) > 10 °C

TCQ is a sensor based continuous query transformed from OCQ by means of location of product. For example, when some of product is moved to another location, the existing TCQ should be redefined for the sensors of a new location. As we can see through the example, it is difficult for TCQ to continuously monitor the storage condition for the movement of product in cold chain environment.

In this study, we introduce a new concept of floating continuous query (FCQ) for solving the problem of re-transformation of continuous query. FCQ is a conceptual view of continuous query independent of product storage. FCQ would be dynamically decomposed into a number of storage bounded continuous queries

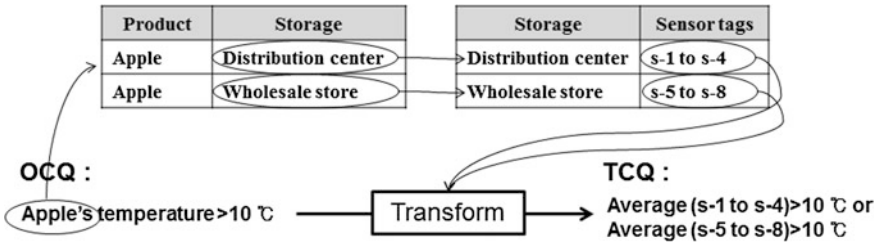


Fig. 25.2 OCQ transformation

(SBCQ). In case of changing of product's storage, SBCQ would be dynamically added (deleted) instead of redefining TCQ. The benefit of dynamic processing of FCQ is to improve the performance of continuous queries.

25.2 Naïve Approach

Naïve approach to monitoring of product storage condition is the transformation of OCQ with considering sensor-based queries. The Fig. 25.3 shows an example of processing of TCQ-1 and TCQ-2. TCQ execution has three steps: (1) sensor tag data collection (2) average computation (3) threshold evaluation. In first step, TCQ-1(TCQ-2) collects sensor tag data whose sensor tag ID is from s-1 to s-8 (s-5–s-8). Second step is to compute the average of collected sensor tag data. Finally the average value is evaluated by threshold on apple's temperature. Since average of (s-1–s-4) is greater than 10, alert report is generated as query result.

Main problem of naïve approach is frequent re-transformation of continuous query caused by product movement. When a product is moved to other storage, TCQ has to be redefined in order to monitor measurements of sensor tags in new storage of product. The TCQ approach has also difficulty of handling of distributed movement of products because of not being able to compute different aggregated values for different storages.

25.3 Processing of Floating Continuous Query

25.3.1 Overview of Our Approach

Our key idea is to support computing different aggregate values for different product storages. We defined a conceptual continuous query independent of product storage as a Floating Continuous Query (FCQ). To implement FCQ, FCQ should be in runtime decomposed into SBCQ. Their definitions are as follows:

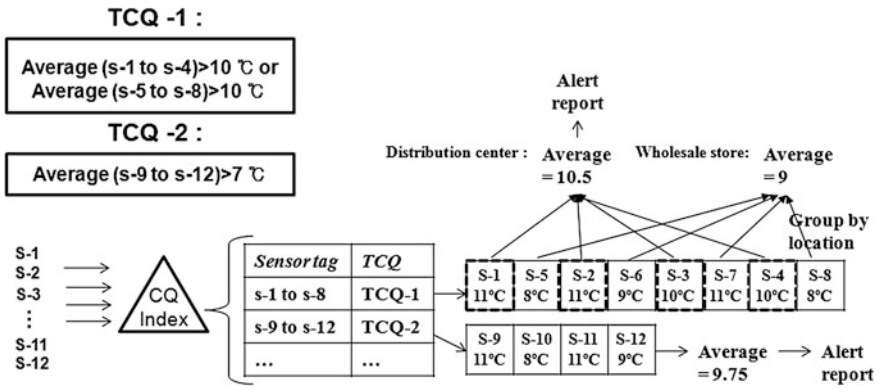


Fig. 25.3 TCQ execution flow

Definition 1. Floating Continuous Query (FCQ). A Floating Continuous Query $FCQ = (p, st)$ defines status threshold (st) of product p. It continuously monitors current status threshold st of product p.

For example, a FCQ is (*apple, temperature* > *10*). This FCQ attempts to monitor status threshold of apple. Since the apple currently is distributed in distribution center (DC) and wholesales store (WS). The FCQ is decomposed into SBCQ-1 (DC, temperature > 10) and SBCQ-2 (WS, temperature > 10).

Definition 2. Storage Bounded Continuous Query (SBCQ). A Storage Bounded Continuous Query $SBCQ = (s, st)$ defines status threshold st on a certain storage s. It continuously monitors status of storage s to checks whether the status of certain storage s violates status threshold st.

Each SBCQ execution monitors the current storage of a specific product. Firstly, each SBCQ collects sensor tag data of the corresponding storage. A group of sensor tag data, which belongs to the same storage, should be stored in one data queue. Secondly, each SBCQ is to compute the average value of sensor tag data in the queue.

For upcoming stream of sensor data, we use predicates on sensor tag to collect sensor tag data. As shown in Fig. 25.4, distribution center has sensor tags (s-1–s-4). All sensor tag data generated by sensor tags (s-1–s-4) are collected in data queue of distribution center.

25.3.2 Dynamic Decomposition of FCQ

FCQ should be dynamically decomposed and updated for product movement. The deletion and addition of SBCQ is driven by product movement. There are two movement types, total movement and partial movement. *Total movement* is a movement type which the entire product is moved to the other storage, while

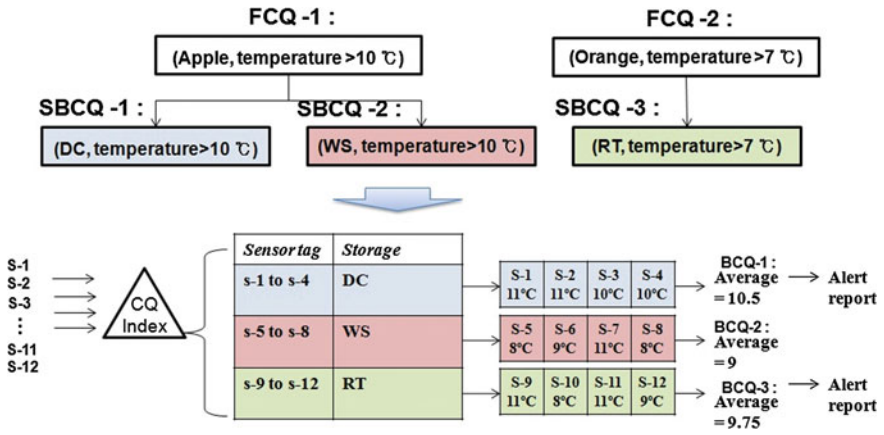


Fig. 25.4 FCQ processing flow

partial movement only moves some part of product to other storages. Different types of product movement lead to different updates on SBCQ.

When a product is totally moved to a new storage, the SBCQ needs not to monitor the original storage any more. For example, apple originally is stored in wholesale store. SBCQ monitors the status of wholesales store. When the apple moves to a retailer store, SBCQ should be updated for monitoring status of the retailer store.

In case of partial movement, the product is divided into two or more storages. SBCQ should not only monitor the original storage but also new storages. To monitor new storage status, a new SBCQ is required to monitor status of new storages (Fig 25.5).

Algorithm 1 describes SBCQ update process. When product p moves from storage A to storage B, the data queue for storage B firstly is created (line 5). If there already is a SBCQ for monitoring product p in storage B, it needs not to generate a new SBCQ for monitoring storage B for product p (line 14–18). Otherwise, a new SBCQ (p, B) is created to monitor storage B for product p (line 9, line 13). If product movement is total, it is not necessary to monitor the old storage A. The algorithm deletes the SBCQ (p, A) for monitoring storage A (line 9, line 16). In addition, the algorithm also manages the related data queue. If the data queue is not used by any other SBCQ, the data queue is deleted (line 10, line 17).

25.4 Performance Evaluation

To verify floating continuous query processing, we conducted two experiments for executing FCQ. In first experiments, we compare query processing cost of our FCQ approach with the naive approach. Second experiments are to compare update cost of two approaches.

Algorithm1. SBCQ update

Input : a product movement pm =(p, A, B, type) denotes that product p moves from storage A to storage B.

Procedure Update(pm)

```
1.  SBCQ SBCQ(p,A) = getSBCQ(p, A);//obtain SBCQ which monitor product p in storage A
2.  SBCQ SBCQ(p,B) = getSBCQ(p, B);//obtain SBCQ which monitor product p in storage B
3.  Queue queue_B = getDataQueue(B); //queue_B is data queue for storage B
4.  If(queue_B ==null)
5.  | Queue_B = generateQueue(B);    //generate new data queue for storage B
6.  End If
7.  If(SBCQ(p,B)== null)
8.  | Case: type = total  // in case of total movement
9.  |    SBCQ(p,B) SBCQ(p,A);  SBCQ(p,B).setStorage(B);  de006Cete(SBCQ(p,A));
10. |    Queue queue_A=getDataQueue(A);
11. |    If(queue_A is not in use)  deleteQueue(queue_A);  EndIf
12. | Case : type = partial    // in case of partial movement
13. |   SBCQ SBCQ(p,B)  SBCQ(p,A) ;  SBCQ SBCQ(p,B).setStorage(B);
14. Else
15. | Case: type = total
16. |   delete(SBCQ(p,A));
17. |   If(queue_A is not in use)  deleteQueue(A);  End If
18. End If
```

Fig. 25.5 SBCQ update algorithm

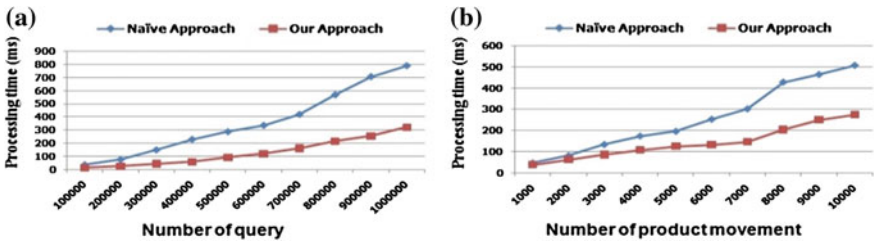


Fig. 25.6 Result of experiments. **a** Query processing evaluation. **b** Query update evaluation

To evaluate continuous query processing performance, we simulate a cold chain environment. In virtual environment, 50,000 sensor tags are deployed into 5,000 storages. 100,000 products are stored in the 5,000 storages.

The experimental results are shown in Fig. 25.6. In first experiments, our approach has shown better performance than the naïve approach. Figure 25.6a shows query processing cost. Figure 25.6b shows query update cost evaluation result. As evaluation result, our approach greatly improves the cost of continuous query processing and query update performance.

25.5 Conclusion

In this paper, we suggest a floating continuous query (FCQ) for monitoring product threshold of upcoming product movement. FCQ can be efficiently processed by the runtime decomposition reflecting product movement. Dynamic decomposition of FCQ removes the overhead of redefinition of transformed CQ and also provides efficiency of computing different aggregated values for different product storages.

Acknowledgment This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012014046).

References

1. Luis RG, Loredana L, Pilar B, Jose IR (2009) A review of wireless sensor technologies and applications in agriculture and food industry: state of the art and current trends. In Sensors, pp 4728–4750
2. Lakshimi VR, Vijayakumar S (2012) Wireless sensor network based alert system for cold chain management. In: International conference on modeling optimization and computing, pp 537–543
3. Jaekwan P, Bonghee H, Chaehoon B (2007) A continuous query index for processing queries on rfid data stream. In: 13th IEEE international conference on embedded and real-time computing systems and applications, pp 138–145

Chapter 26
A New Approach for Task Scheduling Optimization in Mobile Cloud Computing

Pham Phuoc Hung, Tuan-Anh Bui and Eui-Nam Huh

Abstract Mobile cloud computing is growing rapidly because its device (i.e., smart phone) is becoming one of the main processing devices for users nowadays. However, there are still some negative impacts that affect cloud access, especially when access to cloud becomes expensive but recent studies are not yet efficient in eliminating these. In this paper, we present an effective task scheduling by collaborating thick–thin clients and cloud to guarantee a better accessibility to cloud network and boost up the processing time in the mobile cloud platform while considering the network bandwidth and cost for cloud service usage. Intensive simulation proves that our method can improve the task scheduling efficiency and is better cost-effective than other works.

Keywords Task scheduling · Offloading · Parallel computing · Thin–thick client

26.1 Introduction

The last decade has witnessed a revolution in the way people access to the Internet thanks to the incredibly fast proliferation of smart devices and latest advancement in mobile communication network. Personal and business tasks are no longer attached to heavy, bulky desktop or laptop computers, etc. (known as thick clients), but can use much smaller and thinner devices (known as thin clients) like

P. P. Hung · E.-N. Huh (✉)
Department of Computer Engineering, Kyung Hee University, Yongin-si, South Korea
e-mail: johnhuh@khu.ac.kr

P. P. Hung
e-mail: hungpham@khu.ac.kr

T.-A. Bui
Louvain School of Engineering, Catholic University of Louvain, Louvain-la-Neuve, Belgium
e-mail: tuan.bui@student.uclouvain.be

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_26, © Springer Science+Business Media Dordrecht 2014

smartphones or tablets, or the Google glass [1], iWatch [2]. Switching work to mobile devices does not come without problems though. Small memory, weak CPU power and short battery life or unpredictable network connectivity [3] are some of the inherent obstacles that prevent mobile devices from satisfying the increasingly sophisticated applications demanded by users. Meanwhile, gaining more and more popularity in recent years, Cloud Computing (CC) offers a considerable complement to mobile terminals by expanding their power and enabling them to perform far beyond their original capabilities. This convergence results in a new paradigm named mobile cloud computing (MCC) [4, 5] allowing mobile devices to run computation-intensive applications, providing end users with rich computing experience.

Much as MCC can enrich mobile end users' experience, it however introduces some additional headaches related to cloud access. Researchers have been working hard to make MCC more accessible, especially when access path to cloud network becomes unavailable or too expensive. In [6], authors argue that smart objects (phones, tablets) can exploit computing resources from nearby nodes to achieve better cloud access. These efforts, unfortunately, are not sufficient to eliminate the above discussed problems of thin clients. Thick clients, including powerful smartphones (multi-cores, big RAM, LTE-enabled), on the other hand, usually come with better hardware and network connectivity. It is understandably suggested that thin clients be coupled with thick clients to achieve a desirable access to cloud, as stated in [7]. This paper aims at extending the work in [7] to utilize the thin-thick client collaboration and cloud in task scheduling to reduce the processing time of the system in the mobile cloud platform. We particularly consider both network contention and finance of cloud users. Simulation results prove that our model has better tradeoff between cost efficiency and workflow execution time than other existing work.

The remainder of this paper is structured as follows: Sect. 26.2 presents some related work which have partly solved the discussed problems. Section 26.3 gives the motivating scenario in which the role of thin-thick client collaboration becomes crucial. System architecture is presented in Sect. 26.4. Section 26.5 details the problem formulation. And Sect. 26.6 specifies the implementation, performance evaluation. The last section concludes our paper and suggests the future work.

26.2 Related Work

There have been numerous studies which attempt to solve task scheduling problems. In [8], authors propose a task scheduling approach for assigning processors to task graph templates prepared in advance. The limitation of this method is not to consider the network contention. Sinnen and Leonel [9] present an efficient task scheduling method based on network contention which, however, does not ponder on monetary cost paid by cloud customers (CCs) for using the resources.

In heterogeneous CC environment, despite numerous efforts, task scheduling remains one of the most challenging problems [10]. Authors in [11] introduce a cost-efficient approach to select the most proper system (private or public cloud) to execute the workflow according to deadline constraint as well as cost saving. Li et al. in [15] present a scheduling algorithm to schedule the application of large graph processing in consideration of both cost and schedule length. Research literature has not yet presented, though, any scheduling approach that concerns both execution time, network contention and the cost for using cloud services in combination of thin-thick clients and clouds. In this paper we try to solve the above shortage by introducing a method that can find an optimal schedule while keeping in mind the aforementioned constraints.

26.3 Motivating Scenario

The scenario illustrated in Fig. 26.1 exemplifies the utilization of thin-thick client collaboration in optimizing task scheduling: A Western girl is visiting a museum in South Korea and she enjoys seeing things that reflect Korean history and tries to understand more about it. Much of the information there, unfortunately, is written in Korean, which causes the girl unable to understand. As she is reading, she uses her voice to order her Google glass she is wearing to take pictures of the content in Korean language and uploads them to the cloud. While the Google glasses itself does not have a good Internet connection (because its bandwidth is too low), it is configured to get connected with the girl's powerful smart phone (in her pocket), which supports LTE connection. The pictures are automatically transferred to the phone which can then upload them to the Internet in a relatively short time, thanks to superb connection speed of LTE network. Once uploaded to the cloud, these pictures are quickly processed to extract text (using some optical character recognition technology) and translate it into the girl's preferred language (e.g. English). Translated content is then returned to the phone before transferring to the glass then the girl can read and understand it now.

26.4 System Architecture

The following section gives an insight of our system architecture to address issues discussed above: Our architecture has two layers, as illustrated in Fig. 26.2, including (1) Cloud Provider layer, which contains Virtual Machines (VMs), and (2) Cloud Customer layer, where thin clients and thick clients reside. In the second layer, there is a thick client m functioning as a centralized management node, also known as a broker, which (1) receives all computation requests of users (2) manages processor's profiles (processing capacity, network bandwidth) as well as computation costs together with results of data query returned from processors,

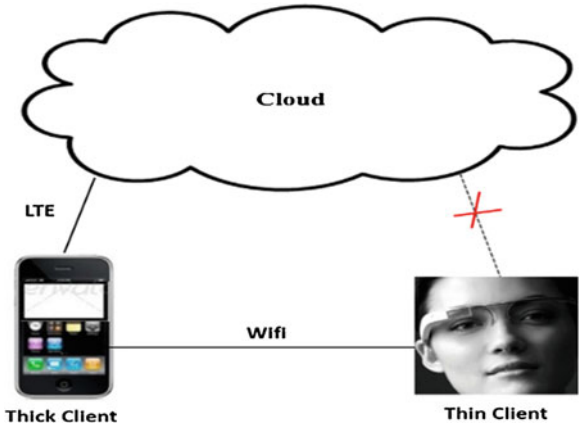


Fig. 26.1 Motivating scenario

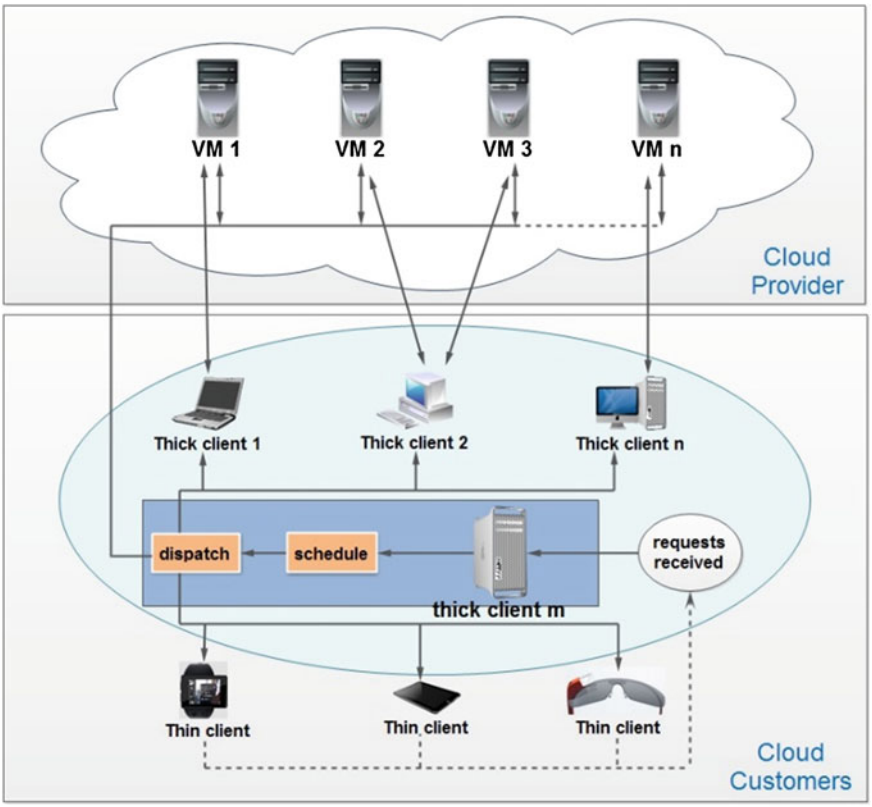


Fig. 26.2 System architecture

and (3) accordingly creates the most reasonable schedule for an input workflow. Especially, it sends data to clouds in a single connection but when VMs send data to cloud customer layer, the data will be divided into different parts with different sizes before being delivered to thick clients in multi connections according to a previous research [7]. Moreover, the system has to satisfy the following requirements:

- A STUN information and a communication library can be shared by these P2P and thick clients to thin clients or vice versa.
- Thick clients should store a copy of persistent data of the cloud, and should keep this loosely synchronized.

26.5 Problem Formulation

In this part, we first define the terms used and then formulate the problem. Eventually, we present a method to solve the above problems.

Definition 1 A task graph (e.g. shown in Fig. 26.3) is presented by a Directed Acyclic Graph DAG G = (V, E, *w, c*) where the set of vertices V = $\{v_1, v_2, \ldots, v_k\}$ represents the set of parallel subtasks, and the directed edge $e_{ij} = (v_i, v_j) \in E$ presents the communication between subtasks v_i and v_j, $w(v_i)$ associated with task $v_i \in V$ represents its computation time and $c(e_{ij})$ represents its communication time between task v_i and task v_j. We presume that a task v without any predecessors, $prec(v) = 0$, is called entry task v_{entry} and the task that does not have any successors, $succ(v) = 0$ is named end task v_{end}. Each task v_i has a different priority because some tasks should be serviced earlier than other ones [12]. It consists of workload w_i which delimits amount of work processed at the computing resources. Besides, it also contains the set of preceding subtasks $prec(i)$, the set of successive subtasks $succ(i)$ of task i, $t_s(v_i, P)$ denotes Start Time and $w(v_i, P)$ means the Execution Time of task $v_i \in V$ on processor P. Hence, the finish time of that task is given by $t_f(v_i, P) = t_s(v_i, P) + w(v_i, P)$. Suppose that the following conditions are satisfied:

Condition 1. A task cannot begin its execution until all its inputs have been gathered sufficiently.

Condition 2. The available time $avail(v_j, P)$ is the time that processor P completes the last assigned task and be ready to execute task v_j. Therefore,

$$avail(v_j, P) = \max_{e_{ij} \in E, v_i \in pred(v_j)} (t_f(e_{ij})) \tag{26.1}$$

Condition 3. Let $[A, B] \in [0, \infty]$ be an idle time interval on processor P, an interval in which no node is executed. A free task $v_i \in V$ can be scheduled on processor P within $[A, B]$ if

$$\max\{A, avail(v_i, P)\} + w(v_i, P) \leq B \tag{26.2}$$

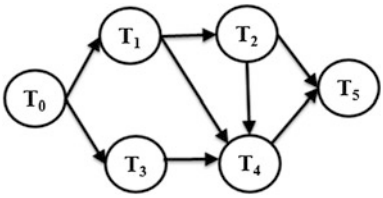


Fig. 26.3 A sample DAG

Definition 2 A processor graph TG = (N, D, H) is a graph that describes the topology of a network between vertices (processors) which are VMs, thick or thin clients. In this model, N is the finite set of vertices, a directed edge $d_{ij} \in D$ denotes a directed link from vertex n_i to vertex n_j, $n_i, n_j \in N$, H is a finite set of hyper edges representing a multi-directional link between vertices. In this graph, each processor i controls processing rate p_i and bandwidth bw_i to communicate with other processors.

Definition 3 Task scheduling S of a DAG G = (V, E, *w, c*) on a target system having network topology TG is to assign processor nodes of that system to task nodes of DAG in order to minimize total execution time. The input of task scheduling is a task graph and process graph. And the output is a schedule which is an assignment of a processor to each task node.

Proposed Method

In this part, we make some following assumptions for our proposed method. Given a task graph G = (V, E, *w, c*) and the processor graph with network topology TG = (N, D, H), our approach, extended from the Contention Aware Scheduling (CAS) algorithm [9], has two steps:

(a) *Determining the task priority to make the order of the tasks*

In this step, each task is set a priority based on the upward rank value of this task in the task graph. Here, a priority of a task v_i is estimated by the length of the critical path leaving the task. Recursively defined, the priority value *pr* of a task v_i is as follow:

$$pr(v_i) = \begin{cases} \overline{w}(v_i) + \max\limits_{v_j \in succ(v_i)} \left[\overline{c}(e_{ij}) + pr(v_j)\right] & v_i \neq endtask \\ \overline{w}(v_i) & v_i = endtask \end{cases} \tag{26.3}$$

where $\overline{w}(v_i)$ is the average computation time of task v_i and $\overline{c}(e_{ij})$ is the average data transfer time between task v_i and v_j, correspondingly:

$$\overline{w}, (v_i) = \frac{w_i}{\left(\sum\limits_{n_k \in N} p_k\right) \Big/ n}, \quad \bar{c}(e_{ij}) = \frac{c(e_{ij})}{\left[\sum\limits_{n_k \in N} bw_k\right] \Big/ n} \tag{26.4}$$

with n is the number of processors in the cloud environment.
Let $c(e_i^{jk})$ be the data transfer time from processor n_j to processor n_k to execute task i, then $c(e_i^{jk})$ is defined as following:

$$c(e_i^{jk}) = \left(di_i^j + \sum_{v \in prec(i), v \in exec(j)} ad_{vi} \right) * \left(\frac{1}{bw_j} + \frac{1}{bw_k} \right) \tag{26.5}$$

In the formula (26.5), di_i^j is the amount of input data stored at processor P_j and used for executing task v_i and ad_{vi} is amount of outgoing data from P_j to process task v_i.
Finally, we sort all tasks with a descending order of pr, which is the length of remained schedule.

(b) *Choosing the most appropriate processor to execute the selected task*

Once all preceding tasks of v_i are completed, the start time of a task presents the time when the last preceding task of v_i is completed. Thence, to determine that start time, the earliest idle interval [A, B] on processor P has to be searched to satisfy *condition* 3. As a result, the start time of task v_i is set as:

$$t_s(v_i, P) = \begin{cases} 0 & v_i = v_{entry_task} \\ \max(A, avail(v_i, P)) & v_i \neq v_{entry_task} \end{cases} \tag{26.6}$$

Thus, the Earliest Start Time (*EST*), and Earliest Finish Time (*EFT*) values of a task v_i executed on a processors P are computed as follow:

$$EST(v_i, P) = \max_{v_j \in prec(v_i), n_k \in N} \left(t_f\left(v_j, P_k\right) \right) + \max_{n_k \in N} \left(c\left(e_i^{kP}\right) \right) \tag{26.7}$$

$$EFT(v_i, P) = w(v_i, P) + EST(v_i, P) \tag{26.8}$$

Besides, the algorithm also considers the cost paid by cloud customers for using cloud resources that are used to execute the tasks. The cost $C(v_i, P_j)$ for task v_i executed at a VM P_j or at a thin client P_j as well as thick client P_j is defined by.

$$C(v_i, P_j) = \begin{cases} C_{process}(v_i, P_j) + C_{wait}(v_i, P_j) + (C_{transfer}(v_i, P_j) + C_{disconnect}(v_i, P_j)) & \text{for VMs} \\ C_{transfer}(v_i, P_j) + C_{disconnect}(v_i, P_j) & \text{for thin-thick clients} \end{cases} \tag{26.9}$$

In the formula (26.9), each cost is calculated as following:
Cost of processing is expressed as:

$$C_{process}(v_i, P_j) = c_1 * \frac{w_i}{p_j} \tag{26.10}$$

where c_1 is the processing cost per a time unit of workflow execution on processor P_j with processing rate p_j .

Let t_{min} be the finishing time of the task which is completed first out of the parallel tasks and there is no available task after this one, c_2 is the waiting cost per time unit, t_i is the finish time of the *ith* task then the cost of waiting time is as:

$$C_{wait}(v_i, P_j) = c_2 * (t_i - t_{\min}) \tag{26.11}$$

Suppose that the amount of money per transferring time unit of outgoing data unit from processor P_j is c_3, then the cost of transferring time is defined as follow:

$$C_{transfer}(v_i, P_j) = c_3 * \left(di_i^j + \sum_{v \in prec(i), v \in exec(j)} ad_{vi} \right) \Big/ bw_j \tag{26.12}$$

We assume that the distribution of disconnection event is a Poisson distribution with parameter μ_T. μ_T is determined by the stability of the network. The expected number of arrivals over an interval of length τ is $E[N_T] = \mu_T * \tau$. Let L be the random variable of the length of an offline event, μ_L is the expected length or mean of length and c_4 is disconnection cost per unit time. Therefore, the expected duration time of disconnection event which can affect to the processing time of task v_i is $\mu_T * \tau * \mu_L$. Hence, the cost of disconnection is as:

$$C_{disconnect}(v_i, P_j) = c_4 * (\mu_T * \tau * \mu_L) \tag{26.13}$$

From this cost, we can calculate the utility function which computes the tradeoff between the cost and *EFT* as formula (26.14).

$$Min \sum_{v_i \in E, P_k \in N} \left(\frac{cost(v_i, P)}{Max[cost(v_i, P_k)]} * \frac{EFT(v_i, P)}{Max[EFT(v_i, P_k)]} \right) \tag{26.14}$$

By considering the above utility function, we can determine the most appropriate processor P which processes task v_i is the one whose combination between *cost(v_i, P)* and *EFT(v_i, P)* should acquire the minimum value.

26.6 Implementation and Analysis

This section presents our experiments via numerical simulations to evaluate the efficiency of our approach and compare its performance with two others: CAS [9], which just takes account of network contention, and Classical Heuristic Scheduling (CHS), which merely concerns EFT. All the parameters are different task graphs G = (V, E, *w, c*) with the increase of the matrix size from 20 to 100 and heterogeneous processor graphs TG = (N, D, H) which is a combination between 15 VMs with the different configurations and 5 thick clients and 4 thin clients located at the local system of CCs for the above algorithms. We developed the simulations in Java with jdk-7u7-i586 and Netbeans-7.2. In the following figures, it is obvious to see that there are some differences between the simulated results.

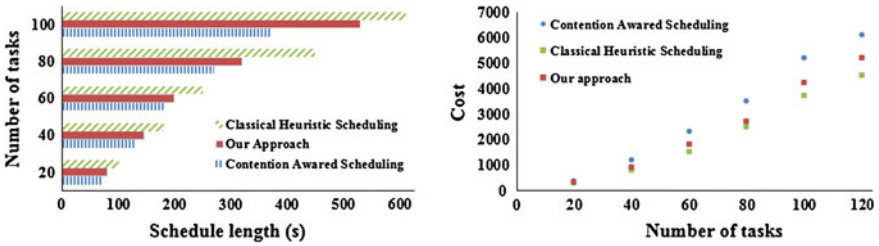


Fig. 26.4 Comparison our approach with others about schedule length and cost

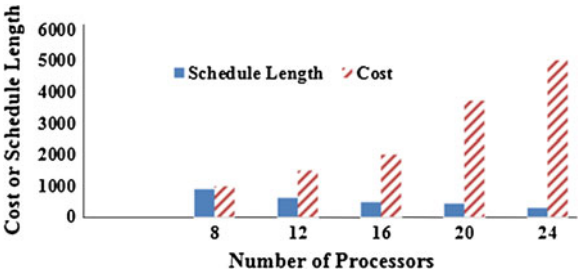


Fig. 26.5 Schedule length and cost

On the left of the Fig. 26.4 shows that CHS get the worst case on schedule length, CAS obtains the best result while our approach is still in the middle. Specifically, our method is 18 % better than CHS. However, on the right of the Fig. 26.4, according to the monetary cost paid by CCs, it can be seen that although CAS provides the best performance, it takes the biggest cost. In the meantime, CHS approach spends the lowest cost while our solution gets balanced between schedule length and cloud cost. Compared with CAS, our method can save nearly 22 % cost for CCs.

Besides, we measured the effect of the increasing number of processors on the cloud cost and the schedule length while fixing the quantity of tasks. The result is reflected in Fig. 26.5 which indicates that the more number of processors is used, the better performance our system gets but the cost CCs have to pay is higher as well.

26.7 Conclusion

This paper discusses an architecture based on the joint work of thin-thick clients and cloud in optimizing the task scheduling in mobile cloud platform. Our work especially tries to bring desired processing time while balancing the network contention and cloud service cost in order to bring a cost-effective solution, which appears to be better than other existing approaches when compared with. We will soon extend the proposed model to run in a various circumstances to achieve higher reliability and better performance.

Acknowledgments This research was supported by the MSIP (Ministry of Science, ICT&Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2013-H0301-13-4006) supervised by the NIPA (National IT Industry Promotion Agency). The corresponding author is Eui-Nam Huh.

References

1. Google glass, http://www.google.com/glass/
2. Apple iWatch. http://www.t3.com/news/apple-iwatch-rumours-features-release-date
3. Vallina-Rodriguez N, Crowcroft J (2012) Energy management techniques in modern mobile handsets. IEEE Commun Surv Tutorials 99:1–20
4. Huang D (2011) Mobile cloud computing. IEEE COMSOC Multimedia Commun Tech Committee (MMTC) E-Lett 6(10):27–31
5. Kumar K, Yung-Hsiang L (2010) Cloud computing for mobile users: can offloading computation save energy. IEEE Comput 43(4):51–56
6. Huerta-Canepa G, Lee D (2010) A virtual cloud computing provider for mobile devices. In: MCS'10, USA
7. Hung PP, Tuan-Anh B (2013) A solution of thin-thick client collaboration for data distribution and resource allocation in cloud computing. In: 2013 International conference on information networking (ICOIN), pp 238–243
8. Wolf J (2008) SODA: an optimizing scheduler for large-scale stream-based distributed computer systems. In: International conference on middleware, pp 306–325
9. Sinnen O, Leonel A (2005) Communication contention in task scheduling. IEEE Trans Parallel Distrib Syst 16(6)
10. Lee Y-C, Zomaya A (2008) A novel state transition method for metaheuristic-based scheduling in heterogeneous computing systems. IEEE Trans Parallel Distrib Syst 19(9):1215–1223
11. Van den Bossche R (2011) Cost-efficient scheduling heuristics for deadline constrained workloads on hybrid clouds. In: CloudCom, pp 320–327
12. Ghanbari S, Othman M (2012) A priority based job scheduling algorithm in cloud computing. ICASCE 50(2012):778–785

Chapter 27
Study on Encryption of Wireless Network-Based Mobility Energy-Enterprise Security Management

Young-Hyuk Kim, Il-Kown Lim, Jae-Pil Lee, Jae-Gwang Lee, Hyun Namgung and Jae-Kwang Lee

Abstract The objective of this study is to propose the encrypted mobility Energy-Enterprise Security Management (E-EMS) for a green race environment currently widely adopted througout the global. E-EMS is a system which can actively monitor and control energy consumption. Unlike wired network, there are certain weakness in terms of surveillance and security for wireless network so it is critical to have information exchange between modules, and administrators for the management of each security module. To overcome the aforementioned limitation of wireless network, this study proposes an integrated security control system equipped with Smart Rack and AES, SEED encryption protocol.

Keywords Energy monitoring · Security control · Mobility

Y.-H. Kim · I.-K. Lim · J.-P. Lee · J.-G. Lee · H. Namgung · J.-K. Lee (✉)
Department of Computer Engineering, Hannam University, Daejeon, South Korea
e-mail: jklee@netwk.hnu.kr

Y.-H. Kim
e-mail: yhkim@netwk.hnu.kr

I.-K. Lim
e-mail: iklim@netwk.hnu.kr

J.-P. Lee
e-mail: jplee@netwk.hnu.kr

J.-G. Lee
e-mail: leejk@netwk.hnu.kr

H. Namgung
e-mail: ghnam@netwk.hnu.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_27, © Springer Science+Business Media Dordrecht 2014

27.1 Introduction

It is currently known about 70 % energy consumed by PCs are rather from standby power. Among them, internet data center (IDC)s is responsible about half of IT related energy costs. Especially, the energy efficiencies of servers for government organizations and corporate companies are very low. About 33 % of consumed energy is lost during transfer and about 35 and 32 % are used for and cooling, respectively. Only about 20–30 % energy is actually used to operate CPUs [1, 2]. Therefore, it is critical to come up with a system which can actively monitor and control energy consumption to enhance the energy consumption efficiency.

The alternate is Energy-Enterprise Security Management (E-ESM). The previous E-ESM exists as various systems including firewall, Intrusion Detection System (IDS), Intrusion Prevention System (IPS) and Virtual Private Network (VPN) for carbon metering and/or power management monitoring.

The previous system is operated in separation of network to completely prevent the leakage of important information. Currently, the separation of network is made usually for wired network and do not emphasize on wireless network. Therefore, despite of an advent of smart devices and mobile convergence network, the previous system is very vulnerable to attacks such as Evil Twrin, MITM, Packet Sniffing and War walking/driving/carting and fails to take advantages of the network separation [3].

In this study, a SmartRack device and encryption application model are designed and a control system adopted with AES is developed, based on the comparison of AES and SEED, to develop an energy security control system which can be mobile at wireless network. SmartRack is composed of a temperature/humidity sensor, air conditioning device, PDU which is in control of power flow and embedded board for networks and encryption. For the management of each and individual security module, it is critical to exchange information between modules, and administrators. Therefore, it is required to have proper encryption to prevent the falsification of security logs, and control information [4]. In Sect. 27.2, the previous security control system and previous studies are reviewed, and the proposed system is designed in the Sect. 27.3. Evaluation of the encrypted performance is reviewed in the Sect. 27.4, and concluded in the Sect. 27.5.

27.2 Related Research

27.2.1 ESM

Security control system is a system which can integrate and control every security events including but not limited to firewall, IDS and Anti-Spam. Its main functions are to prevent the intrusion of networks and to collect security related information to provide an integrated security report. The existing security control system

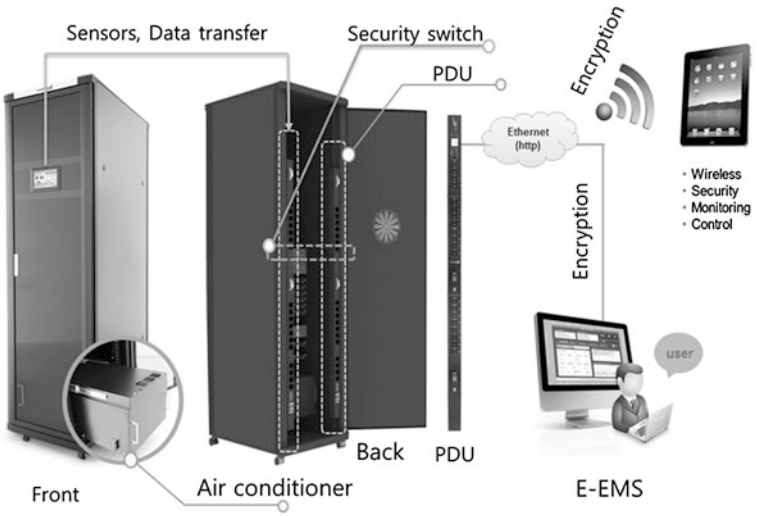


Fig. 27.1 E-EMS system

provides the collected security events to security administrator and with Risk Classification Methodology, Normalization/Rule base Event Collection, Abnormal Detection/Reaction and Integrated Policy Management, it allows to take an immediate action against cyber threats in real time. However, since it is only possible to collect and exhibit information detected by integrated security devices, detection information of the previous ESM is limited to only the installed security devices. Moreover, due to a large volume of security events, accumulating in real-time, it requires a large amount of time to check individual intrusion case even though a specialist monitors the whole situation. Also, even with one detection error, it is difficult to detect a relation of attacks made in a long term (Fig. 27.1).

ESM is an open Multi Vendor-based enterprise integrated security control platform and it integrates security devices such as F/W, IDS and Web F/W to collect security information from the security devices and show threat information in accordance with predetermined patterns [5].

27.2.2 Security Information and Event Management (SIEM)

SIEM (Fig. 27.2) is one of the representative integrated security control technology and it monitors threats against internal networks as well as internal behaviors. HP taking over ArcSight, IBM taking over Q1Labs and McAfee taking over NitroSecurity are representative companies in this field. SIEM receives statistics information, events, threats and danger data from security devices, servers and network devices including firewall, IDS, IPS and Anti-Virus, and analyzes

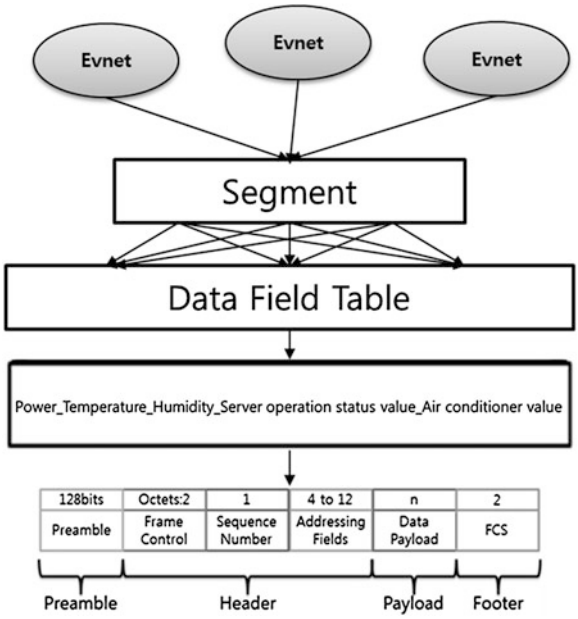


Fig. 27.2 Protocol architecture

correlations between the received information to check whether the detected situation is threat and take an immediate action upon analysis as well as control of logs. The vital module for such product is monitoring and the corresponding monitoring can be made either at desk tops as well as mobile devices of administrators.

27.3 Mobility E-EMS Design

Mobility E-EMS is an integrated control system which can monitor energy consumption of IT facilities which are operated under the scheme of Green-IDC, Green-Building, and Green-City, and convert and control the carbon emission with three dimensional analysis of data collected from temperature/humidity sensors and reduce the carbon emission through the intelligent control of energy facilities based on the analysis results.

Components of the data sensor includes but not limited to servers, DGMS, IT systems, communication networks, power, temperature/humidity sensor system, security and operation system, and control system.

Figure 27.1 shows components of the E-EMS proposed in this study.

Computing resources of IDC are composed as rack. Since this system is in charge of monitoring but controlling, wireless communication should be made with mobile devices in a safe manner. Therefore, the comparison of AES and

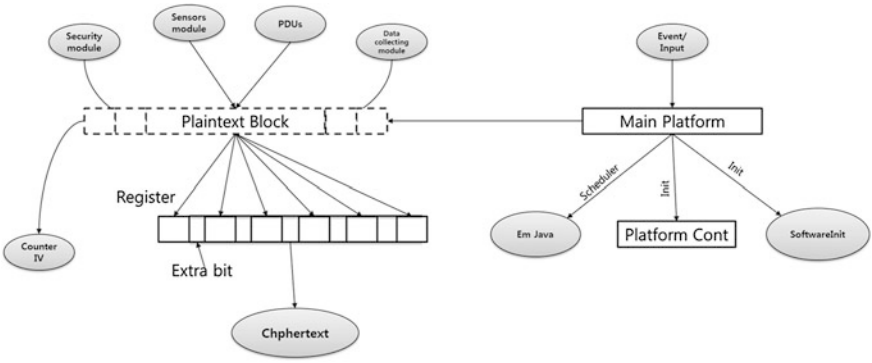


Fig. 27.3 Whole process flow

SEED was made for this study and since AES was found to be more suitable, AES encryption is applied for communication in this study (Fig. 27.3).

E-EMS utilizes a sort of the separator separation with '.'. Information on temperature, humidity and current status of an air conditioning system, sent by sensors installed on the Smart Rack monitored by PDU are composed to be one data frame with an separator of '.'.

Protocol starts as recording of the first figure of individual value of data payload which follows a sequence number included the header at PDU.

In other words, if smart rack internal temperature exceeds certain temperature of which an air conditioning system requires the sophisticated control and the temperature meets the situation value requires the attention of an administrator, a server saves relevant information at DB and sends a message to an administrator immediately. However, if an air conditioning operates normally, a server discards the received data without saving.

For management of modules and security modules, it is critical to exchange information between modules, and administrators. In order to prevent any of falsification of security logs and control logs in this stage, encryption must be made and it is made at PDU of SmartRack. Since a PDU module is embedded with a module for processing the integration with a control system, the control system and a PDU-DB Server saves data on the exchange of energy information, raw data and control signals. For the exchange of data, encryption protocol is required. For this study, the comparison of AES and SEED was made and it was found that AES-256 algorithm has faster encryption speed upon Key Schedule and less no. of round steps so AES was decided to be used for this study.

Figure 27.4a is an snapshot of the operation with SEED encryption at Eclipse, and 27.4b is an snapshot of the operation of a coded encrypted by AES-256. It was set that encryption would be maintained till the output of a control system (Fig. 27.5). In order to achieve that, Embedded Java, which is usually used for a product requiring a large volume of data, is employed for the implementation

Fig. 27.4 Data Encryption conversion code of SEED-256 (**a**) and AES-256 (**b**) on screen

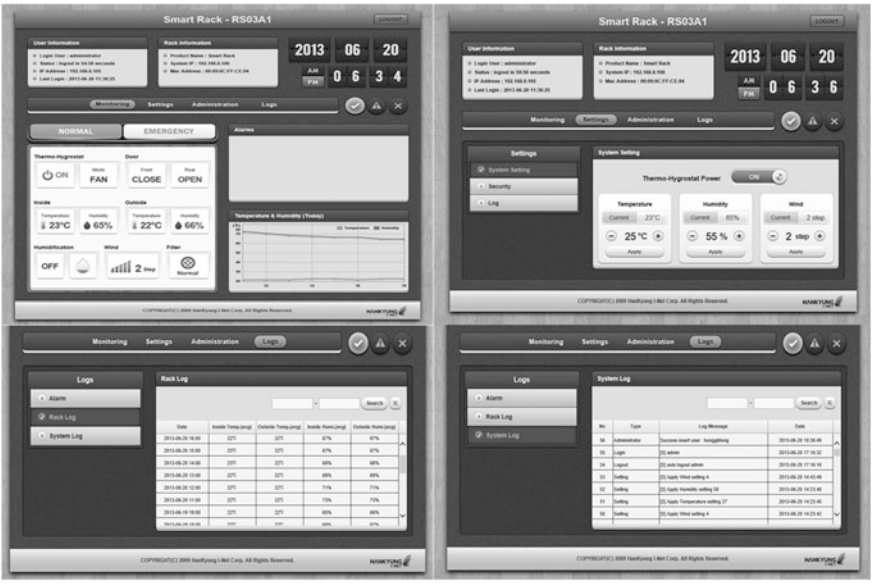


Fig. 27.5 Implemented control system

environment of the proposed system. Since Embedded Java is designed exclusively for JVM, it has outstanding portability and expandability.

Figure 27.5 is a final web monitoring screen which would be shown to an administrator. On this screen, internal conditions could be monitored, controlled and managed. The consistency of user information, internal status, and current time and date could be checked, and status of an air conditioning system, closing/opening of a door on the front/back of Smart Rack, internal temperature/humidity and fan speed on the back part and an air conditioning system can be monitored as well. Moreover, it is possible to set up the encryption between DB server and PDU server, the encryption period limitation and the encryption system.

27.4 Conclusion

This study aims for providing the communication safe and secured in wireless network to provide mobility to E-EMS. Therefore, the comparison of AES and SEED was made and it was found that AES-256 algorithm has faster encryption speed upon Key Schedule and less no. of round steps so AES is used for the proposed system.

EMS could be considered as general security control system but E-EMS is a security control system exclusive for power measurement and control. Consistent communication is required between modules, security modules, and administrators, and communications should be made in wireless network in a safe manner to provide the mobility to administrator. Our future goal is to apply AES only to wired communication and to apply the encryption protocol, designed exclusively for mobile devices, to wireless communication. With HIGh security and light weigHT (HIGHT), 64 bit block encryption for mobile devices, we aim to provide low-power and high speed but reliable communication in wireless network.

References

1. NIA (2013) Cool summer ICT for overcoming power crisis. Green IT Focus, vol 6
2. Kwon WO, Kim SW (2008) PC power is leaking. IITA Technol Weekly 1344:1–14
3. Lee DS (2013) Security management service framework in mobile convergence network. Ph.D Thesis, Sungkyunkwan University
4. Ha OH (2009) A study on conversion security control system for industrial security. J Inf Secur 9(4):1–6
5. Kim MJ, Kim KN (2011) A study of mining esm based on data-mining. J Inf Secur 11(6):3–8
6. No BK, Kim DW, Kim KS, Kim IK (2013) Intelligent cyber security technology and issue. PM Issue Report 2013 1(3)
7. Lee JP, Kim YH, Lim IK, Lee JG, Lee JK (2013) A study on the encryption protocol for security vulnerability and improvement in the carbon metering systems. The Korean Institute of Communications and Information Sciences conference, pp 114–115

Chapter 28
Secure and Efficient Deduplication over Encrypted Data with Dynamic Updates in Cloud Storage

Dongyoung Koo, Junbeom Hur and Hyunsoo Yoon

Abstract Cloud service providers adopt a deduplication technique to minimize resource utility costs. However, it is one of the most challenging issues to manage the outsourced data in a storage-efficient way when users encrypt data for preserving privacy and frequently update it. When the data is updated, file-level deduplication makes entire copy of updated file although there are small modifications. Block-level deduplication solves this problem, but it requires metadata larger than the outsourced blocks. To address this problem, we propose a hybrid deduplication scheme that minimizes storage overhead. Our scheme performs file-level deduplication along with isolation of only updated blocks with augmented metadata. The analysis results show that our scheme minimizes storage costs while guaranteeing secure update with efficient verification.

Keywords Cloud storage · Deduplication over encrypted data · File-level deduplication · Block-level deduplication · Storage overhead

D. Koo · H. Yoon
KAIST, Daejeon, Republic of Korea
e-mail: dykoo@nslab.kaist.ac.kr

H. Yoon
e-mail: hyoon@nslab.kaist.ac.kr

J. Hur (✉)
Chung-Ang University, Seoul, Republic of Korea
e-mail: jbhur@cau.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_28, © Springer Science+Business Media Dordrecht 2014

28.1 Introduction

For efficient utilization of cloud storage costs, cloud service providers adopt data deduplication technique to save storage by keeping a single copy of duplicate data and providing users a link to the stored data [1].

In deduplication approaches, privacy issues also have gained a significant attention [2, 3]. As a promising technique, several encryption schemes have been extensively studied in file or block level for data privacy as well as reduction in storage costs [4–6], without consideration of dynamic updates of outsourced data in the cloud storage.

In the cloud service, authorized users may modify, insert, or delete the outsourced data on occasion. However, previous approaches are yet to provide deduplication over encrypted data efficiently in the environment. File-level deduplication makes entire copy of updated file although there are small modifications. Block-level deduplication solves this problem, but it requires metadata larger than the outsourced blocks.

Related work. Xu et al. [5] proposed file-level deduplication over encrypted data by exploiting Merkle tree, while Bellare et al. [6] introduced Message-Locked encryption as a generalization of convergent encryption [7]. Although several extended studies provided data privacy via encryption, these approaches cannot apply file-level and block-level deduplications at the same time.

Our contribution. We present a hybrid deduplication scheme over encrypted data where file-level and block-level deduplications are applied simultaneously. The proposed scheme achieves significant reduction in storage space. Once the outsourced data are updated, only the updated blocks are uploaded and maintained. In addition, data owners can retrieve and decrypt the data with their own keys.

28.2 Scheme Construction

28.2.1 Cryptographic Background

Bilinear map. Given two multiplicative cyclic groups $\mathbb{G}$ and $\mathbb{G}_T$ of prime order p with a generator g of the group $\mathbb{G}$, bilinear map $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$ is an efficiently computable algorithm satisfying the following properties: (1) Bilinearity: for any $s, t \in \mathbb{Z}_p$ and any $u, v \in \mathbb{G}$, $e(u^s, v^t) = e(u, v)^{st}$, (2) Non-degeneracy: $e(g, g) \neq 1$ for the generator g of $\mathbb{G}$.

Bilinear Diffie-Hellman (BDH) assumption. BDH problem is to compute $e(g, g)^{abc} \in \mathbb{G}_T$ provided $g, g^a, g^b, g^c \in \mathbb{G}$ for $a, b, c \in \mathbb{Z}_p^*$ for bilinear map e. If the probability, of which polynomial-time algorithm solves BDH problem on $\langle p, \mathbb{G}, \mathbb{G}_T, e \rangle$, is negligible then $\langle p, \mathbb{G}, \mathbb{G}_T, e \rangle$ is said to satisfy the BDH assumption.

28.2.2 System Description

A cloud storage system consists of the following three entities:

1. Cloud service provider (CSP): A CSP is an entity providing storage, retrieval, and verification services for outsourced data. We assume that it is honest-but-curious. Thus, the CSP follows a prescribed protocol but tries to learn useful information during the process for illicit profit. For storage services, it maintains *entry table* for deduplication and *ownership list* for dynamic updates at the cloud storage.
2. Data uploader: This is a data owner trying to outsource data to the cloud storage. For privacy concerns, the initial uploader encrypts the data before outsourcing. When a subsequent uploader attempts to upload the same data, proof-of-ownership process would be performed instead of actual uploading of the data. Henceforth, both initial and subsequent uploaders are referred to as data owners.
3. Updater: This is a data owner trying to modify the outsourced data. The updater can manage its own version of updated data independently. All data owners can verify integrity of their outsourced data regardless of updates.

28.2.3 Deduplication Over Encrypted Data with Dynamic Updates

System setup. A trust initializer runs *Setup* algorithm to generate system public parameters: two multiplicative cyclic groups $\mathbb{G}, \mathbb{G}_T$ of prime order p with a generator g of the group $\mathbb{G}$, a bilinear map $e : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$ defined over $\langle p, \mathbb{G}, \mathbb{G}_T, g \rangle$, and a cryptographic hash function $h : \{0,1\}^* \rightarrow \mathbb{Z}_p^*$. Then, $\langle p, \mathbb{G}, \mathbb{G}_T, g, h \rangle$ is publicized as *param*.

Initial upload. When an initial uploader u_i tries to outsource a file F, its hash value $h(F)$ is sent to the CSP. If $h(F)$ was not in the entry table, the CSP requests u_i to upload actual data. On receipt of upload request, u_i generates a key pair (ek_F, dk_F) by running *EKGen* algorithm. u_i chooses a random value $k \in \mathbb{Z}_p^*$ and computes $ek_F = h(e(g,\ g)^k)$ and $dk_F = g^{k \cdot \Sigma^{-1}}$ where $\Sigma = \sum_{k=1}^{n} B_k \cdot dk_F$ can be used by subsequent uploaders to derive ek_F using Σ.

Then, u_i encrypts F by running *Encrypt* algorithm for preserving privacy. Each block of F is encrypted via secure symmetric encryption *SE* under the encryption key ek_F such as $C_F = (C_{B_1}, \ldots, C_{B_n})$, where $C_{B_k} = SE_{sk_F}(B_k)$ for $1 \leq k \leq n$. u_i constructs a Merkle tree over the ciphertext C_F and maintains only the value of root node in the tree. This value is used for later integrity verification.

Prior to outsourcing, u_i determines its own private/public key pair ($sk_{i,F}$, $pk_{i,F}$) through *UKGen* algorithm as $sk_{i,F} \xleftarrow{\$} \mathbb{Z}_p^*$ and $pk_{i,F} = g^{sk_{i,F}^{-1} \cdot \Sigma}$ to decrypt the cipheretxt

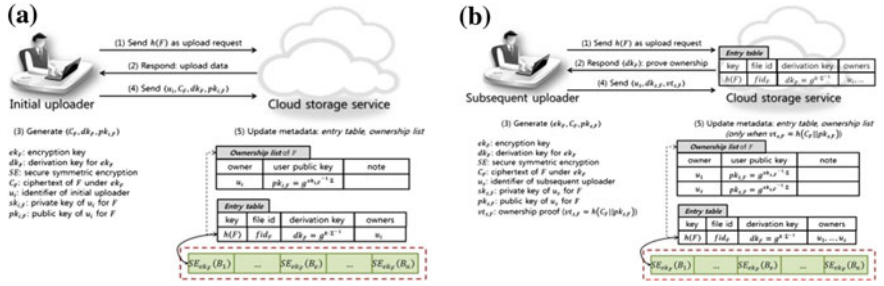


Fig. 28.1 Data outsourcing. **a** Initial upload of $\boldsymbol{F}$. **b** Subsequent upload of $\boldsymbol{F}$

later. u_i finishes data upload by sending $\langle u_i, C_F, dk_F, pk_{i,F} \rangle$ to the CSP. When the CSP receives upload responses from u_i, it inserts new fields to the entry table and ownership list, respectively, as depicted in Fig. 28.1a.

Subsequent upload. When the CSP receives $h(F)$ as an upload request of file F already existing in the entry table, the CSP transmits $\langle dk_F \rangle$ to the subsequent uploader u_s for an ownership proof of F. Then u_s runs *EKDrv* algorithm to derive the encryption key $ek_F = h\left(e(g, dk_F)^{\Sigma}\right) = h\left(e(g,g)^{k}\right)$. u_s computes a cipheretxt C_F which is the same as outsourced data via *Encrypt* algorithm, and constructs Merkle tree over C_F and keeps the value of root node in the tree.

u_s runs *UKGen* algorithm for $\langle sk_{s,F}, pk_{s,F} \rangle$ and keeps $sk_{s,F}$ secret for later retrieval. After that, to prove the ownership of F, u_s computes a verification term $vt_{s,F} = h\left(C_F || pk_{s,F}\right)$. and sends $\langle u_s, pk_{s,F}, vt_{s,F} \rangle$. On receiving it, the CSP runs *PoW* algorithm. If $h\left(C_F || pk_{s,F}\right) \neq vt_{s,F}$, *PoW* returns *False* and the CSP ignores u_s' request. Otherwise, *PoW* returns *True* and the CSP inserts u_s and $pk_{s,F}$ into entry table and ownership list, respectively, as in Fig. 28.1b.

Update of outsourced data. All data owners can update their data. When the updater u_u modifies a block B_v in the outsourced file F, u_u generates $\left(ek_{B'_v}, dk_{B'_v}\right)$ pair by choosing a random $k' \in \mathbb{Z}_p^*$ via *EKGen*. The public key $pk_{u,B'_v} = g^{sk_{u,F} \cdot B'_v}$ is computed with the private key $sk_{u,F}$ unchanged. Then, $C_{B'_v}$ can be obtained as a ciphertext of the modified blk B'_v through *Encrypt* algorithm. u_u finishes update request by sending $\left\langle u_u, C_{B'_v}, dk_{B'_v}, pk_{u,B'_v} \right\rangle$ with $m\left(B'_v\right)$ implying a modification of B_v to B'_v. Insertion can be performed in the same way as modification except that u_u sends $i(B_{v1})$ for an insertion of the data after v-th block in the outsourced file. u_u can delete blocks just by sending $\langle u_u \rangle$ with $d(B_v)$ to remove the v-th block from the file. On every update, u_u reconstructs and keeps Merkle tree over updated cipehretxt.

On receipt of update request from u_u, the CSP updates corresponding fields in the entry table and ownership list according to a type of the update request.

Audit for outsourced data. Data owner can send audit request to the CSP for checking the integrity of the outsourced data. The CSP constructs Merkle tree over

the stored data and delivers the value of root node in the tree to the requester. Since the CSP maintains entry table and ownership list, the latest version of Merkle tree can be correctly constructed regardless of updates. By running *Audit* algorithm, the owner can verify the integrity of the outsourced data. If the owner receives the same value as one kept secret by the owner, the *Audit* returns *True* implying integrity is guaranteed. Otherwise, *False* is returned, which means the integrity is violated.

Retrieval of outsourced data. Data owners u_o trying to retrieve the outsourced data F receives $\langle C_F, dk_F, pk_{o,F} \rangle$ from the CSP. Then, u_o derives ek_F by computing $ek_F = h\left(e\left(ek_F, pk_{o,F}\right)^{sk_{o,F}}\right) = h\left(e(g,g)^k\right)$ and runs *Decrypt* to obtain the plaintext F by applying theymmetric decryption algorithm under the key ek_F.

28.3 Evaluations

Since a hybrid approach for encrypted data is first introduced, we compare the proposed scheme with block-level deduplication and file-level deduplication using Message-Locked Encryption [6] in terms of efficiency and security.

28.3.1 Performance Analysis

Key management overheads. User-level key management in the proposed scheme allows each data owner to determine the key by its own criteria. However, data owners are forced to store the key determined by an initial uploader in the other approaches, which may cause inefficient key management overhead.

Storage overheads. Storage costs for a single file are summarized in Table 28.1. The file-level deduplication requires entire copy for updated data, which makes the storage size linear to the number of updates. The block-level deduplication requires metadata for each block for decryption, which is the largest storage overhead. On the other hand, the proposed scheme has minimal storage overhead in the cloud storage, since only updated blocks are augmented in the cloud storage.

Computation overheads. All approaches except the proposed scheme cannot provide user-level key management. When we assume file-level deduplication encrypts n blocks to generate a ciphertext, it requires re-computation of a new ciphertext in file level for each data update. Block-level deduplication needs to verify each block to guarantee the integrity since there is no way to summarize the whole file. The proposed scheme requires minimal computations for *Audit* and *Update* as in Table 28.2. Although more costs are incurred in *EKGen*, it is performed only once during the whole lifetime of the data in our scheme.

Table 28.1 Storage costs comparison

	File-level deduplication	Block-level deduplication	Proposed scheme
Cloud storage	$nS_c + 2S_p + un(S_c + 2S_p)$	$n(S_c + 2S_p) + u(S_c + 2S_p)$	$nS_c + 2S_p + u(S_c + 2S_p)$
Data owner	S_p	S_p	S_p

S_c size of ciphertext for one block, S_p size of element in $\mathbb{Z}_p^*$, n number of blocks in file, and u number of updates performed on the file

Table 28.2 Computation costs comparison

	EKGen	*UKGen*	*En/Decrypt*	*Update*	*Audit*
File-level deduplication	C_h	–	nC_{SE}	unC_{SE}	C_h
Block-level deduplication	nC_h	–	nC_{SE}	uC_{SE}	nC_h
Proposed scheme	$C_M + C_E + C_e$	C_E	nC_{SE}	uC_{SE}	C_h

C_h hash operation, C_M multiplication operation, C_E exponentiation operation, C_e pairing operation, and C_{SE} operation of secure symmetric encryption/decryption

28.3.2 Security Analysis

In order for polynomial-time adversary to learn information about F, it should recover the encryption key ek_F. However, only data owner of F can compute Σ and it might be removed from local storage after upload. The only way for the adversary to recover ek_F is to derive dk_F and $\{pk_{o,F}\}_{o\in\mathcal{O}}$, where $\mathcal{O}$ is a set of data owners of F. As long as data owners keep their secret key safely, any adversary cannot derive ek_F from $dk_F = g^{k\cdot\Sigma^{-1}}$ and $\left\{pk_{o,F} = g^{sk_{o,F}^{-1}\cdot\Sigma}\right\}_{o\in\mathcal{O}}$ without knowledge of Σ due to discrete logarithm assumption and BDH assumption.

With regard to secure verification, due to the properties of cryptographic hash, the CSP cannot generate a valid proof via the correct Merkle tree unless the CSP stores the corresponding file in its entirety. Therefore, the proposed scheme is secure in terms of data privacy and integrity.

28.4 Conclusion

In this paper, a hybrid deduplication over encrypted data is presented, which enables block-level deduplication as well as file-level deduplication with reduction in storage while providing privacy. In addition, user-level key management allows data owners to determine their own keys instead of maintaining a set of independent keys for each block/file outsourced by others. The proposed scheme is best suited for cloud storage services where data owners frequently update the outsourced data.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0016584 and No. 2013R1A2A2A01005559).

References

1. Meyer DT, Bolosky WJ (2012) A study of practical deduplication. ACM Trans Storage 7(4):14
2. 2013 Data Breach Investigations Report, http://www.verizonenterprise.com/resources/reports/rp_data-breach-investigations-report-2013_en_xg.pdf
3. Malicious insider attack to rise, http://news.bbc.co.uk/2/hi/7875904.stm
4. Srinivasaan K, Bisson T, Goodson G, Voruganti K (2012) iDedup: latency-aware, inline data deduplication for primary storage. In: FAST, 14 pp, IACR
5. Xu, J., Chang, E.C., Zhou, J.: Weak Leakage-Resilient Client-side Deduplication of Encrypted Data in Cloud Storage, In: ASIACCS, pp.195-206, ACM, (2013)
6. Bellare M, Keelveedhi S, Ristenpart T (2013) Message-locked encryption and secure deduplication, In: EUROCRYPT, 29 pp, Athens
7. Douceur J, Adya A, Bolosky W, Simon P, Theimer M (2002) Reclaiming space from duplicate files in a serverless distributed file systems. In: ICDCS, pp 617–624
8. Dropbox, https://www.dropbox.com/
9. Mozy, http://www.mozy.com/

Chapter 29
Tool Selection: A Cloud-Based Approach

Chirpreet Singh, Qun Shao, Yuqian Lu, Xun Xu and Xinfeng Ye

Abstract Cloud manufacturing is an emerging manufacturing model. It provides organizations with the ability to virtualize their manufacturing resources and offer them as scalable cloud-based services. Tool selection is a crucial function for machining operations. It determines whether a consumer's manufacturing task can be realized in a machining system. This paper proposes a system for carrying out tool selection for a typical cloud manufacturing system. It uses an ontology to store information about manufacturing resources and automatically generates queries for searching suitable tools in the ontology according to a customer's specification. The searching of the tools is carried out by a reasoner on the ontology. From the tools that match the user's requirements, the system selects the most appropriate tool according to the user's preferences.

Keywords Ontology · Cloud manufacturing

C. Singh · X. Ye (✉)
Department of Computer Science, The University of Auckland, Auckland, New Zealand
e-mail: x.ye@auckland.ac.nz

C. Singh
e-mail: csin042@aucklanduni.ac.nz

Q. Shao · Y. Lu · X. Xu
Department of Mechanical Engineering, The University of Auckland,
Auckland, New Zealand
e-mail: qsha771@aucklanduni.ac.nz

Y. Lu
e-mail: ylu633@aucklanduni.ac.nz

X. Xu
e-mail: x.xu@auckland.ac.nz

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_29, © Springer Science+Business Media Dordrecht 2014

29.1 Introduction

Cloud manufacturing is an emerging manufacturing model. It gives organizations the ability to virtualize their manufacturing resources as cloud-based services for consumption over the Internet [19]. An integral part of the cloud manufacturing paradigm is resource virtualization. It is concerned with digitally representing the manufacturing resources and services, and making them available as cloud-based services.

Machining is one of the essential manufacturing processes and tool selection is an important stage of planning a machining operation. It is therefore imperative that a cloud manufacturing system have a robust tool selection mechanism. This paper presents a system developed to virtualize resources using an ontology that facilitates cutting tool selection for a machining task. The system allows users to input their manufacturing criteria as a standardized manufacturing operation file. It then analyses the inputs and automatically generates description logic queries for the purpose of selecting the most appropriate tool from the ontology.

This paper is organized as follows. Section 29.2 describes the related works. The proposed tool selection system is discussed in Sect. 29.3. Conclusions are given in Sect. 29.4.

29.2 Related Work

The right choice, rational use and effective management of cutting tools have a great impact on the efficiency and quality of a production process. To a large extent, cutting tools dictate the geometry and quality of the machined surface, the tolerance of product dimensions and the machining cost of a part. On selecting a cutting tool, several factors need to be taken into account, e.g. tool life, the hand of tools, the geometry of tools and number of edges of a tool. In a traditional process planning stage, the selection of a cutting tool is mainly based on the experience of a domain expert, which can be error-prone and time-consuming. The large number of tools in the cloud environment call for an intelligent cutting tool selection system.

As an important part of process planning, cutting tool selection issues have always been a topic of research. Zhang and Alting [20] proposed a list of more than 100 Computer-Aided Process Planning (CAPP) systems. Integrating with Standard for Exchange of Product data (STEP) [10] feature recognition and approaches of artificial intelligence, Amaitik and Kilic [1] designed a CAPP system named ST-FeatCAPP, which is capable of generating a process plan and the corresponding STEP-NC [11] file in XML format. Similar works can also be found in [7, 8, 16]. On the other hand, a number of companies have also designed

various commercial software tools. Coroguide[1] developed by Sandvick Coromant helps clients to retrieve cutting tool information rapidly.

However, there are some problems with the existing systems. Current CAPP systems usually do not handle semantic data. For example, STEP AP224 XML documents are used to describe feature information. The cause of the problem is that an XML Schema mainly describes the syntax, instead of the semantics, of a data structure and the relationships between data. Furthermore, the proprietary software tools developed by companies often have a limited scope of operational use.

In order to address the aforementioned limitations, the use of ontologies in the manufacturing industry has been studied by numerous researchers. Kim et al. [12] presented an ontology-based framework for assembly information sharing. Lee et al. [15] developed an ontology-based architecture to facilitate collaboration between the different stakeholders in a manufacturing value chain. Their aim was to share concepts and design information between heterogeneous systems. Kumara [9] used an ontology-based methodology to support product family design. It stated that, when design and manufacturing information is required to be shared and reused, ontologies provide a particularly advantageous platform to do so. Pandit and Zhu [17] used an ontology-based decision framework to support product design and manufacturing. The works reported in [2, 4, 21] map EXPRESS-based data models to (Web Ontology Language) ontologies. Distinct to these existing works, the system in this paper automatically generates description logic queries for searching the ontology according to users' requirements. Thus, users of a typical cloud manufacturing platform can easily use the system even if they do not know the description logic, and be guaranteed that a logically *correct* set of matching tools will be provided.

29.3 A Tool Matching Scheme

The system consists of three components: (a) a user interface, (b) a tool selection module, and (c) a manufacturing resource information module. The manufacturing resource information module is implemented as an OWL-DL ontology. It stores the definitions of the concepts used in the cloud manufacturing system and the information about each individual cutting tool available in a specific cloud manufacturing system. Additionally, it is responsible for executing the queries that retrieve the tools that meet the users' requirements. The user interface is a browser-accessible Web application. Users search for available cutting tools by submitting a production specification through the interface. The tool selection module analyses the user's specification of a production task to retrieve, from the ontology, the appropriate tools that are needed for carrying out the task. Then, the

[1] http://www.coroguide.com/

system generates queries for retrieving the suitable cutting tools from the resource information module. If there are multiple tools available for carrying out an operation, the tool selection module evaluates the available tools to decide the one that best suits the user's specific production needs.

29.3.1 Resource Virtualization

Resource virtualization is about representing the information of a physical resource. For example, the representation of a drilling tool might include its dimensions, the type of coolant to be used, the cost of the tool, etc. A manufacturing cloud often consists of numerous partners, each of which might have their own methods of describing their virtualized resources. For this reason, interoperability is a key issue that must be addressed in cloud manufacturing.

The EXPRESS data modelling language [9] is a part of the STEP standard that is widely used by engineering industries to digitally describe products. However, the EXPRESS standard has two major shortcomings [4].

- It lacks formal *rigorousness*. That is, EXPRESS is not based on a mathematical theory. Thus, it is impossible to infer implicit information from the information that is explicitly specified.
- It has limited reusability and interoperability. Applications from the engineering domains are almost exclusively the only types that are able to understand and manipulate EXPRESS-based data models. As cloud manufacturing spans over various domains, e.g. design, machining, assembly, logistic, etc., EXPRESS has inherent interoperability problems, in particular over the Internet.

Due to these limitations, EXPRESS-based data models are not suitable for representing virtualized resources in cloud manufacturing. It has been argued that data can become smarter through the use of higher semantics from technologies such as ontologies [12]. Ontologies allow for concepts of a particular domain, and the relationships between those concepts, to be formally described. A key benefit of using ontology is that it allows for a common understanding of terms to occur between different entities. That is, ontologies facilitate interoperability. Thus, ontology of manufacturing resources has been used in this paper for resource virtualisation.

The ontology language used in this paper is OWL-DL which is a W3C standard [3]. OWL-DL is based on description logic [5]. This means that mathematical inference can be carried out on the ontology to discover implicit information.

An OWL ontology is a collection of axioms, each of which can provide an explicit assertion on three main ontological constructs: individuals, properties, and classes. OWL individuals represent the objects of a particular type of *things*. For example, an individual could be a specific instance of a twist drill. OWL properties are relationships between two things, namely individuals (e.g. tool name) or data values (e.g. diametre). OWL classes are sets of individuals and are described using

logical descriptions and restrictions. Classes are typically organized into hierarchies using taxonomical operators such as super-classes and sub-classes.

Accurately modeling a hierarchy of tools and their relationships with other tools is a complex task due to the level of interconnections required. Thus, creation of a manufacturing resource ontology is often tedious and error-prone. As a large number of manufacturing resources have already been described in various STEP-based standards, instead of creating the ontology manually from scratch, the ontology is instead procured by converting these existing STEP-based resources to an ontology. The set of specifications chosen for this paper is the ISO 14649 (STEP-NC) set of standards [13]. The main advantages of using the models defined within ISO 14649 are: (a) it is an established, continually developed, worldwide effort to standardize machining operations, (b) it contains specifications of all the appropriate tools, and the tool hierarchy, and (c) it contains an exhaustive set of machining features for computer numerical control (CNC) machine tools. Thus, the models in ISO 14649 will ensure that the resource virtualization will be generic enough to account for all standardized CNC tool operations.

29.3.2 Automatic Query Generation

To ensure that an object is manufactured to the specifications set out by the designer, it has become a common industry practice to computerize manufacturing tasks. In the STEP-NC standard, a manufacturing task is defined as a set of computerized *working steps*. Each *working step* is a digitized machining operation that can be performed autonomously by a CNC machine. To operate a CNC machine, a valid STEP-NC instance file that contains the set of *working steps* is required. A STEP-NC instance file specifies the machining operation of each *working step*, the tool required for carrying out the operations, and parameter values for the specific machining operations. Our system requires the user to input a STEP-NC instance file that describes the manufacturing operations required to realize the user's product. To execute the operations, the cutting tools that are capable of carrying out the operations need to be found in the manufacturing resource ontology.

In order to retrieve the set of tools that match the tools specified in the user's STEP-NC instance file, the system needs to (a) parse the STEP-NC instance file to retrieve the information about each of the tools for carrying out the machine operations, and (b) construct description logic (DL) queries according to the retrieved information. The DL queries are used to search for suitable tools in the ontology. A DL query is effectively a set of OWL class expressions. An OWL class expression is used to represent sets of instances (OWL individuals) by explicitly specifying the conditions on those instances' properties [3]. That is, if an instance of a particular class satisfies a set of property conditions, it is considered to be an instance of the respective class expressions that match it. For each machining operation described within the user's input file, our system

automatically constructs a DL query for finding the tool that is capable of carrying out the machining operation.

For different tools, the parameters might need to be matched in different ways. Some parameters, e.g., `number_of_teeth` of a tapered end mill, require precise matching. That is, the number of teeth must be exactly the same as the value specified in the user's input. Some parameters, e.g., the *diameter* of a tapered end mill, may only require a range-matching. For instance, if the user's request specifies that a flat bottom hole of diameter 20 mm be machined, any end mill whose diameter is less than 20 mm would be able to carry out the operation. These matching features are stored in the ontology for various tool types. They are retrieved from the ontology to help construction of the queries for tool selection. The procedure used to generate the tool matching DL queries is as follows:

- For each *working step*, determine which type of tool is required to complete the machining operation, and append its tool class to the tool selection DL query. For example, append `tapered_endmill` for a tapered end mill.
- For each tool type, retrieve the matching property of the parameters of the tool from the ontology, e.g. *exact* or *greater than*.
- For each parameter of the tool, retrieve the value of the parameter from the STEP-NC file, and append the class expression to the DL query according to the value and the matching property of the parameter. For example, append `number_of_teeth value ''4.0''^^decimal` for the parameter `number_of_teeth` according to the extracted value 4.0 of the type decimal and matching feature of *exact*.

An example of a constructed query in Manchester OWL syntax is shown below. The query is used to find all tapered end mills that have (a) a diameter *no greater than* 20 mm, and (b) four teeth.

```
(tapered_endmill) and (diameter some decimal [<= ''20.0''^^decimal])
and (number_of_teeth value "4.0''^^decimal)
```

29.3.3 Selecting the Most Suitable Tool

The result of a query might return multiple tools that satisfy the need of a machining job. Our system evaluates the tools to find out the one that best meets the requirements. The Multiple Criteria Decision Making and Simple Additive Weighting techniques [6] are commonly used in comparing different entities. In these techniques, a user assigns a weight to each of the factors being considered when ranking the entities. The factors are normally the properties of a tool, e.g. cost of the tool, the precision, etc. For example, if a user regards the precision is twice as important as the cost of the tool, the user can assign 2 as the weight for the

precision and 1 as the weight for the cost. For each attribute, the similarity between a user's expected value and the value offered by the tool is calculated using the formulas below. $V^{\max}$ and $V^{\min}$ are the maximum and the minimum value of the property that the user is willing to accept. V is the value of the property offered by the tool. If the maximum and the minimum value are equal, the similarity value SV is set to 1.

$$SV = \frac{V^{\max} - V}{V^{\max} - V^{\min}} \quad \text{and} \quad SV = \frac{V - V^{\min}}{V^{\max} - V^{\min}}$$

The left formula is used for negative property, i.e. the higher the value, the lower the quality, e.g. the price of the tool. The right formula is used for positive attribute, i.e. the higher the value, the higher the quality, e.g. precision, etc.

The ranking score assigned to a tool is calculated according to the formula below. In the formula, P denotes the set of properties being considered when ranking a tool, W_i is the weight assigned to property i, and, SV_i is the similarity value of property i. The tool with the highest score will be given to the user, i.e.

$$Score = \sum_{i \in P} (SV_i \times W_i)$$

29.3.4 Implementation

The system was implemented using Java, specifically JavaFX. After converting the tools information stored in an ISO 14649 (STEP-NC) file to an OWL-DL ontology, the ontology is stored in a clear text ASCII file. New instances of the tools are inserted into the ontology directly through the user interface of the system. Our system uses the OWL-API to manipulate the resource ontology [18]. When the system initializes, the resource ontology is loaded into the system.

A software reasoner is an important component of an OWL-based ontology system. A reasoner can infer implicit knowledge in ontology. In order to obtain *correct* query results, queries on OWL ontologies are executed by a reasoner. This is because the query results often exist as the implicit knowledge in the ontology. For example, although instances of twist drill and tapered end mill classes are not directly specified as milling cutters, when querying for milling cutters, all instances of twist drill and tapered end mill should be shown. This is because twist drill and tapered end mill are subtypes of the milling cutter class. The reasoner used in the system is the Pellet reasoner [18].

29.4 Conclusions

The cutting tools ontology has been successfully constructed based on which tool-matching can be carried out using a reasoner. The system extracts the specifications of the required tools from the descriptions of the machining operations and automatically generates description logic queries for carrying out search on the ontology. The system shows that cloud manufacturing resources can be effectively modeled and searched using ontology.

References

1. Amaitik SM, Kiliç SE (2007) An intelligent process planning system for prismatic parts using STEP features. Int J Adv Manuf Technol 31(9–10):978–993
2. Barbau R et al (2012) Ontostep: enriching product model data using ontologies. Comput Aided Des 44(6):575–590
3. Bechhofer S, et al (2004) Owl web ontology language reference. W3C Recommend 10:2006-01
4. Beetz J et al (2009) Ifcowl: a case of transforming express schemas into ontologies. Artif Intell Eng Des Anal Manuf 23(01):89–101
5. Horrocks I, Patel-Schneider PF (2003). Reducing owl entailment to description logic satisfiability. In: The semantic Web-ISWC 2003. Springer, Berlin, pp 17–29
6. Hwang C-L, Yoon K (1981) Multiple criteria decision making. Lecture notes in economics and mathematical systems. Springer, Berlin
7. Garrido Campos J, Marín Martín R (2010) Modelling and implementing circular sawblade stone cutting processes in STEP-NC. Robot Comput Integr Manuf 26(6):602–609
8. Gizaw M, Abdul Rani AMB, Yusof Y (2013) Turn-mill process plan and intelligence machining operations selection on STEP. Asian J Sci Res 6(2):346–352
9. ISO 10303-11 (1994) Industrial automation systems and integration-product data representation and exchange part 11
10. ISO 10303 AP203 (1994) Industrial automation systems and integration – product data representation and exchange—part 1: overview and fundamental principles, Switzerland
11. ISO 14649 (2003) Industrial automation systems and integration – physical device control—data model for computerized numerical controllers—part 1: overview and fundamental principles, Geneva, Switzerland
12. Kim K-Y, Manley DG, Yang H (2006) Ontology-based assembly design and information sharing for collaborative product development. Comput Aided Des 38(12):1233–1250
13. Kramer T, Xu X (2009) Step in a nutshell. In: Xu X, Nee AYC (eds) Advanced design and manufacturing based on STEP. Springer, Berlin, pp 1–22
14. Kumara SR (2006) A methodology for product family ontology development using formal concept analysis and web ontology language. J Comp Inf Sci Eng 6:103
15. Lee J, Chae H, Kim C-H, Kim K (2009) Design of product ontology architecture for collaborative enterprises. Expert Syst Appl 36(2):2300–2309
16. Ouyang HB, Shen B (2012) STEP-NC oriented process planning optimization based on hybrid genetic algorithm. Comput Integr Manuf Syst CIMS 18(1):66–75
17. Pandit A, Zhu Y (2007) An ontology-based approach to support decision-making for the design of ETO products. Autom Constr 16(6):759–770
18. The OWL API, http://owlapi.sourceforge.net/, accessed on 10/11/2013

19. Xu X (2012) From cloud computing to cloud manufacturing. Robot Comput Integr Manuf 28(1):75–86
20. Zhang H, Alting L (1994) Computerized manufacturing process planning systems. Chapman & Hall, New York, London
21. Zhao W, Liu J (2008) OWL/SWRL representation methodology for express-driven product information model. Comput Ind 59(6):580–589

Chapter 30
The Self-induction System for Fuel-Efficient Driving Using OBD II

WonAng Jang, JinGwan Park, JuWan Kweon, ChangYong Lee and DoHoon Lee

Abstract Driving habit affects emission of greenhouse gas and fuel consumption. In order to manage those issues, we can utilize the internal information of vehicle using on-board diagnostic II(OBD II). In this paper, we propose an evaluation model to monitor driving habit or behavior and fuel consumption using the OBD II. Proposed model shows this works well under abnormal driving condition as well as that of normal driving. We also propose a self-induction system for safe driving and monitoring a vehicle on android mobile platform.

Keywords Monitor system · Vehicle · OBD II · Driving behavior

30.1 Introduction

Recently, there have been many attempts for reducing greenhouse gas emissions from transportation and vehicle's fuel consumption. Therefore we paid attention to the use of the internal information of vehicle's for those issues. However it is

W. Jang · J. Park · J. Kweon · C. Lee · D. Lee (✉)
Department of Computer Science and Engineering, Pusan National University, Pusan, South Korea
e-mail: dohoon@pusan.ac.kr
URL: http://visbic.pusan.ac.kr

W. Jang
e-mail: wonang@pusan.ac.kr

J. Park
e-mail: pjk8906@naver.com

J. Kweon
e-mail: peinkmls@naver.com

C. Lee
e-mail: puga21@naver.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_30, © Springer Science+Business Media Dordrecht 2014

difficult for users to collect vehicle's internal information. In order to overcome the problem there have been attempts to use on-board diagnostic II (OBD II), which is an equipment for sensoring a vehicle's self-diagnosis [1–3].

The previous diagnostic equipment was very big, expensive, and difficult to handle. Recently, however, it is getting smaller and cheaper, and can be installed to its connector in the vehicle. There have been studies to get electronic control unit(ECU) data by OBD II and communicate between OBD II connector and CDMA module. For the comunication between the connector and the other devices, wireless method such as BLUETOOTH and WI-FI are used extensively [4].

There have been applications using Bluetooth, such as Torque Pro, OBD car doctor pro, OBDdroid scan pro, caroo pro, and etc. All of them provide similar information such as fuel efficiency, gathered by many sensors. Especially Torque pro provides driver's route and caroo pro provides additional function such as vehicle's blackbox data [5, 6]. Fueoogle [2] by Nam Pham provides most efficient route to sensor vehiclular fuel consumption, and Eco-driving by Kanok [3] evaluates change in driving habit.

Real time sensoring data from OBD-II are difficult to understand, and we have to convert the data into accessible information for driver to use fuel efficiency model. The fuel efficiency model [7] by Lightner works well under normal driving environment but gives inadequate value under abnormal case such as sudden acceleration and breaking.

In this paper, we propose a modified fuel efficiency model which works well under abnormal driving condition, such as a burst speeding and sudden breaking, as well as normal driving environment. In addition, we develop a self-induction system to guide safe driving behavior and fuel efficient driving habit.

30.2 System Architecture

Figure 30.1 shows flow diagram of proposed system. The System transmits PID to OBD-II module and OBD-II module returns ECU raw data to driver system. SS-index calculates the combination of fuel efficiency value, rapid acceleration value and sudden stop value Safe and Saving index (SS-index) signifies driver's safe driving habit and fuel efficiency. The system transmits SS-index to Smart-phone using Bluetooth, and driver will change depending on visualization of SS-index in Smart-phone. The system gathers 3-axis acceleration sensor values using bluetooth and sends an S.O.S to emergency phone number along with GPS coordinate, when system detects accident.

OBD-II data, which the proposed system use, is Table 30.1 and decides fuel efficiency value, rapid acceleration value, and sudden stop value to calculate SS-index.

Moment fuel efficiency value calculate Mile per Gallon (MPG) using all rows of Table 30.1. The constant value are given in Table 30.2. A mile per gallon [7], MPG, is defined by

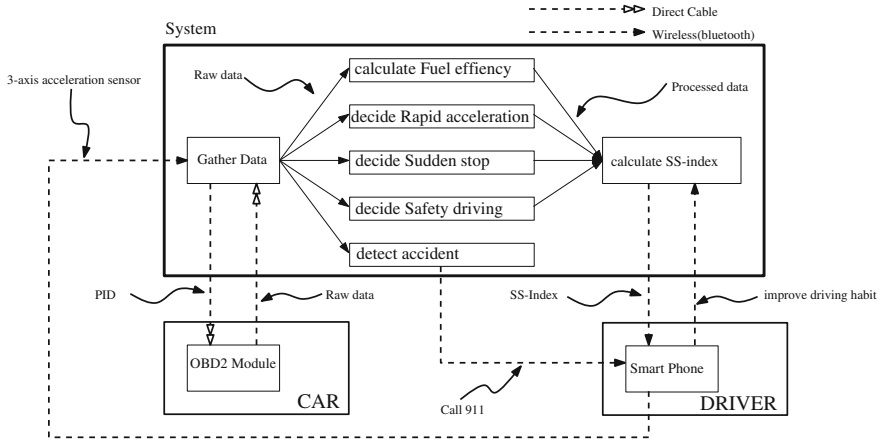


Fig. 30.1 System flow diagram

Table 30.1 Parameters used in OBD-II

Num.	Symbol	Unit
1	Engine RPM	RPM
2	Vehicle speed (VSS)	Km/h
3	Intake manifold absolute pressure (MAP)	Kpa
4	Intake air temperature (IAT)	‘C

$$\text{MPG} = \frac{AFR * PPG * GPP * VSS * MTK}{SPH * MAF} \tag{30.1}$$

All constants of Table 30.2 except AFR are values for unit transformation. AFR is an ideal ratio of air and fuel. According to Bruce Lightner [7], confidence of Eq. 30.1 is high because ECU regularly maintains AFR. VSS can be offered from OBD-II described in Table 30.1, but MAF might not be offered if there aren’t sensors in car.

A mass air flow, which is proposed by Bruce Lightner using RPM, MAP, IAT, VSS and ideal gas equation [8] MAF is defined by $\text{MAF} = \frac{28.97 * VE * ED * \left(\frac{imap}{12}\right)}{8.314}$, where VE is engine volume efficiency and ED is engine displacement Constant, 28.97 means molecular weight in the air and constant, 8.314 is fuel constant. VE and ED depends on installed engine in each car.

An engine volume efficiency, VE, is defined by $\text{VE} = \frac{\pi}{4} * len(bore)^2 * len(Stoke) * NoC$,where *len(bore)* and *len(stroke)* denote the length of bore and stroke, and *Noc* denotes the number of cylinder. ED of KIA NEW MORNING, which is experimental vehicle, is about 2 l and VE calculates the value of 0.85.

Table 30.2 Constants for calculating MPG

Num.	Symbol	Value	Description
1	AFR	14.7	Grams of air to 1 g of gasoline—ideal air/fuel ratio
2	PPG	6.17	Pounds per gallon—density of gasoline
3	GPP	4.54	Grams per pound—conversion
4	MTK	0.621371	Miles per hour/kilometers per hour—conversion
5	SPH	3,600	Seconds per hour—conversion

IMAP denotes RPM and MAP values, and is defined by $\text{IMAP} = \frac{RPM * MAP}{AT}$.

RPM and MAP can directly be offered from OBD-II but Absolute Temperature (AT) is a conversion value converted from absolute temperature. An absolute temperature, AT, is defined by $\text{AT} = \frac{273.16}{IAT}$.

Lastly MPG of Eq. 30.1 converts Kilometer per liter (KPL) by KPL = 0.425144 * MPG.

A sudden acceleration, SA, is defined by

$$\text{SA} = \begin{cases} +1, & \left(RPM_{pre} < \delta_{SA} \cap RPM_{now} \geq \delta_{SA}\right) \\ 0, & else \end{cases}$$

where δ_{SA} is threshold value for evaluating sudden acceleration and we use the value 3,000 according to the experimental estimation. SA increases by one if prior collected RPM is under δ_{SA} and current collected RPM is above δ_{SA}.

A sudden break, SB, is defined by

$$\text{SB} = \begin{cases} +1, & \left(\frac{VSS_{pre} - VSS_{now}}{DT} > \delta_{SB}\right) \\ 0, & else \end{cases}$$

where δ_{SB} is threshold value for evaluating sudden break and we use the value 0.01 according to the experimental estimation. SB increases by one if the ratio of the decrement of speed and time-varying degree is above δ_{SB}. Namely, if speed decreases by more than 1 km per hour duing time-varying degree of δ_{SB}, SB increases by one. DT is in millisecond.

A SS-index, SS, is defined by SS = BP-PP. A base point, BP, is defined by $\text{BP} = \frac{AFE}{OFE} * 100$, where AFE is average fuel efficiency and OFE is official fuel efficiency. AFE calculates the mean of collected moment fuel efficiency once every driving. OFE is offered by automaker. A penalty point, PP, is defined by $\text{PP} = \frac{SA * rate_{SA} + SB * rate_{SB}}{RT}$, where rate_{SA} and $rate_{SB}$ (<20) are the degrees of effect in the fuel consumption. The point of rate increases by 1 whenever the accumulated time (SA and SB) reaches 5 min.

A detect accident, DA, is defined by

$$\text{DA} = \begin{cases} detect, & \sqrt{(x^2 + y^2 + z^2) - (9.81^2)} > \delta_{DA} \\ 0, & else \end{cases}$$

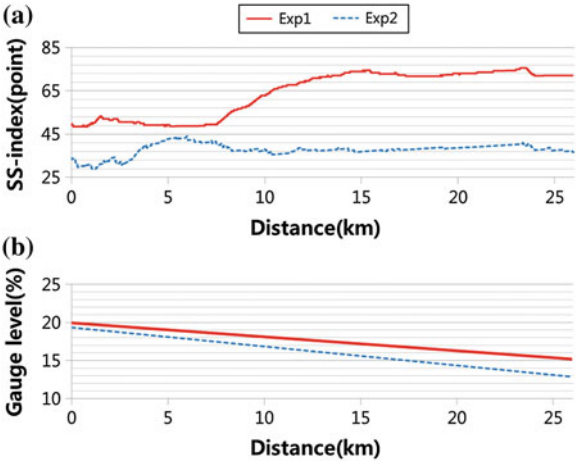


Fig. 30.2 The experimental result of driving environment of 26 km. **a** SS-index. **b** Fuel gauge level

where δ_{DA} is threshold value for evaluating detect accident and we use the value 10 according to the experimental estimation. DA is detected when acceleration value, that is difference between 3D-acceleration sensor value and acceleration of gravity value, exceeds δ_{DA}.

30.3 Discussion

For the implementation and experimental test, we use FUN:DI which is Bluetooth-OBDII diagnostic equipment and Android 4.0v. We examined under two types of driving environments to evaluate effect of reduction in fuel efficiency with SS-index and Fig. 30.2 shows the result. The data comes from the driving distance 26 km in identical road. A experiment car has an engine displacement of 1,600 and a gas tank of 60 l. In Fig. 30.2, Exp1 shows the variance of SS-index (a) and fuel consumption (b) according to the driving distance when we intentionally induce high SS-index without SA and SB. Also Exp2 shows the variance of SS-index (a) and fuel consumption (b) when we intentionally induce low SS-index with many SA and SB.

From the first experiment, we examined SA twice, none for SB, and average SS-index was 63 points. Figure 30.3 shows the result of SA and SB for each experiment. The next one, we also examined that SA was 17, and SB was 2. The average SS-index was 39 points. Comparing the fuel consumption according to the driving distance, we noticed that the Exp1's fuel gauge level was 2.7 % higher than Exp2's fuel gauge level for the 26 km driving as shown in Fig. 30.2b. The result tells us to get benefit of fuel consumption and the amount of 1.62 l of fuel used in vehicle.

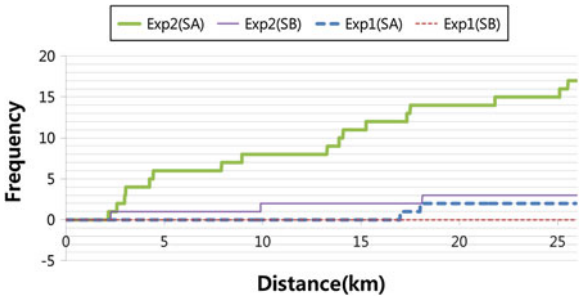


Fig. 30.3 The frequency of SA and SB for 26 km

30.4 Conclusion

Driving habit affects fuel consumption and greenhouse gas generation. Most of the eco-driving studies have been concentrating on giving advice such as taking class and training not a system. So we analyzed how to affect the driving behavior, especially sudden acceleration and breaking, with fuel efficiency, and proposed a self-induction system to evaluate driving behavior by SS scoring model.

The scoring model improved the mileage system that has a limitation in estimating fuel efficiency under sudden behaviors in driving. We expect that the system will guide better driving habits to give the evaluation score, and also reduce the fuel consumption.

References

1. Ganti RK, Pham ND, Ahmadi H, Nangia S, Abdelzaher TF (2010) GreenGPS: a participatory sensing fuel-efficient maps application. In: MobiSys '10 proceedings of the 8th international conference on mobile systems, applications, and services. ACM Press, New York, pp 151–164
2. Pham ND, Ganti RK, Nangia S, Pongthawornkamol T, Ahmed S, Abdelzaher TF, Heo J, Khan M, Ahmadi H (2009) Fueoogle: a participatory sensing fuel-efficient maps application. Technical report, University of Illinois Urbana Champaign, Illinois
3. Boriboonsomsin K, Vu A, Barth M (2010) Eco-driving: pilot evaluation of driving behavior changes among U.S. drivers. In: Faculty Research, University of California Transportation Center, UC Berkeley, California
4. Jang WA (2013) Wireless OBD2 API to get vehicle inner information. Master thesis, Pusan National University
5. Torque Pro, https://play.google.com/store/apps/details?id=org.prowl.torque&hl=en
6. Caroo Pro, https://play.google.com/store/apps/details?id=com.pokevian.prime&hl=en
7. Bruce D (2005) AVR-based fuel consumption gauge. In: AVR 2004 design contest, Issue 183, Oct 2005
8. MAP and MAF-Based Air/Fuel Flow Calculator, http://www.lightner.net/obd2guru/IMAP_AFcalc.html

Chapter 31
Intuitive Hand Motion Recognition and NUI/NUX Framework

Gwanghyung Lee, Dongil Shin and Dongkyoo Shin

Abstract Up to now, typical motion recognition methods have used markers. The recognition methods were to receive coordinate input values of each marker as relative data and to store each coordinate value into the database. Methods using markers could store and utilize accurate values in the database but as ubiquitous era comes, there was no time enough to handle the preparation process for recognition. To compensate for this problem, we don't use markers and implement real time motion recognition framework using Kinect camera. Especially the framework of hand mouse and fingers recognition framework is implemented. Also, we implemented for anyone to handle NUI/NUX framework easily and intuitively.

31.1 Introduction

With diversification of image display device, the demand for new interfaces is increasing. To communicate with computers and humans, the study of simple and intuitive control method such as the current commercial mouse and keyboard, touch screen, space touch is growing up.

G. Lee · D. Shin · D. Shin (✉)
Department of Computer Engineering, Sejong University, 98 Gunja-Dong, Gwangjin-Gu, Seoul 143747, South Korea
e-mail: shindk@sejong.ac.kr

G. Lee
e-mail: khlkhlll@gce.sejong.ac.kr

D. Shin
e-mail: dshin@sejong.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_31, © Springer Science+Business Media Dordrecht 2014

Recently, software using KINECT has been developed for the elderly who has discomfort legs [1], and the system has been developed to select the object in the virtual world using a camera [2].

Voice of the Android phone and i-phone, tablet PC and so on is reliable enough to be used in. But humans delivered to the other party their will using voices as well as nonverbal means and voice alone is difficult to communicate in a lot of noise in the area around. Studies have been reported that non-verbal means such as hand gestures, body gestures and facial expressions convey more information [3]. In this paper inspired by this point, we propose and implement multi-modal NUI/NUX framework controlling voice, body motion, and facial expression simultaneously.

31.2 Related Works

Because this NUI is used for the natural motion interface without using mice, keyboards, pens and, etc., "Natural" terminology is used. While as GUI, the interface shown with eyes provides visible and clear environment, NUI is hard to clearly make a confirmation with eyes and if NUI/NUX interface is used, additional study is needed [4]. Also, the design should have scalability of gesture set and recognition rate in mind for performance of the development software system [5].

31.2.1 Methods Using Markers

Previously, in order to study natural motion recognition, markers are used for high quality. In a film, Avatar, Human motions could be recognized perfectly by attaching about 40 markers only to characters' face and about 6–8 markers only to each human joint. This is called motion capture and after attaching sensors (markers) to various parts of the body, what virtual characters behave as the same motion through position values of sensors is the core of this technology. Then four years later, now, as this motion capture technology is developed, not only face recognition but also facial wrinkles are expressed by attaching about 100 markers to face in a game "Beyond: two souls" as shown in Fig. 31.1.

In this paper, to implement NUI/NUX framework, cumbersome markers are not used and methods recognizing human motion are studied using only camera.

31.2.2 Multi-modal User Interface

Multi-modal user interfaces combine the different interfaces such as motion, face, voice and etc. recognition interfaces and can complement each other. The representative multi-modal interface, "Put-That-There" controlled systems through

Fig. 31.1 Game "Beyond: two souls", developed using markers and released in 2013

pointing and voice as user inputs [6]. Also, North American and European car manufacturers consider user-friendly UI/UX importantly as development of next-generation technology based external environment recognition technology [7].

31.3 Design and Implement of NUI/NUX Framework and Experiment Results

NUI/NUX framework designed in this paper is multi-modal interface and its kind is motion, voice, facial expression recognition and so on. Also, we utilize NUI/NUX framework designed in various device using XML. Figure 31.2 is the overall look of the implemented framework. After that, NUI/NUX framework is explained sequentially by each modal of its architecture.

31.3.1 Hand Mouse Gesture Recognition

Previously if windows were moved in the monitor, a mouse was used but what in this paper is proposed is to implement intuitive hand gesture NUI/NUX framework, hand mouse. As several examples of an intuitive NUI, when users want to take pictures, they take body gesture as if they press the shutter of the camera. When users want to control the size of the window, they hold the window frame and take action to control it.

As shown in Fig. 31.3, specifically we implemented "Click" event to stretch right hand above 0.5 m forward the body, "Drag" event to keep this status and "Double-Click" event to occur "Click" again in 0.1 s. Also, as a mouse icon clicks an object, actions can be taken to get feedback. In the right side of Fig. 31.4, the upper number of the big size numbers is shown in the distance from Kinect camera to the user's chest and the lower number of them is shown in the distance

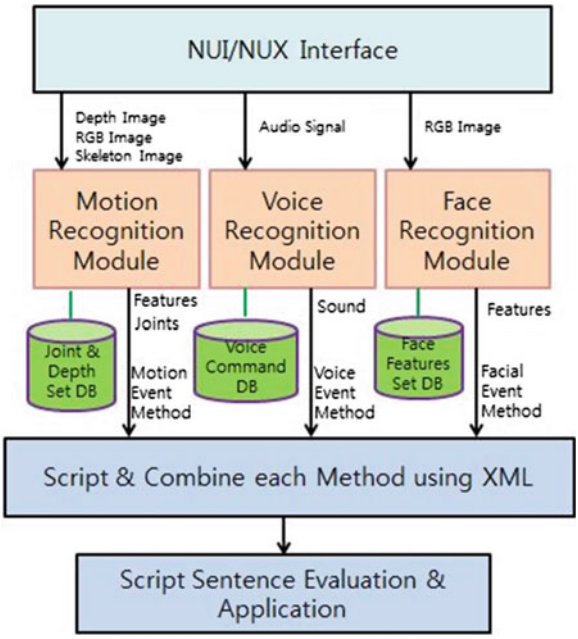


Fig. 31.2 NUI/NUX framework system architecture

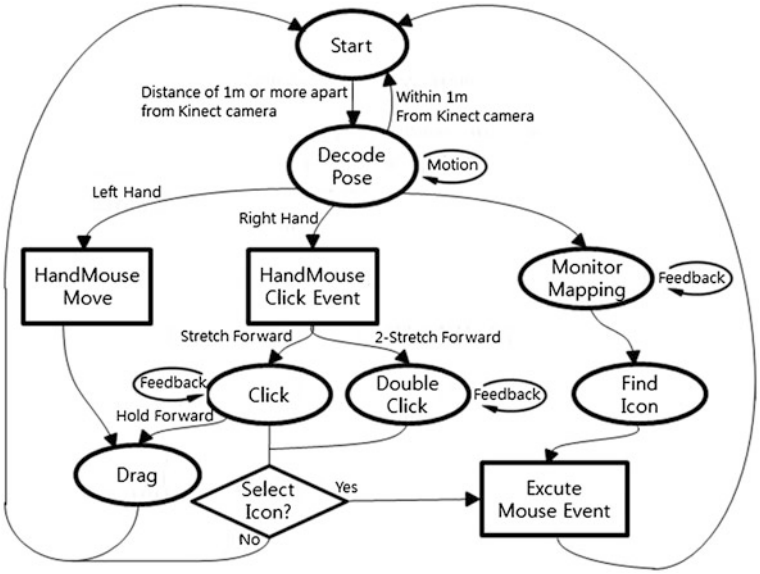


Fig. 31.3 Flow chart of the hand mouse framework implemented at second

Fig. 31.4 Framework results and mouse area comparison of between an adult and a child

from Kinect camera to the user's right hand. As hand mouse must be easily accessible to everyone, mouse area is made dynamically. Especially, White rectangle reflects it and it is made to multiply user's height by 9/8 and to extend it proportionally each segment of it. With this, people of all ages can use mouse events easily using NUI/NUX framework.

The hand mouse is useful for recognizing big motion but not useful for recognizing sophisticated motion. For example, if a virtual keyboard is made in the NUI/NUX framework of the augmented reality using cameras, sophisticated motion will be required to press each key. Therefore, to implement more precisely the hand mouse, finger recognition is preceded especially.

Let the length of between the central coordinate values of the hand and the first knuckle of the finger be the threshold and draw a circle with a radius of it and then users can recognize the number of the fingers which meet the circle. Thus, as shown in Fig. 31.5, it is represented by the six cases. Using this, delicate hand gesture NUI/NUX framework can be studied.

The method which can check for Bending of the hand draws a circle getting larger from the center of the hand and then users can find the point of contact between the hand and the circle. This course is a prerequisite for the length of a finger and then algorithms must be embedded which find the length of the rest.

The left side of Fig. 31.6 is to represent the status of the NUI/NUX framework after running it and the right side shows mapping to a monitor. Because, in the left figure, white rectangle in the small window is mouse area, its upper left corner is mapped to the upper left corner of the monitor and its lower right corner is mapped to the lower right corner of the monitor (red line). By making Parallel movement as (+x, +y) from the client area coordinate (0, 0) in the small window to upper left corner of the white mouse area, it can be mapped from the small window to the monitor.

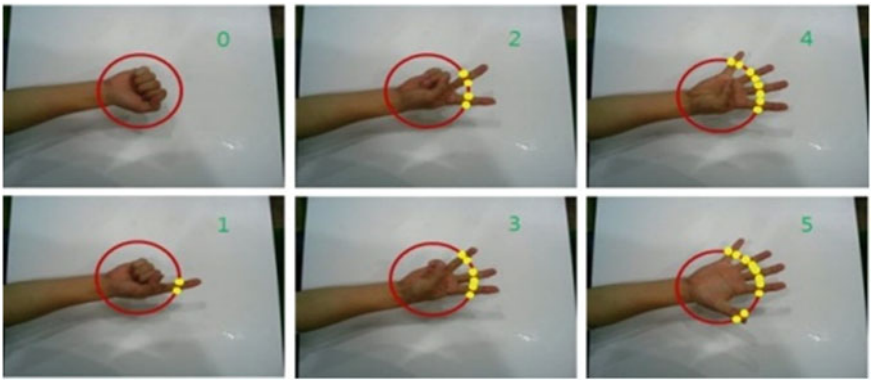


Fig. 31.5 Circles with a radius of the threshold value and the contact point of fingers

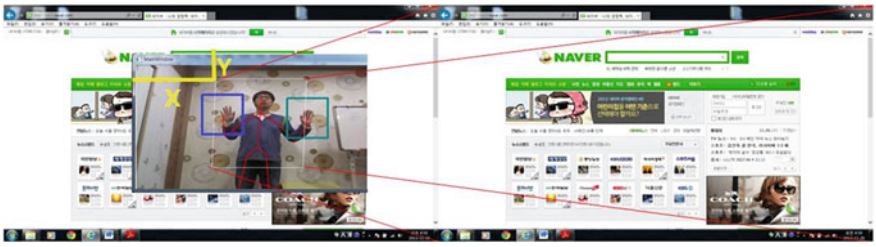


Fig. 31.6 Mapping between the implemented area and a monitor

31.4 Conclusion and Future Research

The purpose of this paper is to recognize intuitive hand gestures using KINECT camera and to map it on the monitor and to propose to design cloud system using XML. Differences between this paper and so far, what has been researched are described as follows.

- Not to use markers, NUI/NUX framework is implemented with bare hand.
- The mouse area is implemented dynamically for people of all ages to use it.
- Mouse event is implemented intuitively and correctly.
- Number of fingers and which fingers are stretched or not are recognized.

In future studies, multi-modal NUI/NUX will be designed to input voice and facial expressions for more delicate and rich interaction between computers and humans. Now as NUI/NUX motion recognition being utilized in a presentation, if the camera catches the center point of human to the target, it is expected that users give presentation remotely as he walks. Also, without doctors' taking and taking

off gloves, there are advantages which are fast nano-robot surgeries using NUI/NUX, and so they are expected to be able to be useful and excellent in the near future.

Acknowledgments This research is supported by Seoul R&BD Program (SS110008).

References

1. Chaing IT, Tsai JC (2012) Using Xbox 360 Kinect games on enhancing visual performance skills on institutionalized older adults with wheelchairs. In: Fourth IEEE international conference on digital game and intelligent toy enhanced learning, pp 263–267
2. Shiratuddinm MF, Wong KW (2011) Non-contact multi-hand gestures interaction techniques for architectural design in a virtual environment. In: The international conference on IT and multimedia at UNITEN (ICIMU 2011), Malaysia, Nov 2011
3. Ohya J, Kitamura Y et al (1993) Real-time reproduction of 3D human images in virtual space teleconferencing. In: Proceedings of '93 IEEE virtual reality annual international symposium, pp 408–414
4. Bau O, Mackay WE (2008) OctoPocus: a dynamic guide for learning gesture-based command sets. UIST 2008
5. Henrique C, Forster Q (2007) Design of gesture vocabularies through analysis of recognizer performance in gesture space. Intell Syst Des Appl 641–646
6. Bolt RA (1980) Put-that-there: voice and gesture at the graphics interface. In: International conference on computer graphics and interactive techniques, Association for Computer Machinery, pp 262–270
7. Roto V, Law E, Vermeeren A, Hoonhout J (2011) user experience white paper. Bringing clarity to the concept of user experience. Result from Dagstuhl seminar on demarcating user experience, 11 Feb 2011

Chapter 32
Evaluating Skin Hereditary Traits Based on Daily Activities

Jehyeok Rew, Young-Hwan Choi, Daehoon Kim, Seungmin Rho and Eenjun Hwang

Abstract Skin aging can be caused by intrinsic or extrinsic factors. Typical extrinsic factors include UV exposure, lifestyle, and environmental pollution. However, the effects on skin aging differ among individuals because of differences in hereditary traits. Identification of such hereditary traits requires much time and effort. In this study, we propose a scheme for identifying personal hereditary traits by correlating skin images with daily activities. We collected skin images and information on the daily activities of subjects. Skin texture ages were calculated based on various skin texture features, and damage scores were assigned based on activity-related parameters such as activity duration and intensity. By tracing skin ages and damage scores for each activity, we ranked their influence on skin aging and defined personal skin hereditary traits based on the results of the analysis.

J. Rew · Y.-H. Choi · D. Kim · E. Hwang
School of Electrical Engineering, Korea University, Anam-dong, Seongbuk-Gu, Seoul, South Korea
e-mail: rjh1026@korea.ac.kr

Y.-H. Choi
e-mail: work48@korea.ac.kr

D. Kim
e-mail: kdh812@korea.ac.kr

E. Hwang
e-mail: ehwang04@korea.ac.kr

S. Rho (✉)
Department of Multimedia, Sungkyul University, Anyang-si, South Korea
e-mail: smrho@sungkyul.edu

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_32, © Springer Science+Business Media Dordrecht 2014

32.1 Introduction

A deep interest in stopping or delaying the skin aging process has existed for a long time. Therefore, scientists and companies have been working on revitalizing the skin or improving skin conditions. According to recent studies, 2 types of factors can induce skin aging—intrinsic and extrinsic factors. Intrinsic (internal) factors, also known as the natural aging process, constitute a persistent process that normally begins in the mid-20s. Extrinsic (external) factors include environmental factors, such as exposure to sunlight or pollution. Various extrinsic factors accelerate the normal skin aging process and eventually lead to premature skin aging. The most influential extrinsic factors have been identified through experiments. Dermatologists have attempted to identify the main causes of skin aging from various perspectives. UV exposure, alcohol and nicotine consumption, skin type, and gender are typical factors that could lead to premature skin aging. In particular, smoking and UV exposure were identified as important factors contributing to premature aging of human skin [1–4]. Nicotine abuse induces the expression of matrix metalloproteinases (MMPs), which degrade the extracellular matrix (ECM) in human skin [5]. Andrea et al. [6] focused on airborne particle exposure and extrinsic skin aging, which has been known to result from chronic exposure to solar radiation and, more recently, to smoke. Kohl et al. [7] addressed the importance of extrinsic factors on skin aging and the detection of its mechanisms. Most factors that cause premature skin aging are normal human activities such as UV exposure due to outdoor activity, smoking, drinking, exercising, and eating. However, the effects of these extrinsic aging factors differ among individuals because of differences in skin hereditary traits. For example, individuals show different reactions to consumption of the same amount of alcohol. Similarly, small amounts of alcohol could cause different damages to the skin.

The identification of such hereditary traits requires much time and effort. In this study, we propose a scheme for discovering personal hereditary traits by correlating skin images and daily activities. Figure 32.1 shows the overall architecture of the proposed system. It is composed of 3 sections: data collection, data evaluation, and data analysis. The data collection section describes the collection of skin images and information on the ordinary activities of the subjects. In the data evaluation section, skin texture age is calculated using the support vector machine (SVM), and the damage score of each extrinsic aging factor is measured. Finally, the data analysis section ranks the damageable extrinsic aging factors using dynamic time warping (DTW).

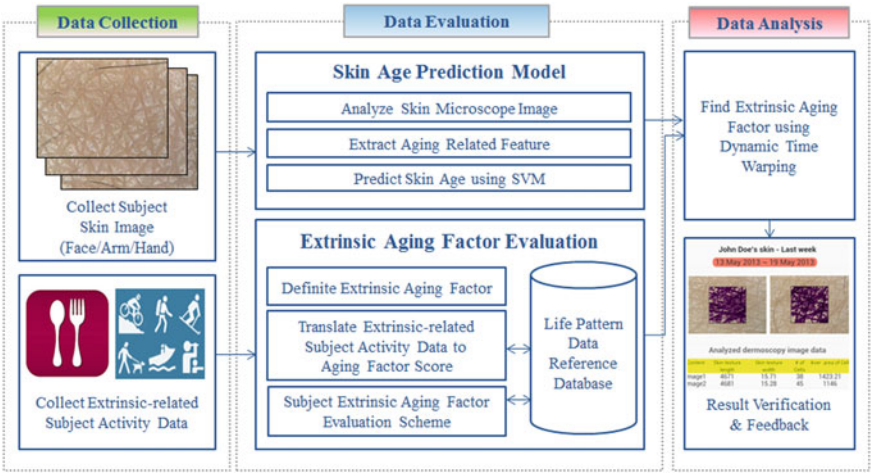


Fig. 32.1 System architecture

32.2 Evaluation of Skin Hereditary Traits

32.2.1 Feature Extraction and Prediction of Skin Age

In our previous studies [8, 9], we proposed the use of texture feature extraction and analysis schemes for evaluating the degree of skin aging from dermoscopy images. We defined several features such as skin texture length, width, depth and cell-related features, and established a prediction model to determine skin age using dermoscopy images of different body parts. For constructing the prediction model, we collected the texture features from dermoscopy images in different age groups and analyzed the features using the SVM.

32.2.2 Collecting Extrinsic Aging Factors

Skin condition can be determined by analyzing different factors such as UV exposure, smoking, exercising, and daily nutrition. Based on previous studies [1–7] and recommendations of dermatologists, we focused on 6 factors in this paper: eating, outdoor activity, drinking, exercising, smoking, and sleeping. These are the activities that many people perform on a daily basis and hence could have a profound effect on skin conditions.

In this paper, we collected information on ordinary life activities of subjects through a website that we created. Each subject recorded his/her daily activities on the website and the records were saved into the database. The database contained detailed information about each activity such as “the amount of calories in the

subject's intake for the day," "the time duration during which the subject enjoyed outdoor life," "the amount of alcohol consumed daily" etc. The server then translated these data into numerical values and calculated the skin damage score.

32.2.3 Calculating Collected Data for Normalization

The Harris-Benedict equation was used to estimate an individual's basal metabolic rate (BMR) and daily kilocalorie requirements. The resulting number is the recommended daily kilocalorie intake to maintain body weight [10]. Equation 32.1 shows the manner in which we calculated skin damage associated with the activity of eating based on the calculated BMR. As shown in the equation, the resulting damage was scaled to the range of 0–1. In other words, if a subject's intake was within a certain limit, then the maximum damage is 1. We maximized the limit to a value that was twice the kilocalories considered to be appropriate as BMR for a subject. The D_{BMR} indicates the recommended kilocalories per day and D_{take} indicates the actual intake in kilocalories per day.

$$D_{eating} = \min\left(\frac{|D_{BMR} - D_{take}|}{D_{BMR}}, 1\right) \tag{32.1}$$

The skin is often damaged by temperature and UV exposure. Temperature information and UV exposure values can be acquired by parsing the web pages of the meteorological administration agency and weather forecasting. We obtained the highest/lowest temperature, sunrise/sunset times, and daily UV exposure values. The UV exposure values were scaled from 0 to 1 where the lower value was the least harmful to the skin. To consider the effect of temperature and UV exposure, we calculated the time duration when the outside temperature was >27 or <5 °C in a day, and the UV exposure values based on the time of day, respectively. In Eq. 32.2, T_{start} indicates the start time of exposure to UV light and T_{end} indicates the end time of UV exposure.

Using Eq. 32.2, we can calculate $D_{outdoor}$ to determine the damage from UV exposure based on time.

$$\begin{aligned} W_{UV_exposure} &= \begin{cases} 0, & if \;\; T \le Sun_{rise} \\ \frac{T - Sun_{rise}}{UV_{max} - Sun_{rise}}, & if \;\; Sun_{rise} < T \le Sun_{max} \\ -\frac{T - Sun_{drop}}{Sun_{drop} - UV_{max}}, & if \;\; Sun_{max} < T \le Sun_{drop} \\ 0, & if \;\; Sun_{drop} \le T \end{cases} \\ D_{outdoor} &= \frac{T_{duration}\left(t < 5\,^{\circ}\mathrm{C} \;\; \text{or} \; t > 27\,^{\circ}\mathrm{C}\right) \times \sum_{T_{start}}^{T_{end}} W_{UV_exposure}}{T_{outdoor}} \end{aligned} \tag{32.2}$$

Alcohol is another important factor that can affect skin status. We therefore assessed the amount of alcohol consumed in terms of calories. However, we

allowed up to 1 glass of wine without any penalty, which amounts to 200 Cal. Equation 32.3 shows the manner in which we calculated the damage associated with drinking activity. Here, $C_{drinking}$ indicates the translated calories of alcohol consumption.

$$D_{drinking} = \begin{cases} D_{init}, & if \;\; 0 \leq C_{drinking} \leq 200 \\ \frac{|C_{Thresh} - C_{drinking}|}{C_{Thresh}}, & if \;\; 200 \leq C_{drinking} \leq 400 \\ 1, & if \;\; C_{drinking} > 400 \end{cases} \tag{32.3}$$

With regard to exercise, dermatologists recommended exercising for 30 min per day. Based on this recommendation and verified knowledge, we calculated the damage associated with exercise activity. To calculate the damage associated with exercise, information on the number of times and duration of exercise per day was required. To obtain the best result from the activity, a subject needs to complete at least 150 min of moderate aerobic activity per week or 75 min of vigorous aerobic activity per week.

Table 32.1 shows the intensity of exercise and calories burned per minute. Moderate exercise includes slow walking, standing and stretching. Intermediate exercise includes regular exercise such as biking, fast walking, and hiking. Vigorous exercise refers to intense exercise that burns calories instantaneously. The M, I, and V indicate the exercise intensity. The T_M, T_I, and T_V indicate the time duration of exercise. Equation 32.4 explains the manner in which we calculated damage associated with exercise activity.

$$D_{exer} = \begin{cases} D_{init}, & if \;\; T_{exer} = 0 \\ -\frac{(M \times T_M) + (I \times T_I) + (V \times T_V)}{T_{rec}} + D_{init}, & if \;\; 0 < T_{exer} \leq T_{rec} \\ 1 & if \;\; T_{rec} \leq T_{exer} \end{cases} \tag{32.4}$$

With regard to smoking, we used a simple rule: if a subject smoked, we used the highest score as the damage score i.e., if the subject is a smoker, we ignored the amount of smoking.

Sleeping plays an essential role in skin regeneration. According to dermatologists, humans need to sleep for 8 h from 10 p.m., because skin regenerates most actively between 10 p.m. and 2 a.m. Therefore, we calculated the damage associated with sleeping activity by considering the duration for which a subject slept and the time at which they went to sleep. Equation 32.5 shows the manner in which we calculated the damage associated with sleeping. In the equation, we considered 2 factors: "whether the subject slept the proper amount of time" and "how many hours overlapped with the recommended skin regeneration time,"

$$D_{sleeping} = \left| 1 - \frac{1}{2} \left(\frac{T_{sleeping}}{T_{rec}} + \frac{T_{user_regen}}{T_{Ref_regen}} \right) \right| \tag{32.5}$$

Table 32.1 Relation between calories and exercise intensity

Exercise intensity	Men (Kcal/min)	Women (Kcal/min)
Moderate (*M*)	3.3	2.3
Intermediate (*I*)	5.5	5
Vigorous (*V*)	16.6	12.5

32.2.4 Evaluating Skin Hereditary Traits

To evaluate the skin hereditary type, we assumed that the extrinsic aging factors are closely related to the skin texture age. We intuitively understand that drinking alcohol and exposure to UV light can cause skin aging. However, drinking alcohol and exposure to UV light may or may not influence skin aging in reality, because people have their own inherited characteristics. Hereditary factors have different effects on individuals exposed to the same extrinsic aging factors. In this section, we propose a method to evaluate the skin hereditary type. The definition of skin hereditary type, which ranges from sensitive to dull, is the rank obtained after defining extrinsic factors. In other words, if a subject's first rank is drinking alcohol, that subject is significantly affected by drinking alcohol, and small amounts of alcohol can cause skin aging. For skin hereditary type evaluation, we used the DTW method. The DTW provides reliable time alignment between reference and test patterns and works well for specific data that have their own characteristics. The advantage of using DTW is that it can analyze sequences of different lengths. Furthermore, the flexibility of the DTW method allows 2 time series that are similar but locally out of phase to align in a non-linear manner. In the present study, we used DTW to identify unique characteristics associated with changes in skin texture age and life activities.

$$\gamma(\mathrm{i},\mathrm{j}) = \mathrm{d}\left(q_i, c_j\right) + \min\{\gamma(i-1, j-1), \gamma(\mathrm{i}-1,\mathrm{j}), \gamma(\mathrm{i},\mathrm{j}-1)\} \tag{32.6}$$

In Eq. 32.6, the distance is designated to depict the greatest similarity between series by calculating the minimum distance between them. Suppose we have two time series, Q and C, of length n and m, respectively, where $\mathrm{Q} = q_1, q_2, q_3, \ldots, q_i, \ldots, q_n$ and $\mathrm{C} = c_1, c_2, c_3, \ldots, c_i, \ldots, c_n$. In our case, Q represents a sequence of normalized skin texture ages and C represents a sequence of normalized extrinsic aging factors.

To align two sequences using DTW, we construct a n-by-m matrix where (*ith*, *jth*) element of the matrix contains the Euclidean distance d(q_i, c_j) between the two points q_i and c_j. Each matrix element (i, j) corresponds to the alignment between the points q_i and c_j. A warping path W is a contiguous set of matrix elements that defines a mapping between Q and C. Calculating cumulative distances $\gamma(i,j)$ of the adjacent elements and minimizing warping path are defined in dynamic time warping method [11]. To calculate the DTW distance, using Eq. 32.7, we normalized the result of the observed skin texture age by transforming raw scores into z-scores to compare the extracted data with the results of the life activity category.

$$Z = \frac{Q - E[Q]}{\sigma(Q)} \tag{32.7}$$

The variable Q of observed skin texture age was standardized by subtracting its expected value E[Q] and dividing the difference by the standard deviation in Eq. 32.7. In the present study, we analyzed the similarities between the observed skin texture ages and the collected 6 life activity results. By using Eq. 32.6, we calculated the sum of the minimum value between the 2 vectors. A sum of the minimum distances that is close to 0 reflects a high similarity between the 2 vectors. However, if the sum of the minimum distances gives a high value, it indicates a low similarity between the 2 sequences. Using previously defined experimental methods, we identified the related extrinsic aging factors for each subject with high similarity. High similarity results (the results with close distance between 2 graphs) indicate factors that are sensitive for aging in subjects based on the observed and analyzed data. In other words, if a subject drank a lot of alcohol on a certain date and the skin texture aged on the subsequent day, then alcohol was ranked high in the hereditary type.

32.3 Experiment

For skin feature extraction and statistical estimation of skin age, the dataset consisted of 834 face, hand, and neck skin images from 238 subjects with ages ranging from 10 to 50 years [8, 9]. In addition, we collected information on the activities of 4 subjects for 20 days and measured the daily condition of their skin. To collect the life activity data of subjects, we surveyed the subjects using web pages. We then used the defined equations to translate those activities into numerical values ranging from 0 to 1. We also collected 4 skin texture images from subjects. To ensure accuracy, on day 1, we collected 10 images of the subject's body parts from the same position. By using DTW, we evaluated hereditary types by comparing the graphs of the extrinsic aging factor and the graph of skin texture age for each body part.

Table 32.2 shows the sum of minimum distances of extrinsic aging factors and observed skin texture age. As shown in Table 32.2, subject #1 shows a low average distance value of 0.68 in the sleeping section. However, the result shows a high distance value in outdoor and smoking activities. According to the definition of DTW, a low average distance indicates high similarity. The order of subject #1's skin hereditary type is Sleeping, Exercising, Eating, Drinking, Smoking, and Outdoor activity. This order reflects the ranks of the less damaging components on the hereditary type list. The hereditary types of each subject are defined in Table 32.2.

Figure 32.2 shows the representative results of the DTW distance between the extrinsic aging factors and skin texture age of each body part. Each body part showed a different similarity distance result for each life activity; however, their rank remained the same as that shown in Table 32.2.

Table 32.2 Average minimum distance

Rank	Subject #1	Subject #2	Subject #3	Subject #4
1	Sleeping (0.68)	Sleeping (0.68)	Eating (0.37)	Sleeping (1.09)
2	Exercising (1.37)	Eating (0.93)	Sleeping (0.58)	Eating (1.17)
3	Eating (1.41)	Exercising (1.11)	Exercising (0.88)	Exercising (1.31)
4	Drinking (1.46)	Drinking (1.31)	Drinking (2.60)	Drinking (1.85)
5	Smoking (4.15)	Outdoor activity (4.58)	Outdoor activity (6.07)	Outdoor activity (4.03)
6	Outdoor activity (6.16)	Smoking (4.91)	Smoking (6.40)	Smoking (4.47)

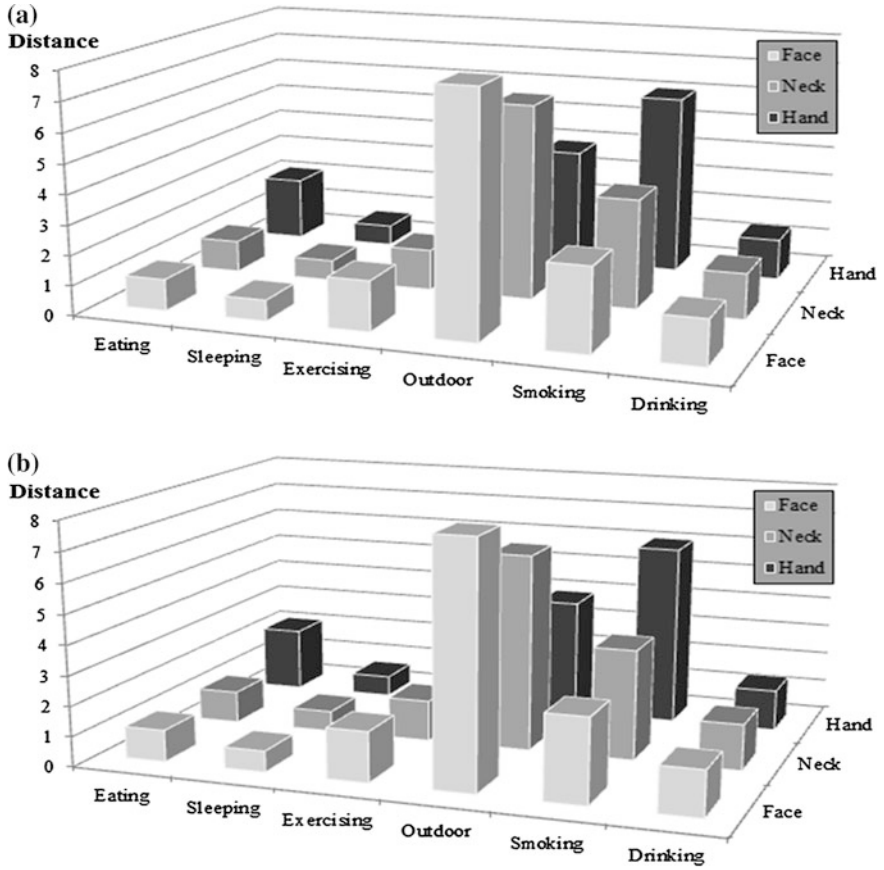


Fig. 32.2 DTW distance between extrinsic aging factors and skin texture age of each body parts. **a** Subject #1. **b** Subject #2

Our observations indicated that most subjects had the highest distance in the categories of outdoor and smoking. In other words, either the subjects were less affected in those 2 categories or there were no status changes in those 2 categories. We assume that they had similar life patterns with no status changes in those 2 categories. In other words, they preserved their skin at temperatures from 5 to 27 °C and their smoking status never changed. Our survey data also show the respective results.

32.4 Conclusion

In this paper, we showed the relation between extrinsic aging factors and skin texture age. By comparing these 2 factors, we evaluated the personal hereditary traits. For quantitative analysis, we used DTW for similarity matching between 2 time series sequences. We examined the skin texture age of subjects by using verified methods and compared the results to the translated subject's life activity data. Based on these data, we proposed equations that can translate life activity data into numerical values under certain conditions. Using these equations, we showed that life activity was related to skin status. The proposed scheme can be used for personal skin diagnosis and as a monitoring system in the future.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2061978) and the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the IT/SW Creative research program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2013-H0502-13-1119).

References

1. Farage M, Miller K, Elsner P, Maibach H (2008) Intrinsic and extrinsic factors in skin ageing: a review. Int J Cosmet Sci 30(2):87–95
2. Bermann P (2007) Aging skin: causes, treatments, and prevention. Nurs Clin North Am 42(3):485–500
3. Morita A, Torii K, Maeda A, Yamaguchi Y (2009) Molecular basis of tobacco smoke-induced premature skin aging. J Invest Dermatol Symp Proc, Nat Publishing Group 14(1):53–55
4. Lei Y, Akimichi M, Takuo T (2001) Skin aging induced by ultraviolet exposure and tobacco smoking: evidence from epidemiological and molecular studies. Photodermatol Photoimmunol Photomed 17(4):178–183
5. Lahmann C, Bergemann J, Harrison G, Young AR (2001) Matrix metalloproteinase-1 and skin ageing in smokers. Lancet 357(9260):935–936
6. Andrea V, Tamara S, Ulrich R, Dorothea S, Mary M, Ursula K, Jean K (2010) Airborne particle exposure and extrinsic skin aging. J Invest Dermatol 130(12):2719–2726
7. Kohl E, Steinbauer J, Landthaler M, Szeimies R (2011) Skin ageing. J Eur Acad Dermatol Venereol 25(8):873–884

8. Choi Y, Tak Y, Rho S, Hwang E (2013) Skin feature extraction and processing model for statistical skin age estimation. Multimedia Tools Appl 64(2):227–247
9. Kim K, Choi Y, Hwang E (2009) Wrinkle feature-based skin age estimation scheme. In: Proceedings of international conference on multimedia and Expo, pp 1222–1225
10. Harris J, Benedict G (1918) A biometric study of human basal metabolism. Proc Natl Acad Sci 4(12):370–373
11. Keogh E, Ratanamahatana CA (2005) Exacting indexing of dynamic time warping. Knowl Inf Syst 7(3):358–386

Chapter 33
A New Fast Encoding Algorithm Based on Motion Activity for High Efficiency Video Coding (HEVC)

Jong-Hyeok Lee, Kalyan Goswami and Byung-Gyu Kim

Abstract High efficiency video coding (HEVC) has three units such as coding unit (CU), prediction unit (PU), and transform unit (TU). It has too many complexities due to improve coding performance. We propose a fast algorithm which can be possible to apply for both CU and PU parts. To reduce the computational complexity, we propose based on rate distortion cost of CU about the parent and current levels to terminate the CU decision early. In terms of PU, we develop fast PU decision based on spatio-temporal and depth correlation for PU level. The Experimental results verify that the proposed algorithm provides up to 51.70 % of time reduction for encoding with a small loss in video quality, compared to HEVC Test Model (HM) version 10.0 software.

Keywords HEVC · Coding tree block (CTB) · Coding unit (CU) · Prediction unit (PU) · CU splitting · PU decision

33.1 Introduction

A New generation video coding standard, called a high efficiency video coding (HEVC) [1], has been developed by the Joint Collaborative Team on Video Coding (JCT-VC) group. The JCT-VC is a group of video coding experts created by ITU-T Study Group 16 (VCEG) and ISO/IEC JTC 1/SC 29/WG 11 (MPEG) in 2010.

J.-H. Lee · K. Goswami · B.-G. Kim (✉)
Department of Computer Engineering, Sun Moon University, Asan, Republic of Korea
e-mail: bg.kim@mpcl.sunmoon.ac.kr

J.-H. Lee
e-mail: ljh0607@mpcl.sunmoon.ac.kr

K. Goswami
e-mail: kalyan_goswami@mpcl.sunmoon.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_33, © Springer Science+Business Media Dordrecht 2014

Although coding efficiency improvement is solved it has more computational complexity than the H.264/AVC [2] because of tools having high complexity and high resolution of sequences. Therefore, reduction of encoding time with convincing loss is an interesting issue.

The video encoding process in the HEVC has three unit for block structure: (a) a coding unit (CU) is basic block unit like macroblock in the H.264/AVC, (b) a prediction unit (PU) for performing motion estimation, rate-distortion optimization and mode decision, (c) a transform unit (TU) for transform and entropy coding. Encoding process encode from Coding Tree Block (CTB) having largest block size to its 4 child CUs, recursively.

In [3], a tree pruning algorithm is proposed that makes an early termination CU. Shen et al. [4] proposed an early CU size determination algorithm. In [5], a motion vectors merging (MVM) method is proposed that reduce the inter-prediction complexity of the HEVC. The MVM algorithm decide the PU partition size by using motion vectors.

This paper is organized as follows: In Sect. 33.2, the early CU splitting termination (CST) and fast PU decision method is described for proposed algorithm. Section 33.3 presents the coding performance of suggested algorithm combined by two methods. In Sect. 33.4, Concluding comments are given.

33.2 Proposed Work

33.2.1 Fast Coding Unit Splitting Termination (CST) Method

The main objective is to select the lowest RD cost before all combinations are finished. We have designed the ratio function of $CU_d(i)$ at depth 'd' can be define as:

$$r_i = \frac{RD_{cost}(CU_d(i))}{RD_{cost}(CU_{d-1})} \quad where \quad \sum_{i=0}^{3} r_i \leq 1 \tag{33.1}$$

When a CU is split, it will be divided up into 4 child CUs. Therefore, we can define a value of ratio function for each newly created CUs. This parameter r_i is a ratio of the RD costs of the current CU and its parent CU. When the ratio function of a child is lower than its siblings, it has low amount of chance to split in the next depth level. Accordingly, this parameter can be used as a threshold for making the decision to split a current CU to next depth level.

In previous algorithm [3], if a CU is coded as SKIP mode then no further splitting of a CU. Motivating from this fact, we have explored other prediction modes of the PU. INTRA and INTER predictions have different kinds of PU types. INTER modes are 2N × 2N, 2N × N, N × 2N, N × N (presently, this mode is

Table 33.1 Classified weight factors for all prediction modes according to the motion activities

Mode	Motion activity	PU_WF (γ_i)
SKIP	Homogeneous region with motionless	0
Inter 2N × 2N	Slow motion	1
Inter 2N × N, N × 2N	Motion between slow or moderate	2
Other inter and intra	Complex motion or texture	3

not used), 2N × nU, 2N × nD, nL × 2N and nR × 2N. INTRA modes are 2N × 2N and N × N. SKIP mode is consisted of 2N × 2N size. A CU is divided into 4 CUs of lower dimensions after the completion of all mode calculations. PU modes are processed recursively from the CTB to all children CU.

Table 33.1 have classified all the prediction modes according to motion activities. This kind of motion activity was explored in good approach in the H.264/AVC codec [6, 7].

Encoded CU has high chance for no splitting in SKIP mode (PU_WF is 0). To more investigate for PU_WF is 1 or 2, We have calculated a local average of the RD cost values of all encoded CUs. In this process, the final PU mode and the corresponding RD cost values are checked after encoded CU. Equations (33.2) and (33.3) perform after encoding of a CU check dimension (d) of CU and PU_WF (***wf***) in terms of satisfying $\boldsymbol{wf} < 3$.

$$sum_d^{wf} = \left(\mathrm{avg}_d^{\mathrm{wf}} \times count_d^{wf}\right) + RDcost_d^{wf}, \tag{33.2}$$

$$avg_d^{wf} = \frac{sum_d^{wf}}{count_d^{wf} + 1} \tag{33.3}$$

An absolute motion vector (MV) is also calculated in proposed method. An average of the motion vectors can achieve from both List_0 and List_1 in the HM reference software. We only use magnitudes of the MV directions for the x and y in order to calculate simply as shown in Eq. 33.4.

$$\mathrm{MV}_{\mathrm{abs}} = 0.5 \times \left(\sum_{i=0}^{1} |MV_x(i)| + |MV_y(i)|\right) \tag{33.4}$$

By using Eq. 33.4, we can know the motion of CU with directly relationship for MV_{abs}. The CU which has low MV_{abs} value has high change that will be not divided in current hierarchy of CU. According to above mentioned method, the proposed algorithm can determine which CU need to splitting early. The CU decision part in our algorithm can be divided into two stages. The first stage is to consider the CTU case. In terms of the CTU, we do not have any information from higher levels. Hence, ratio function cannot be calculated. In the second stage, other higher level CUs are considered.

This proposed CU splitting termination (CST) algorithm is check the PU mode weighting factor initially. If PU_WF is 0 meaning SKIP mode, the decision is

taken directly that there is no need splitting. If PU_WF is 1 or 2, then we cannot make any decision directly. Hence, this method is divided into two stages in order to use different checking. For the CTU case, this algorithm only use MV_{abs} and local average RD cost of equivalent dimension of PU for PU_WF = 1 and 2. If RD cost of current PU is less than average RD with same dimension and MV_{abs} is 0, Encoder is select that there is no need splitting in the CTU case for PU_WF = 1 and 2.

Otherwise, ration function, MV_{abs} and local average RD cost of suitable dimension of PU for PU_WF = 1 and 2 are used in non-CTU case. A non-CTU case is same process. Since it can use ration function with higher level's information, if the RD cost is less than average RD cost of suitable dimension of PU and ratio function (R) is less than 0.25, then it also decide that CU is no need to split.

33.2.2 Fast Prediction Unit Decision (FPU) Method

In this section, we propose an effective prediction unit selection method based on correlation and block motion complexity (BMC) that can be applied together with the CST above proposed. A natural video sequence has high spatial and temporal correlations. To analysis these correlations, we define predictor set (Ω) such as PU_1 to PU_7. Figure 33.1 shows position from current CU for each PU_i.

To define block motion complexity, we have designed weight factor for according to location from current CU as Table 33.2. Table 33.2 indicates to land more weight in spatial correlation and upper level (depth 'd−1') from current depth 'd' than other position. Based on the designed mode weight factor (Table 33.1) for motion activities, we defined a block motion complexity (BMC) [8] according to the predefined mode weight factors (γ) and a group of predictors in Ω in order to estimate the PU mode of the current CU.

The BMC has been represented as follow [8]:

$$BMC = \frac{\sum_{i=1}^{N} W(i,l) \cdot k_i \cdot r_i}{\sum_{i=1}^{N} W(i,l) \cdot k_i}, \tag{33.5}$$

$$W(i,l) = w_i + f(i,l), \tag{33.6}$$

$$f(i,l) = round(l \cdot T_{level}) \cdot T_t \cdot Tk_i \tag{33.7}$$

where N is the number of PUs equal to 7, $W(i, l)$ is function of the weight factor for each mode. γ_i is weight factor for mode of PU_i in Ω. The k_i denote which PUs have available. If current CTB is in boundary region in the frame, then above or left PU information are not available. When PU_i is available, k_i is set to 1; otherwise, k_i is equal to 0.

Equation 33.6 is an adaptive weighting factor function. The value of w_i is weight factor for PU_i. It has property of $\sum_{i=1}^{N} w_i = 1$. A value of l denotes the

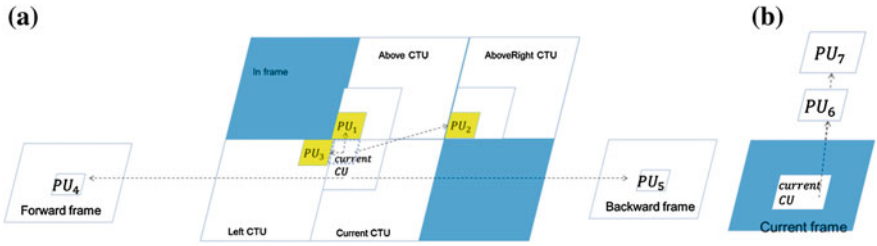


Fig. 33.1 The position of adjacent PUs from current CU ($\boldsymbol{CU_0}$) for **a** spatio-temporal and **b** depth correlation

Table 33.2 Weight factor for PUs (default setting and changing weight factor adaptively according to level of temporal layer)

Default setting	Index (i) in classified PUs	1	2	3	4	5	6	7
	w_i	0.2	0.2	0.2	0.05	0.05	0.2	0.1
Changes of weight factor for PUs	Index (i) in classified PUs	1	2	3	4	5	6	7
	w_i	0.2	0.2	0.2	0.1	0.1	0.1	0.1

temporal decomposition layer. The HEVC consist of three layers in 8 GOP according to distance between frames. Accordingly, temporal layer can represent $\text{Level}_{1,2,3,4}$ having 8, 4, 2, and 1 distance. If current frame close with reference frame, it has strong correlation than long distance case. Therefore, setting weight factors is changed when PU has 1 and 2 distance adaptively as second row in Table 33.2. $\text{Level}_{1,2,3,4}$ is set to 0.02 and is used for normalization of temporal level as l. Tk_i is a control value for changing the weight factors. Tk_i is a control value used to increase or decrease weight factors. It is set to 1, −2, or 0. PUs having the temporal correlations with PU_4 and PU_5 are set to be 1. Also, for a PU_6, it is set to be −2. The others are set to 0.

$$\begin{cases} BMC < Th_1, & \textit{Slow motion}, \\ Th_1 \le BMC < Th_2, & \textit{Medium motion}, \\ Th_2 \le BMC, & \textit{Complex motion or texture} \end{cases} \tag{33.8}$$

where Th_1 and Th_2 are set to 1 and 3 as the mode weight factor for slow motion and complex motion or texture in PU_WF of Table 33.1. If the BMC is slow motion, the proposed method performs SKIP mode and Inter 2N × 2N. In case of medium motion, mode decision process performs modes of slow motion, Inter 2N × N and N × 2N. When motion is complex, all modes are performed, then mode search process selects the best one.

Table 33.3 The performance summarization of original HM reference software 10.0 and the proposed algorithm

	CST			FPU [8]			MVM [5]			Proposed		
	ΔBit (%)	ΔY − PSNR	ΔT (%)	ΔBit (%)	ΔY − PSNR	ΔT (%)	ΔBit (%)	ΔY − PSNR	ΔT (%)	ΔBit (%)	ΔY − PSNR	ΔT (%)
Avg. of Class A	0.5	−0.03	−15.86	0.61	−0.03	−38.35	0.87	−0.03	−22.98	1.93	−0.09	−43.66
Avg. of Class B	−0.4	−0.03	−15.34	0.38	−0.02	−45.81	0.50	−0.01	−20.11	0.20	−0.06	−51.70
Avg. of Class C	0.66	−0.05	−25.56	0.64	−0.04	−41.06	0.70	−0.02	−25.48	2.80	−0.14	−48.42
Avg. of Class D	0.44	−0.07	−17.04	1.06	−0.05	−44.76	0.91	−0.02	−23.12	3.42	−0.18	−51.09
Total average	0.3	−0.04	−18.45	0.67	−0.04	−42.50	0.75	−0.02	−22.92	2.09	−0.12	−48.72

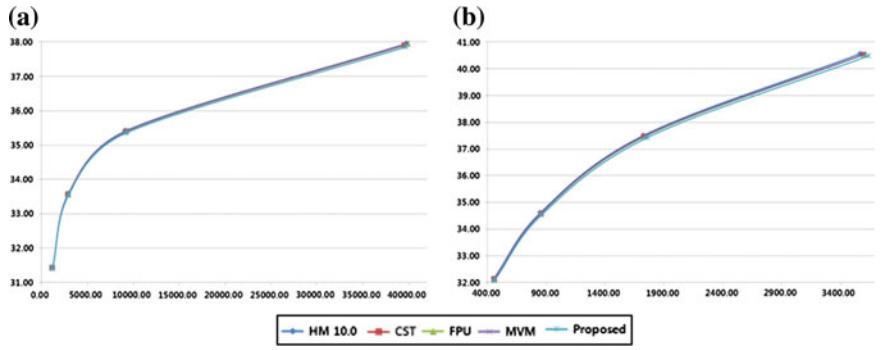


Fig. 33.2 Rate-distortion (*RD*) *curves* for **a** BQTerrace and **b** BasketballDrill sequences for Class B and Class C in random access, main condition

33.3 Experimental Results

The proposed algorithm was implemented on HM 10.0 (HEVC reference software). Test conditions were random access using RA-Main. Standard sequences with 50 frames were used from two or three sequences per each Class with various QP values (22, 27, 32, 37). Details of the encoding environment can be seen in JCTVC-L1100 [9].

Table 33.3 shows results for comparisons between the original HM 10.0 software and the proposed algorithm without any fast options. Moreover, we have considered 3 other algorithms (namely CST, FPU and MVM) which are shown in this table. The time reduction performance of the proposed method is almost 48.72 % on average with some loss as 2.09 % and 0.12 (dB) losses both in bit-rate and Y-PSNR, respectively. It can be noted from the table, that our proposed algorithm gives superior result in terms of time reduction compared with other fast algorithm. However, it suffers from marginal bit rate and PSNR loss. But if we consider the time reduction, then the loss is negligible.

Figure 33.2 shows the RD performance. In Fig. 33.2a, it is shown that for the bigger QP values larger loss are generated for bit-rate. However, the proposed method is very similar to the original HM 10.0 software. There is negligible loss of bit-rate in Fig. 33.2b.

Although our algorithm has some loss for bit-rate and PSNR, it has significant time saving performance 43.66 %, minimum, and 51.70 %, max. In performance of Class B, the proposed algorithm can achieve very little loss as 0.2 % of bit-rate and 0.06 (dB) of PSNR, and 51.70 % of good time saving performance without any fast options.

Even if our proposed algorithm has some loss, it shows performance more than a other fast algorithm about time reduction. Our method also has good structure that can be running with other fast options in HEVC standard.

33.4 Conclusions

We have proposed a new fast coding algorithm both of early CU splitting termination (CST) and fast PU decision (FPU). The CST algorithm are used the RD costs for different CU dimensions and motion complexity for PU level. The FPU method is based on spatial, temporal and depth correlation information and adaptive weighting factor design using temporal distance. By combine with CST and FPU, our algorithm achieved, on average, a 48.72 % of time saving over the original HM 10.0 software with 2.09 % of bit-rate loss. Our algorithm is useful to where needed much time reduction like real-time video encoding systems.

Acknowledgments This research was supported by the KCC (Korea Communications Commission), Korea, under the ETRI R&D support program supervised by the KCA (Korea Communications Agency) (KCA-2012-11921-02001).

References

1. Sullivan GJ, Ohm JR, Han WJ, Wiegand T (2012) Overview of the high efficiency video coding (HEVC) standard. IEEE Trans Circuits Syst Video Technol 22(12):1649–1668
2. Wiegand T, Sullivan GJ (2007) The H.264/AVC video coding standard. IEEE Signal Process Mag II:148–153
3. Choi K, Jang ES (2012) Fast coding unit decision method based on coding tree pruning for high efficiency video coding. Opt Eng Lett
4. Shen L, Liu Z, Zhang X, Zhao W, Zhang Z (2013) An effective CU size decision method for HEVC encoders. IEEE Trans Multimedia 15(2):465–470
5. Sampaio F, Bampi S, Grellert M, Agostini L, Mattos J (2012) Motion vectors merging: low complexity prediction unit decision heuristic for the inter prediction of HEVC encoders. In: International conference on multimedia and Expo, July 2012, pp 657–662
6. Hosur PI, Ma KK (1999) Motion vector field adaptive fast motion estimation. In: International conference on information, communications and signal processing (ICICS'99)
7. Zeng H, Cai C, Ma KK (2009) Fast mode decision for H.264/AVC based on macroblock motion activity. IEEE Trans Circuits Syst Video Technol 19(4):1–11
8. Lee JH, Park CS, Kim BG, Jun DS, Jung SH, Choi JS (2013) Novel fast PU decision algorithm for the HEVC video standard. In: IEEE International conference on image processing, pp 1982–1985
9. Bossen F (2013) Common test conditions and software reference configurations. In: Joint collaborative team on video coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11 12th meeting, Jan 2013

Chapter 34
Smart Network Services Using Ticket Issuance in the TSM (Trusted Software Monitor) System

Su-Wan Park and Jeong Nyeo Kim

Abstract A smart network environment is a system where the user can use the network's services anytime, anywhere. To establish such an environment, many studies are continuously being conducted on wireless communication technology and mobile devices. The company that provides smart network services should have an established system for authentication and authorization of user. Besides, other studies have been focused on the authentication and efficiency of the mobile device and smart device. This paper aims to provide fast authentication and service in use of mobile/smart device by providing simultaneous authorization and authentication about user and device, and it is referred to as Trusted Software Monitor (TSM) system. In TSM system, one feature is that the mobile/smart device contacts to the external authentication server through the home authentication server every time. Another one is to use ticket to provide authentication between them.

34.1 Introduction

Further developments in information technology (IT) have given people various services, with the wired Internet service shifting more than wireless communication. In the future's ubiquitous environment, the user should be able to use of network services wherever the user wants. Effective provision of such services requires an efficient user authentication mechanism, authorization of the service authority, and accounting of user service fees. Thus, the user registers his or her

S.-W. Park (✉) · J. N. Kim
Electronics and Telecommunications Research Institute, #161, Gajeong-dong, Yuseong-gu, Daejeon, Republic of Korea
e-mail: parksw10@etri.re.kr

J. N. Kim
e-mail: jnkim@etri.re.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_34, © Springer Science+Business Media Dordrecht 2014

mobile device to the home authentication server to manage efficiently and utilizes it to gain authentication and authorization to use the service [1]. However, it is impossible to authorize the user if the user accesses from an external network which is on the outside of home network. Thus, it is impossible to provide the service to user [2, 3]. Therefore, mobile device accesses an external authentication server through the home authentication service, whereas the home authentication service makes it possible for the external authentication server to provide mobile service to the user through its authentication information. Likewise, security technology is applied to ensure the safety of data transmitted through the external authentication service. Therefore, this paper aims to provide fast authentication and service by providing simultaneous authorization and authentication. In Sect. 34.2, we discuss the requirements of existing authentication and authorization methods. In Sect. 34.3, we provide details of our proposed method, and Sect. 34.4, the proposed method is analyzed based on requirements mentioned in Sect. 34.2. Lastly, we cite directions for future studies and conclusion in Sect. 34.5.

34.2 Security Requirement

The most important aspect in TSM is authentication, where the rightful user is confirmed based on the authorization of the service provision and consideration of the service charge. Authorization refers to the authority to access and use the service after authentication [4–8]. The TSM applies various authentication protocols, one of which is the utilization of ID and password. It involves checking if the ID and password supplied are registered in the home authentication server, and how safely the ID and password are submitted to the home authentication server. When the mobile device which is accessed from a user requests for authentication through an external network, it is necessary to provide a security method to transmit data from the external network. One way to ensure safety is having the mobile device and home authentication server shared symmetric keys and use encryption. The key distribution for the communication and service is provided to the home authentication server as it becomes the center. The external authentication server can use the key to authenticate the mobile device, and the mobile device can again create the key through key creation factors transmitted from the home authentication server. This process requires the following security requirements [9, 10]:

- Only the rightful entity has to be able to check the message.
- The forged message which is transmitted needs to elimination and must not be able to modulate on center.
- If forgery and modulation are occurred, there must be no possibility of leak of confidential information.
- The rightful user should be able to confirm that he or she is the user for the rightful server.

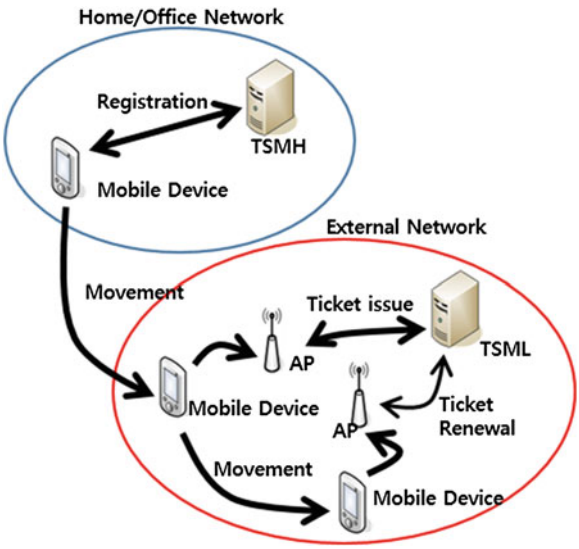


Fig. 34.1 Proposed method framework

- It should not be possible for any third party to make fraudulent use of the message transmitted by the user.
- A third party should not be able to impersonate as the rightful user of the authentication server.
- Ciphers used by the user and external authentication should be verifiable to ensure if they match.

Authorization is the setting of authority to authenticate information, and determines the range or extent service is to be provided. For this authorization message, the following security items are required:

- Forging or modulating the authorization message should be impossible.
- The rightful entity should be able to verify the authorization message.
- It should be invulnerable to any replay attack.

This paper proposes an authentication and authorization method that satisfies with the above criteria and will discuss the ticketing procedure as a method to provide fast mobility (Fig. 34.1).

34.3 Proposed Method

In this study, we present a ticketing method to provide service according to the movement of the mobile device, whose safety is ensured through the advanced registration at the home authentication server. Also, as it considers the movement between different APs in the External network, efficiency in the movement is

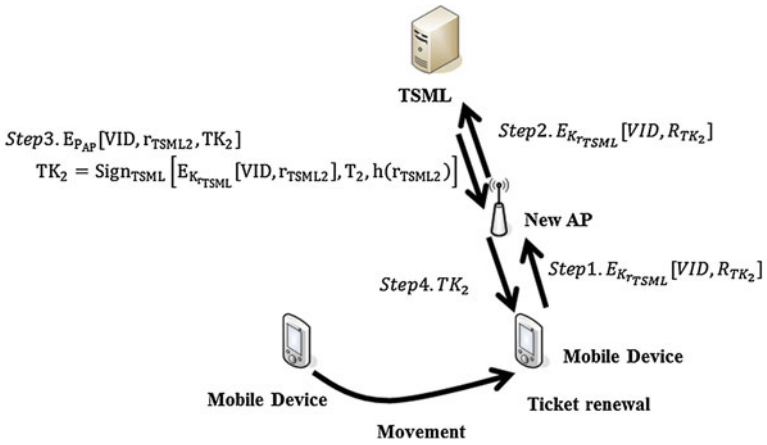


Fig. 34.2 Ticket renewal protocol diagram

ensured between APs with the ticket that the external authentication server issues. The proposed method consists of four phases, namely registration, authentication, authorization, and ticket renewal. When seeing Fig. 34.2. The mobile device is registered in home authentication server. The mobile device moves toward the external network and issues the ticket.

34.3.1 System Symbol

The system symbols used in this proposed method are as follows.

- *: Entity of the communication (User: user, TSMH: home authentication server, TSML: external authentication server)
- MID: ID of the mobile device user
- VID: Virtual ID used in the external network
- h(): the public one-way hash function
- r*: Random number that * selected
- T*: Time stamp
- S_No: Serial number that the home authentication server generates and assigns to the registered mobile device
- Sign*: Value of the electric signature of *
- P*: *'s public key

34.3.2 Registration Protocol Phase

The registration phase is the registration of the mobile device to the home authentication server that key distribution is made through a safe communication path. Through the following process, the mobile device of user is registered to the home authentication server.

Step 1: The mobile device sends mobile device ID (MID), random number (r_{U1}) and the digital signature value ($Sign_{U1}(h(MID||r_{U1}))$ to home authentication server (TSMH).

Step 2: The home authentication server checks the digital signature and registers MID and $h(MID||r_{U1})$. After that, it assigns a serial number (S_No) to the mobile device and transmits the hashed value (h(S_No)) and the digital signature value of the home authentication server ($Sign_{TSMH}(h(S_No))$.

34.3.3 Authentication Protocol Phase

The user may require service from the external network with a registered mobile device. Therefore, authentication is required to the home authentication server in order to confirm if the user is the rightful user. The home authentication server establishes the key that can be used in the external authentication server, which makes it possible for the external authentication server and the mobile device to engage in safe communication. At this point, the communication method should not be vulnerable to forging, retransmission, or modulation.

Step 1: The mobile device transmits the ID of the mobile device (MID), random number (r_{U2}), and the hash value ($h(h(S_No)||h(MID||r_{U1})||r_{U2})$) to a home authentication server.

Step 2: The home authentication server verifies the transmitted data and if verification is correctly made, it establishes a key ($K_{mn} = h(h(MID||r_{U1}) \; ||T_1)$).

Step 3: The home authentication server transmits the created key (K_{mn}) and the mobile device authentication message ($E_{P_{TSML}}[K_{mn}, T_1, sign_{TSMH}(MID)]$) to the external authentication server using the public key of the external authentication server.

34.3.4 Authorization Protocol Phase

The authorization phase is where the external authentication server provides service to the mobile device wherever it moves within the external network. Also,

with the provided session key, safety communication path is ensured in the communication of the mobile device.

Step 1: The external authorization server (TSML) creates a virtual ID ($VID = h(MID||r_{TSML})$) and a ticket ($TK_1 = Sign_{TSML}[E_{K_{mn}}[VID, r_{TSML}], T_1, h(r_{TSML})]$) of the accessed mobile device.
Step 2: The external authorization server codes the contents ($E_{P_{AP}}[VID, r_{TSML}, TK_1]$) as a cipher using AP's public key and transmits it to AP.
Step 3: AP decodes the message as a private key of AP, and the virtual ID (VID) becomes the ID of the mobile device. The shared symmetric key uses r_{TSML} as input value.
Step 4: AP provides ticket (TK_1) to the mobile Device.
Step 5: The mobile device creates a key (K_{mn}) with the ticket (TK_1) and can obtain r_{TSML} as the key input value.

34.3.5 Ticket Renewal Protocol

Ticket renewal makes it possible for the mobile device to use for consistent service by renewing the existing ticket, which uses to access the new AP as it moves the AP within the external network.

Step 1: The mobile device connects to the new AP and transmits a virtual ID (VID) and ticket requirement (R_{TK_2}) using the session key ($E_{K_{r_{TSML}}}[VID, R_{TK_2}]$) used previously.
Step 2: The new AP requires a renewed authorization ticket as it provides message transmitted from the mobile device to the external authentication server (TSML).
Step 3: External authentication server checks the transmitted content, renews the ticket ($TK_2 = Sign_{TSML}[E_{K_{r_{TSML}}}[VID, r_{TSML2}], T_2, h(r_{TSML2})]$), and reissues it ($E_{P_{AP}}[VID, r_{TSML2}, TK_2]$) to AP.
Step 4: New AP obtains the ticket (TK_2) as it decodes the message as private key and provides the renewed ticket (TK_2) to the mobile device.

34.4 Analysis of the Proposed Method

We verify if the proposed method satisfies the security requirements proposed in Sect. 34.2 (Table 34.1).

Table 34.1 Analysis of proposal method

Classification	Technology of the proposed method		Note
Efficiency	Mobile device	Utilization of symmetric key and hash (K_{mn}, r_{U1}, r_{U2}, r_{TSML})	Computation of the mobile device utilizes symmetric key technology, random number and hash function only
	Authentication server	The user management ticketing	In user management, service is provided to the rightful in the composition of the ticket
Security	• The user can use the service anonymously(utilization of VID)		Anonymity is possible for additional privacy when using of the service
	• When moving in the same external network, previous TSML has no right		Without a separate authentication process, it is impossible to utilize TK_2 because it doesn't know the session key

- Confirmation of only rightful entity

Each message is transmitted as it is coded as public key and symmetric key (Kmn), and the user and the server can safely communicate through the advanced registration phase. In the later renewal of the key, a new key is issued to make tampering the previous message impossible, and the user and the server can safely communicate through the advanced registration phase. In the later renewal of the key, a new key is issued.

- Forgery and modulation

To forge or modulate the message, the signature should be forged as well as the random numbers (r_{U1}, r_{U2}) of the user in the initial registration process. Also, it should be possible to verify the random numbers in each message establishment process, as it provides the hash value of the random number, it is impossible to forge or modulate the message again even for the rightful entity. For example, to change the value of the ticket, both the r_{TSML} and VID, $h(r_{TSML})$ value have to be changed, as well as the signature value.

- Rightful server and user confirmation method

The proposed method presents two means of checking the rightful server and user. One is the signature based on the public key and the other is the random number value, which is registered in advance. By registering ID and random number value (r_{U1}, r_{U2}) in advance, the rightful user can utilize it in the communication as the user knows the random number value afterwards. Moreover, security is assured as the value of the random number which is completely inaccessible to any third party, making it impossible for a third party to create a message. In the process of key creation, it is possible to confirm for the previous user with the coded message utilizing the previous session key.

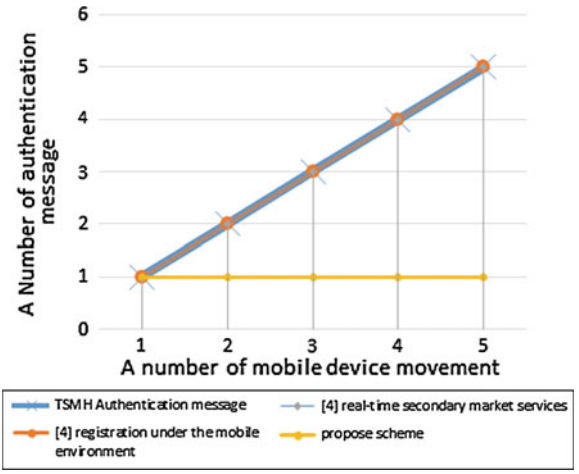


Fig. 34.3 A number of TSMH authentication message

- Authentication of created key

With a newly created session key, it should be checked if the same key is created. Coding a random message involves transmitting and rechecking the decoded message and making it possible to authenticate if the same session key has been created.

The proposed method decreases overhead of home authentication server (TSMH). But overhead of external authentication server increases by this reason. Publication of the ticket is because the external authentication server provides that. If authentication time at home authentication server time takes 0.5, external authentication server will take 1, because of accessing home authentication sever. The overhead of external authentication server increases, as result, we can see that overhead of authentication time is better than the existing methods.

When seeing Fig. 34.3, the proposed method decreases overhead of home authentication server (TSMH).

34.5 Conclusion

In this proposed method, the home authentication server authenticates the mobile device accessing an external network and transfers the device's information for creating an authorization ticket to the external authentication server. Using the ticket created in the external authentication server, the device accessed to the AP is authorized and then receives the service. Also, if user moves within the external network, ticket renewal is provided. The proposed system presents added efficiency; as it no longer requires the issuance of a new session key, approach the home network, and another authentication. With this method, it is possible to go

beyond the limits of the mobile device with safe and fast service. Although this method does not include encryption key management, the management methods of the keys utilized in the encryption will be treated in the next study. Especially, the use of encryption key will increase along with the growing of accessing mobile devices. Therefore, it would be necessary for future works that focus on key management method for various mobile devices.

Acknowledgments This work was supported by the IT R&D program (10043959, Development of EAL 4 level military fusion security solution for protecting against unauthorized accesses and ensuring a trusted execution environment in mobile devices) of KEIT/MSIP (Ministry of Science, ICT and Future Planning), Korea.

References

1. Park JM, Bae EH, Pyeon HJ, Chae K (2003) A ticket-based AAA security mechanism in mobile IP network. In: ICCSA 2003, pp 210–219
2. RFC 2865-Remote authentication dial in user service (RADIUS)
3. IEEE Standard 802.1X-2001 (2001) IEEE standard for local and metropolitan area networks-port-based network access control, June 2001
4. Tang Q, Mitchell J (2005) On the security of some password-based key agreement schemes. Cryptology ePrint Arch
5. Chen Y, Boult T (2002) Dynamic home agent reassignment in mobile IP. In: IEEE-WCNC02
6. Zhou Y, Wu D, Nettles SM (2005) On the architecture of authentication, authorization and accounting for real-time secondary market services. Int J Wireless Mobile Comput 1–8
7. RFC 2905-AAA authorization application examples
8. RFC 2904-AAA authorization framework
9. Kang SI, Lee DG, Lee IY (2008) mobile network service using ticket issuance in the AAA system. In: International symposium on ubiquitous multimedia computing, 2008. UMC '08, pp 141–144
10. Park S, Kim JN, Shin SU (2013) Efficient DRM mechanism of scalable contents based on H.264/SVC in convergence environment. Multimedia Tools Appl

Chapter 35
Macro Modeling Approach for Semi-digital Smart Integrated Circuits

Zaniar Hoseini, Kye-Shin Lee and Byung-Gyu Kim

Abstract This work presents a macro modeling approach for semi-digital smart integrated circuits. The proposed macro model models the behavior of the key circuit block used for semi-digital smart integrated circuits which is the time-to-voltage converter. Furthermore, the macro model can accurately analyze the non-idealities and error sources which can be used as the guideline for design optimization and calibration scheme implantation of semi-digital smart integrated circuits.

Keywords Macro modeling · Semi-digital design · Smart integrated circuits

35.1 Introduction

The semi-digital smart integrated circuit design trend is expected to be the desirable future solution for a variety of high performance electronic system applications including ubiquitous wireless networks, intelligent sensor systems, smart appliances, and health care monitoring systems [1–3]. This reduces the total manufacturing cost, improves the system performance, and minimizes the chip area. However, to enable power and technology scalable designs, the smart

Z. Hoseini · K.-S. Lee (✉)
Department of Electrical and Computer Engineering, The University of Akron, Akron, OH 44325, USA
e-mail: klee3@uakron.edu

Z. Hoseini
e-mail: sh101@zips.uakron.edu

B.-G. Kim
Department of Computer Engineering, Sun Moon University, Asan-si, Chung-nam, South Korea
e-mail: bg.kim@mpcl.sunmoon.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_35, © Springer Science+Business Media Dordrecht 2014

integrated circuits force the analog circuits to be integrated with digital circuits. Although digital circuits can be effectively integrated even with different technology nodes, analog circuits still face a lot of design challenges when they are highly integrated with digital circuits. Several reasons for this are as follows: the design deals with transistors which are basically non-linear circuit elements, the design requires good understanding of DC biasing and small signal analysis, needs good control over non-idealities due to process variation, device mismatch, and various noise sources. Furthermore, unlike the digital circuit design, major aspects are highly depending on the insight of the designer instead of the circuit design tools. Moreover, layout is another factor that affects the circuit performance. In order to overcome the above mentioned problems, the analog blocks must drastically reduce the analog circuit components while maintaining the performance, which leads to a *digital like* design concept [4–7].

So far, we have applied the semi-digital design concept to realize a successive approximation (SAR) type and a cyclic type analog-to-digital converter [8, 9], where the time-to-voltage converter (TVC) and voltage-to-time converter (VTC) are the two key circuit blocks for the semi-digital smart circuits. However, to extend the application of semi-digital smart circuits, a more clear understanding of the key circuit blocks are needed, which enables design optimization and calibration scheme implementation. Toward that end, we developed macro models for the TVC and VTC considering the major non-idealities and error sources. Furthermore, both macro model and transistor level circuit for TVC and VTC were implemented using the CMOS 0.35 μm and the CMOS 65 nm technologies, and the performance is verified through circuit level simulations.

35.2 Error Analysis of TVC and VTC Circuits

Figure 35.1 shows the block diagram and a typical circuit implementation of TVC and VTC structures. In the TVC circuit of Fig. 35.1a, when the input pulse goes high, the current source starts to charge the capacitor with a constant current value. Therefore, the final voltage of the capacitor is proportional to the pulse width of the input pulse. Similarly, in the VTC circuit of Fig. 35.1b, the initial voltage V_0 of the capacitor is discharged linearly by a constant current source, therefore, the time it takes for capacitor to get discharged, is proportional to the capacitor initial voltage. The circuit equations of charging in TVC and discharging in VTC circuits are given in (35.1) and (35.2), respectively.

$$V_C(t) = \frac{I}{C} \cdot t \tag{35.1}$$

$$V_C(t) = V_0 - \frac{I}{C} \cdot t \tag{35.2}$$

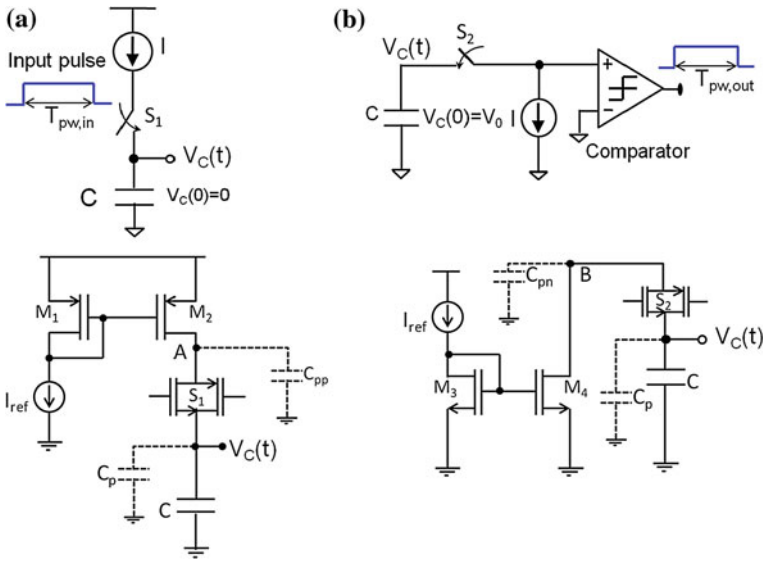


Fig. 35.1 Block diagram and circuit implementation of **a** TVC and **b** VTC structures

According to (35.1) and (35.2), the ideal charge and discharge process of the capacitors with a constant current source occurs linearly, however, as shown in the circuit implementation of Fig. 35.1, there are several error sources such as parasitic capacitances in different nodes, infinite output resistance of current mirror (in this case just transistor output resistance), switch on resistance, charge injection and clock feed through that cause error in TVC and VTC blocks and degrade the accuracy of overall system. The focus of this paper is to give a qualitative analysis on the effects of the first two mentioned error sources for CMOS 0.35 μm and CMOS 65 nm technologies.

In the TVC circuit shown in Fig. 35.1a, when the input pulse is low, M_2 is off because $V_{DS2} = 0$ and the parasitic capacitor C_{pp} at the drain of M_2 is charged to V_{DD}. The initial charge of capacitor C is also supposed to be zero. When the input pulse goes high, since the on resistance of the switch is small, a very fast charge sharing (because of very small time constant) occurs between C_{pp} and the capacitor C (and also the CMOS switch parasitic capacitance C_p at the output node). As a result, depending on the size of C_{pp} and the V_{DD} voltage, an initial error offset voltage is generated at the output. The other parasitic capacitor which is connected to the output node has minor effect on the accuracy of circuit since it is very small compared to C. When M_2 turns on, it starts to charge the capacitor C and the voltage on this capacitor increases. In this situation, drain to source voltage of M_2 changes continuously and leads to changes in the current of M_2 which is desired to be constant.

Therefore, another important parameter that adds nonlinearity to the TVC is the infinite output resistance of the current mirror. Note that since this error has a

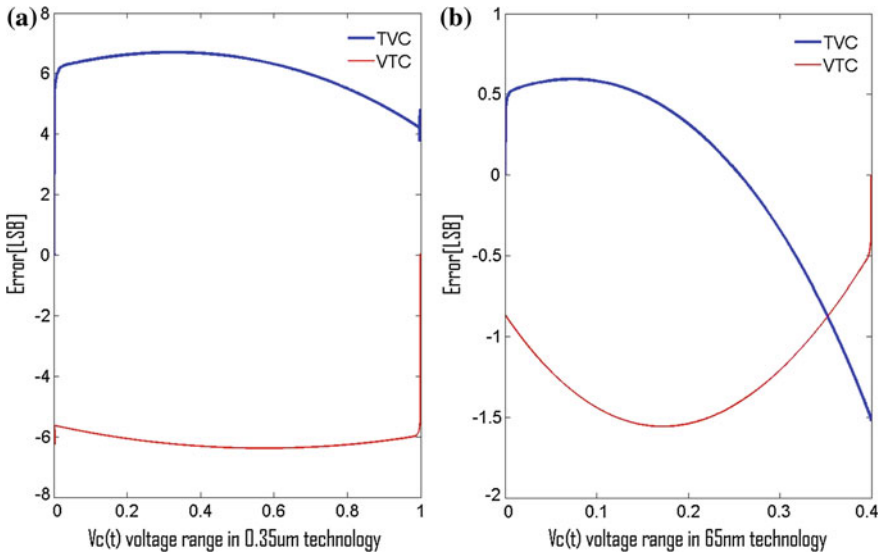


Fig. 35.2 The error between $V_C(t)$ of TVC and VTC circuits with $V_C(t)$ values from Eqs. (35.1) and (35.2), respectively in (**a**) 0.35 μm technology and (**b**) 65 nm technology

nonlinear and variable nature, more complicated calibration techniques are required to reduce its effects compared to the previous error source. When the voltage at node A is lower than the drain voltage of M_1, drain current of M_2 is larger than its desired value and when V_A exceeds the drain voltage of M_1, the current drops below the desired value. The characteristics of this error mostly depend on the channel size of transistor and the V_C voltage range. In order to evaluate the error performance of TVC circuit, voltage difference between the circuit simulation result for V_C and its corresponding ideal value from Eq. (35.1) has been calculated and expressed based on a portion of LSB in two technologies and are shown in Fig. 35.2. In the circuit implementation, the transistor sizes are scaled according to technology properties, yet the W/L ratios are kept equal. The pulse width of input pulse is adjusted so that the V_C voltage ranges from 0 to 1 V and 0 to 0.4 V in 0.35 μm and 65 nm technologies, respectively. A resolution of 8 bits is considered for error calculation; therefore, the LSB value for each technology is equal to (V_C voltage range)/$2^{\text{resolution bits}}$.

When the input pulse goes high, the ideal value of V_C from (35.1), is around zero, however, as explained before, in circuit simulation, due to error caused by the charge sharing of C_{pp}, there is an initial large error around +6 LSB for 0.35 μm and +0.5 LSB for 65 nm between the ideal value and circuit simulation result. Regarding the smaller size of transistors in 65 nm technology, it seems that the voltage error caused by parasitic capacitor at drain of M_2 in this technology is less critical compared to 0.35 μm technology. When the transistor M_2 starts to charge the capacitor C, the voltage

Variation in node A leads to channel length modulation effect which causes variation in charging current and hence changes the error value. From Fig. 35.2, it is clear that the current variation in 65 nm technology is more than 0.35 μm due to smaller channel length of transistors used as current mirror. Note that in the charging case depending on the final desired value of V_C, parasitic capacitance error and current error may be additive or subtractive and therefore leads to error of zero in some values of V_C. In 0.35 μm technology, since there is enough voltage room to keep M_2 in saturation, it is possible to increase the output resistance of current mirror by making the width of M_2 small and its L large while keeping the parasitic and voltage range of V_C at reasonable values. However, in 65 nm technology, due to low voltage operation, the ratio of W/L cannot become very small due to constrain that exist on overdrive voltage of M_2.

The error performance of VTC circuit of Fig. 35.1b is also evaluated and is shown in Fig. 35.2b. The initial voltage V_0 of capacitor C is 1 V and 0.4 V in 0.35 μm and 65 nm technologies, respectively. In this circuit, when the CMOS switch is off, the transistor M_4 is off, so the parasitic capacitor C_{pn} at its drain is discharged to ground. When the CMOS switch is connected, a very fast charge sharing takes place between the capacitor C and C_{pn} that causes an error offset voltage drop at the beginning of discharge process. As a result, after connecting the switch, while the ideal value of the capacitor voltage V_C according to (35.2) is very close to V_0, as shown in Fig. 35.2; the corresponding voltage value in the circuit implementation is less than the initial voltage value and hence there is an error of −6 LSB and −1.6 LSB in 0.35 μm and 65 nm technologies, respectively. Then, transistor M_4 turns on and discharges the capacitor C. In this stage, the voltage decrease in the drain source of M_4 causes the current variation in current mirroring. As mentioned before, the parasitic capacitance error is more evident in 0.35 μm technology while the circuit of 65 nm technology is mostly affected by infinite output resistor error. Note that in order to keep M_4 in saturation when C is completely discharged a constant offset voltage is added to its charge in both TVC and VTC that is not shown here.

35.3 Proposed Macro Model

Figure 35.3 shows the proposed macro model for analyzing the behavior of TVC and VTC circuits. In order to keep the model simple, fast and practical, in this stage just two main error sources; parasitic capacitors and current mirror infinite output resistance are included. The macro model circuit of TVC block is shown in Fig. 35.3a. In this macro model, the current mirror is replaced with r_{o2} which is a rough estimation of the output resistance of M_2, current source I_{ref} and a DC voltage source V_{d1} which is the drain voltage of M_1. In the circuit of Fig. 35.1a, the voltage difference between node A and drain of M_1 causes error between the I_{ref} and the current of M_2. Therefore, in the model of Fig. 35.3a, this current error is created by the difference between V_{d1} and the voltage at node A divided by r_{o2}.

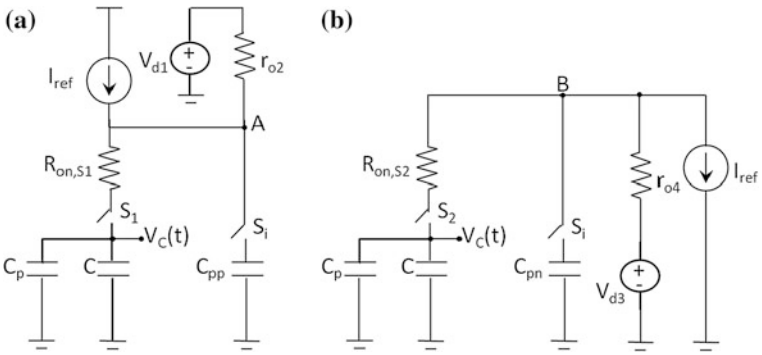


Fig. 35.3 Proposed macro models for **a** TVC circuit and **b** VTC circuit

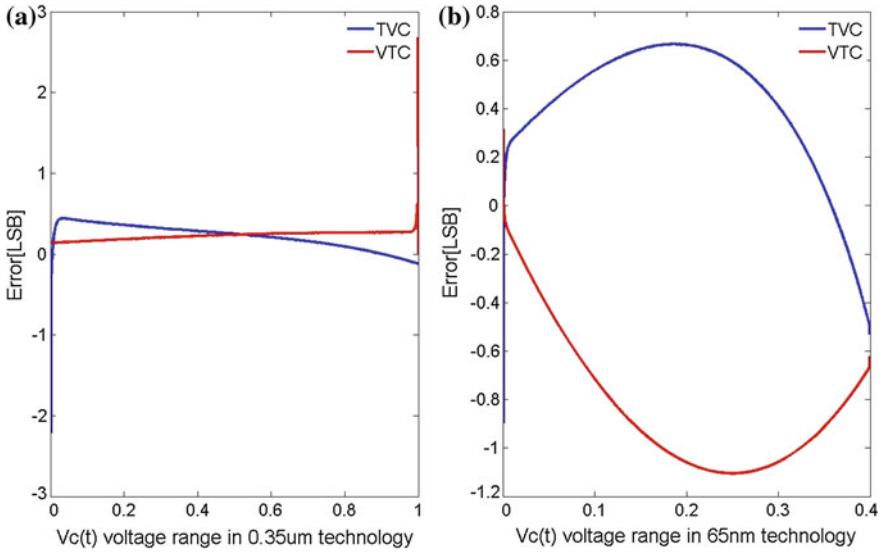


Fig. 35.4 The error between $V_C(t)$ of TVC and VTC circuits and their corresponding macro model results in **a** 0.35 μm technology and **b** 65 nm technology

The parasitic capacitor C_{pp} is also connected to node A and has the initial voltage of V_{DD}. When the CMOS switch is off, the charge of C_{pp} in the circuit of Fig. 35.1a remains at V_{DD} and does not change unless the switch is connected, thus in the model, switch S_i is placed in series with C_{pp} to keep its charge constant until CMOS switch (S_1 and S_i in the model) are connected. The CMOS switch is also replaced with a resistor R_{on} and ideal switch S_1. The values for r_{o2}, R_{on} and parasitic capacitors are set based on size and operation point of transistors in Fig. 35.1a. In the TVC model, the initial charge of capacitor C and C_p are set to be zero.

Similarly, the VTC circuit shown in Fig. 35.1b is modeled in Fig. 35.3b. The current mirror transistor M_4 is replaced by $I_{ref,}$ r_{o4} and V_{d3} which is the drain voltage of M_3. The parasitic capacitor C_{pn} with the initial voltage of zero is in series with a switch to prevent charging before the CMOS switch (S_1 and S_i in model) is connected. In the macro model the capacitor C and C_p have initial charge of V_0.

The accuracy of proposed macro models when modeling the behavior of the real circuit depends on the complexity of model and also the values that are selected for the model components. The accuracy evaluation for TVC and VTC models is shown in Fig. 35.4. In this case, both the real circuit and the model are simulated, and the voltage difference between the $V_C(t)$ in the circuit and $V_C(t)$ of the model is shown for each value of $V_C(t)$ in two different technologies. By considering a resolution of 8-bits and V_C voltage range from 0 to 1 V in 0.35 μm and 0 to 0.4 V in 65 nm technology, the LSB is equal to $1/2^8$ for the 0.35 μm technology and $0.4/2^8$ for the 65 nm technology. As shown in Fig. 35.4, for the TVC, the maximum error between the model and real circuit is less than 0.5LSB for the 0.35 μm technology and around 0.7 LSB for the 65 nm technology. The results of VTC simulations also show the maximum error of 0.5 LSB for the 0.35 μm technology and −1.1 LSB for the 65 nm technology. According to the simulation results, the proposed macro models show a reasonable level of accuracy in modeling the behavior of real circuit while maintain the simplicity in their structure.

35.4 Conclusions

In this work, we proposed a macro modeling approach for semi-digital smart integrated circuits. The key circuit blocks for the smart integrated circuit are the VTC and TVC, respectively. First, to understand the behavior and major error sources, transistor level VTC and TVC circuits are implemented with both CMOS 0.35 μm and CMOS 65 nm technologies. Furthermore, based on the circuit simulation results, the main error sources of VTC and TVC circuits are identified, and the macro model of VTC and TVC circuits are built. The macro model is built using simple circuit elements, that can accurately describe the behavior of the VTC and TVC. Finally, the accuracy of the macro models are verified by comparing the macro model outputs with the actual circuit outputs. Results show the proposed macro model can effectively model the behavior of the actual circuit. Overall, the proposed macro model can be effectively used for design optimization and calibration scheme design of semi-digital smart integrated circuits using VTC and TVC as the key building blocks.

References

1. Si W et al (2008) A single-chip CMOS Bluetooth v2.1 radio SoC. IEEE J Solid State Circ 43:2896–2904
2. Teo T et al (2010) A 700uW wireless sensor node SoC for continuous real-time health monitoring. IEEE J Solid State Circ 45:2292–2299
3. Su Y et al (2012) A 52 mW full HD 160 degree object view point recognition SoC with visual vocabulary processor for wearable vision applications. IEEE J Solid State Circ 47:797–809
4. Fiorenza J, Speke T, Holloway C, Lee H (2006) Comparator based switched-capacitor circuits for scaled CMOS technologies. IEEE J Solid State Circ 41:2658–2668
5. Brooks L, Lee H (2007) A zero crossing based 8-bit 200 MS/s pipeline ADC. IEEE J Solid State Circ 42:2677–2687
6. Speke T, Holloway P, Sodini C, Lee H (2009) Noise analysis for comparator based circuits. IEEE Trans Circuits Syst I 56:541–553
7. Wulff C, Ytterdal T (2009) CBSC pipeline ADC with comparator preset and comparator delay compensation. IEEE NORCHIP 2009:1–4
8. Zhang G, Lee K (2012) SAR ADC using single-capacitor pulse width to analog converter based DAC. In: IEEE international symposium on circuits systems, pp 2365–2368
9. Hoseini Z, Lee K (2013) An 8-bit 500kS/s semi-digital cyclic ADC with time mode residue voltage generation. In: IEEE Midwest symposium on circuits and system, pp 832–835

Chapter 36
Hybrid Performance with Pixel Values' Transition and Curve Fitting for Improved Stereo Matching

Kwangmu Shin, Daekeun Kim and Kidong Chung

Abstract We propose method that it can get disparity values uniformly even though the targeted stereo images is distorted by particular effects of illumination. The proposed method has key feature. This doesn't perform additional process as like preprocessing. That is, this has robust performance about distortion occurred through external illumination and carmera parameters. As a result, we could derive the maximum hybrid performance. The verification was showed by the good subjective and objective evaluations result.

36.1 Introduction

The stereo matching makes people cubic effects that it uses binocular disparity of human [1–5]. So the key point of stereo matching is to extract disparity information of images. In details, there are relatively different movement differentials between background and foreground objects in stereo images. We can exploit this point enough. That is, the background with long distant from camera view point gets small disparity value as the movement differential is short. On the other hand, the foreground objects with short distant from camera view point gets relatively large disparity value. If we get somewhat precise disparity value, also we can calculate depth information through evenly gray-level scaling.

K. Shin (✉) · D. Kim · K. Chung
Department of Computer Engineering, Pusan National University, Busan, Republic of Korea
e-mail: sin@pusan.ac.kr

D. Kim
e-mail: koftek@melon.cs.pusan.ac.kr

K. Chung
e-mail: kdchungm@melon.cs.pusan.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_36, © Springer Science+Business Media Dordrecht 2014

The acquisition of stereo images put to use particular stereo camera. Whereas it can be to happens external effects as like illumination. And then it is considerably hard to acquire precise disparity values. Ordinarily the illumination distortion is occurred by absolute illumination and camera parameter values' difference.

In this paper, we have emphasis on the negative effect of absolute illumination difference. We propose method that it can get disparity values uniformly even though the targeted stereo images is distorted by particular effects of illumination. The proposed method has key feature. This doesn't perform additional process as like preprocessing. That is, this has robust performance about distortion occurred through external illumination and camera parameters. More in details, firstly, we exploit delicately pixel values' transition in particular block-based window. In other words, we could call as the observation of subtle changes. Next, we exploit curve fitting in particular block-based window for more improved stereo matching result. Again, we could call as the observation of coarse changes. As a result, we could derive the maximum hybrid performance.

This paper are organized as follows. Section 36.2 reviews the prominent stereo matching introduced previously. STEP I: Getting of Pixel Values' Transition State and STEP II: Adoption of Curve Fitting for More Improved Performance are presented in Sect. 36.3. Experiments and performance results are presented in Sect. 36.4. Finally, concluding remarks and future works were descripted in Sect. 36.5.

36.2 The Prominent Stereo Matching introduced Previously

36.2.1 Adaptive Support Weight (ASW) [6]

The ASW exploits weight values which uses color and distance information. And then this finds responding area in stereo matching. This method has roughly higher performance than SAD, SSD etc. But this method tends to centralize color's similarity. Consequently, it had weakness when the distortion is occurred with external factors as illumination related with color and camera parameters. That is, we couldn't acquire more precise disparity values. Eq. (36.1) is to represent ASW method with formal expression. We hope that it is to be referred for understanding.

$$w(p,q) = \exp\left(-\left(\frac{\Delta C_{pq}}{\Upsilon_c} + \frac{\Delta g_{pq}}{\Upsilon_p}\right)\right) \tag{36.1}$$

36.2.2 Adaptive Normalized Cross Correlation (ANCC) [7, 8]

The ANCC performs preprocessing as like Gamma value correction, normalization of Log-chromacity Color. The processes as like these need to minimize

unnecessary illumination's effects. Consequentially, this preprocessing causes precise stereo matching results. The method exploits basically weight values which uses color and distance information. In this aspects, we could regard that this method is similar to ASW.

$$ANCC_{logChrom_R}(f_p) = \frac{\sum_{i=1}^{M} w_L(t_i) w_R(t_i) \left[R_L'''(t_i)\right] \times \left[R_R'''(t_i)\right]}{\sqrt{\sum_{i=1}^{M} |w_L(t_i) R_L'''(t_i)|^2} \times \sqrt{\sum_{i=1}^{M} |w_R(t_i) R_R'''(t_i)|^2}} \quad (36.2)$$

Equation (36.2) is to represent ANCC method with formal expression. This method could decide the responding region with mechanism of this expression and support to select the responding region with help of weight. R_L''' and R_R''' are reprocessed pixels information of both windows for protection of fattening effects the preservation of edges. Through above processes, the ANCC method could extract uniform disparity values in redundant illumination's changes.

36.3 The Hybrid Method with Pixel Values' Transition and Curve Fitting

36.3.1 STEP I: Getting of Pixel Values' Transition State

Most images have analogous color information in neighborhood pixels. It is same despite of being effected by overall illumination. So, the difference between particular pixel and neighborhood pixel values would be analogous if it is responding region. Our method start at simple and intuitive idea. Figure 36.1 depicts STEP I: Getting of pixel values' transition state. We hope to refer Fig. 36.1 about specific processes.

Figure 36.2 depicts getting algorithm of pixel values' transition state with simple expression. Particularly, in this figure, the left image shows existing window and the right image shows window moved to arrow direction. That is, this figure shows whole flow of method performance.

In Fig. 36.2, the sum of difference value between A−B and A′−B′ and difference value between C−D and C′−D′ is calculated in direction of left and right. The sum of difference value between A−C and A′−C′ and difference value between B−D and B′−D′ is calculated in direction of up and down.

$$\begin{aligned} &\sum_{(i,j)\in W} (|I_1(i,j) - I_1(i,j+1)| - |I_2(i,j) - I_2(i,j+1)|) \\ &+ \sum_{(i,j)\in W} (|I_1(i,j) - I_1(i+1,j)| - |I_2(i,j) - I_2(i+1,j)|) \end{aligned} \quad (36.3)$$

Equation (36.3) expresses STEP I with the formal expression. In this equation, i and j mean the coordinate of pixel, w means window and I_1 means left image and

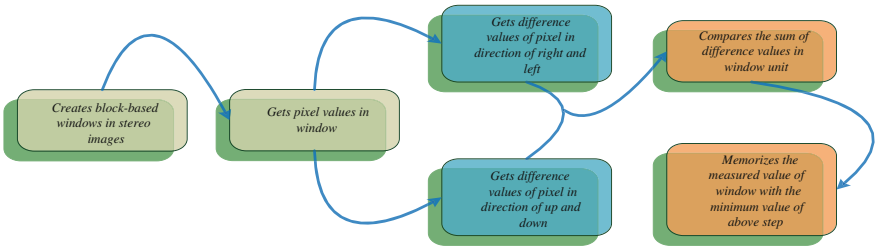
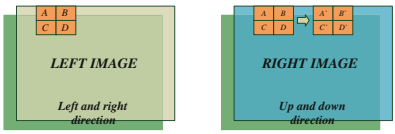


Fig. 36.1 Getting processes of pixel values' transition state



$$\min[(|A-B|-|A'-B'|)+(|C-D|-|C'-D'|)+(|A-C|-|A'-C'|)+(|B-D|-|B'-D'|)]$$

Fig. 36.2 Getting algorithm of pixel values' transition state with simple expression

I_2 means right image. Whereas, occasionally we couldn't find out precise disparity values in edge regions. In that case, it might be excess of native maximum disparity values or 0's disparity value. This is because of using only neighborhood pixel information while it doesn't includes comprehensive changes in block-based window. We adopted additionally curve fitting method for reflecting overall change of window.

36.3.2 STEP II: Adoption of Curve Fitting for More Improved Performance

The curve fitting finds out the most adjacent curve using given coordinates' information. It is represented by formal polynomials.

Figure 36.3 depicts STEP II: Adoption of curve fitting for more improved performance. We hope to refer Fig. 36.3 about specific processes.

If the curve fitting is performed in the region of left window, particular polynomial expression is created. We store all the coefficients. Next, we set window in right image based window's position in left image. And then we move as search range and perform curve fitting about right window. The polynomial expression obtained as like this is comparable polynomial expression of left window. We could get the difference of coefficients between the most degrees. And we decide as responding region that it has the most minimum difference of relevant coefficient. As a result, this method could show the overall pixels change of in-window.

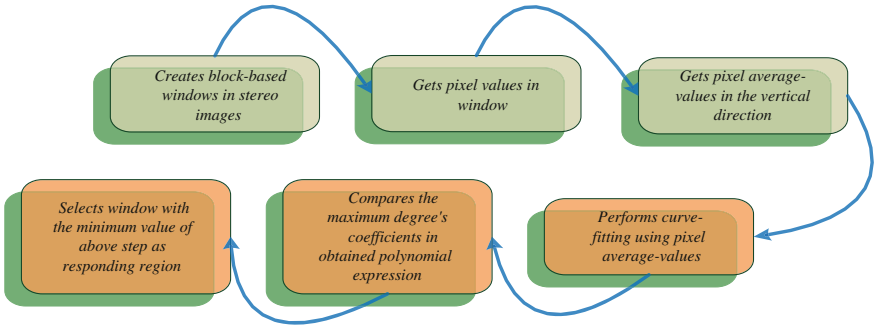


Fig. 36.3 Adoption processes of curve fitting

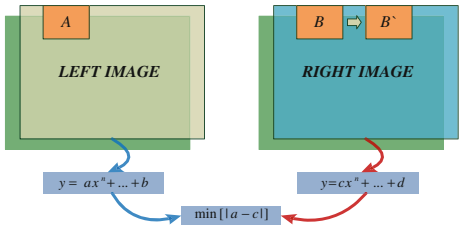


Fig. 36.4 Adoption algorithm of curve fitting

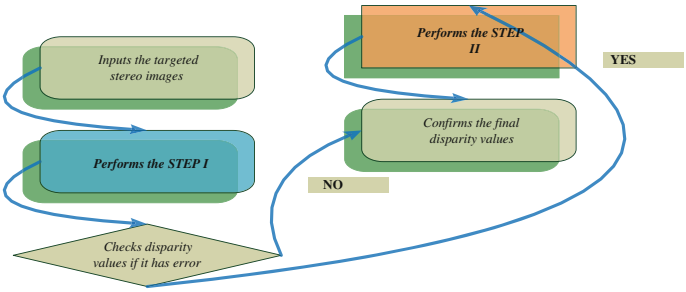


Fig. 36.5 The hybrid method with pixel values' transition and curve fitting

Therefore we could select the more precise responding region. Figure 36.4 depicts simply adoption algorithm of curve fitting.

36.3.3 The Final Algorithm of Proposed Method

Figure 36.5 depicts the hybrid method with pixel values' transition and curve fitting. That is, this figure shows the final algorithm of proposed method.

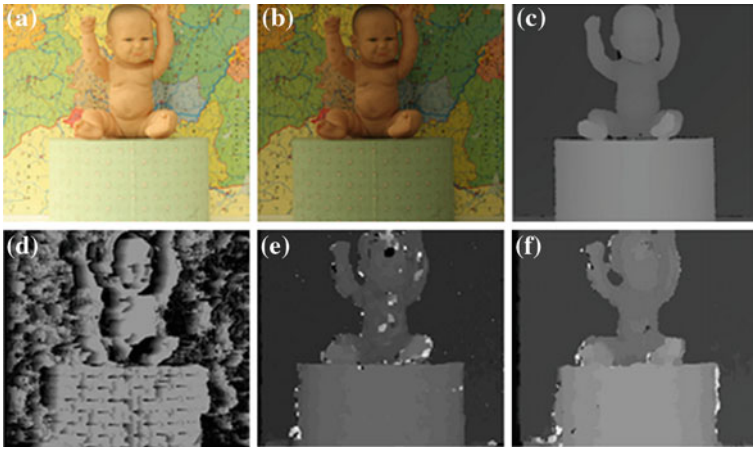


Fig. 36.6 The Baby1 in stereo images

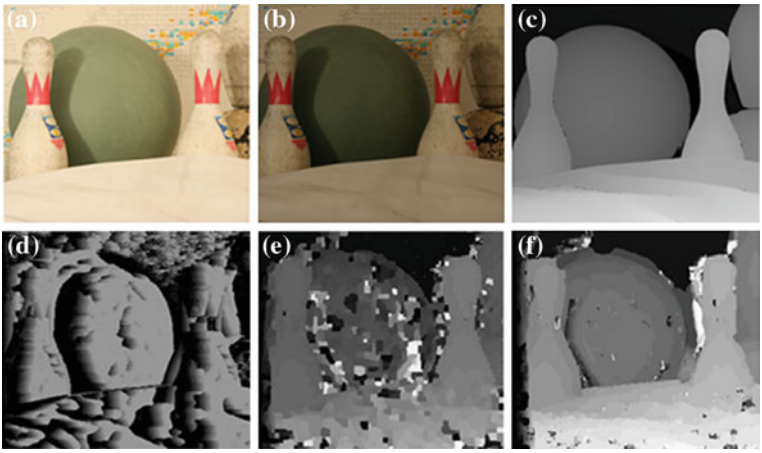


Fig. 36.7 The Bowling1 in stereo images

36.4 The Experiments and Performances

36.4.1 The Experiments Environment

The window size is 11 × 11 in pixel unit. The experiments images' source is Middlebury Studio [9]. The comparison group includes ASW and ANCC and targetted testing stereo images are Baby1, Bowling1, Flowerpots. We performed the subjective and the objective evaluations to increase experiments' reliability.

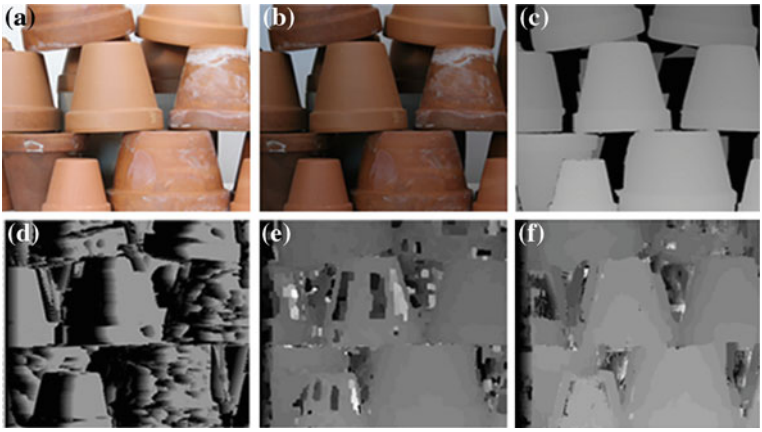


Fig. 36.8 The Flowerpots in stereo images

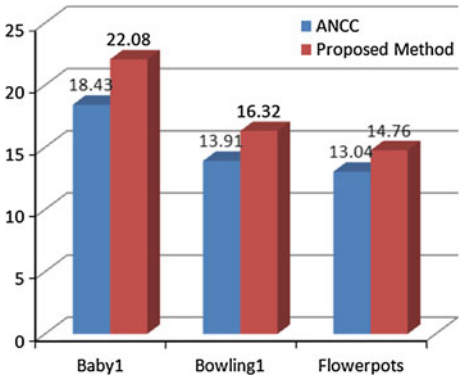


Fig. 36.9 The PSNR numerical values' comparison with Baby1, Bowling1, Flowerpots

36.4.2 The Subjective Evaluations

Figure 36.6 describes the Baby1 in stereo images with different illumination values. As you see, we could verify that the proposed method has more clear, precise disparity image compared with ASW, ANCC. Figure 36.7 describes the Bowling1 in stereo images with different illumination values. As you see, we could verify that the proposed method has more clear, precise disparity image compared with ASW, ANCC. Figure 36.8 describes the Flowerpots in stereo images with different illumination values. As you see, we could verify that the proposed method has more clear, precise disparity image compared with ASW, ANCC.

36.4.3 The Objective Evaluations

Figure 36.9 depicts the PSNR numerical values' comparison with Baby1, Bowling1, Flowerpots. As you see, the proposed method precedes in aspect of PSNR numerical values. ANCC method's PSNR average value is 15.33 and the proposed method is 17.72. Consequently, the method of this paper is about 17 % higher than ANCC method.

36.5 Conclusions and Future Works

We propose method that it can get disparity values uniformly even though the targeted stereo images is distorted by particular effects of illumination. The proposed method has key feature. This doesn't perform additional process as like preprocessing. That is, this has robust performance about distortion occurred through external illumination and camera parameters. As a result, we could derive the maximum hybrid performance. In experiments performed previously, we could confirm through relatively superior result compared with typical local stereo matching. Moreover, the experiment methods include the subjective and objective evaluation comprehensively. But we think that the experiments testing quantity is insufficient. We are ongoing additional verification as future works.

References

1. Hu, X, Zhang C, Wang W, Gao X (2010) Disparity adjustment for local stereo matching. In: IEEE conference on computer and information technology (CIT), Sept 2010
2. Iqbal M, Morel O, Meriaudeau F (2010) Choosing local matching score method for stereo matching based-on polarization imaging. IEEE Conf Comput Autom Eng (CAE) 2:334–338
3. Yang Q, Wang L, Ahuja N (2010) A constant-space belief propagation algorithm for stereo matching. In: IEEE conference on computer vision and pattern recognition (CVPR), June 2010, pp 1458–1465
4. Zhao M, Li X, Lang R (2010) Improved adaptive stereo matching using double dynamic programming with correlation of row and column. In: 2nd International conference on signal processing systems (ICSPS), July 2010, pp 95–100
5. Na I, Choi J, Jeong H (2009) Robust fast belief propagation for real-time stereo matching. IEEE Conf Adv Commun Technol (ACT) 2:1175–1179
6. Yoon K-J, Kweon IS (2006) Adaptive support-weight approach for correspondence search. IEEE Trans Pattern Anal Mach Intell (PAMI) 28(4):650–656
7. Heo YS, Lee KM, Lee SU (2011) Robust stereo matching using adaptive normalized cross-correlation. IEEE Trans Pattern Anal Mach Intell (PAMI) 33(4):807–822
8. Heo YS, Lee KM, Lee SU (2008) Illumination and camera invariant stereo matching. In: IEEE conference on computer vision and pattern recognition (CVPR), June 2008
9. http://vision.middlebury.edu/stereo/. 2011

Chapter 37
ActiveSTB: Scalable Scheme Over Wireless Networks

Heung Ki Lee, Kyung Jin Ahn, Jaehee Jung and Gangman Yi

Abstract Quality of streaming service is determined by bandwidth constriction and error rate of link in wireless environments. The layered video coding provides attraction for streaming service at wireless environments through adaptation of streaming size. However, these dependency also increases the possibility that streaming data are thrown out even though they are transmitted successfully. It wastes a lot of link bandwidth and drops the efficiency of transmission. To understand the structure of layered streams increases the performance of layered video streaming. In this paper, we proposed the active streaming schemes of layered video streaming depending on the available bandwidth. The intermediate nodes figured out the structure of layered video streaming and then change the format of scalable streaming scheme. The complicated relationship between layered streams increases the availability of layered stream, but also risks to waste the transmitted layered streams. Additionally, we added the comparison between simple referred layered streams and complicated layered streams.

Keywords Scalable stream · Layered stream · Wireless connection

H. K. Lee · K. J. Ahn · J. Jung
Samsung Electronic Co., Suwon, South Korea
e-mail: heungkilee@gamil.com

K. J. Ahn
e-mail: kjinni@gmail.com

J. Jung
e-mail: cleme76@gmail.com

G. Yi (✉)
Gangneung-Wonju National University, Gangwon, South Korea
e-mail: gangman@cs.gwnu.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_37, © Springer Science+Business Media Dordrecht 2014

37.1 Introduction

In the Internet environment, the volume of video streaming traffic has been increased dramatically. Current networks cannot support the such growth of streaming traffic. Especially, the mobile access network is critical to decide the quality of service to provide end users. In mobile networks, the several users share the scarce wireless resources for streaming application. Such competitions decrease the quality of streaming service. First, when media data are not transmitted on time or lost, the quality of streaming is reduced. Also, the lost or delayed media data may lead to error propagation to the following data until some error protected frame. Second, the nature of wireless connection cannot ensure the sufficient and consistent bandwidth for streaming service. Therefore, adaptive video streaming schemes can give the beneficial to streaming service through Internet.

Adaptive media streaming are classified into two groups including layered streaming scheme and adaptive streaming scheme. In adaptive streaming scheme, we should contain redundant media into the storage. It wastes the storage, and creates the cost to maintain the massive storage. For example, the huge power of data center is required. The layered streaming scheme increases the efficiency of the strategy in containing media streams. Layered streaming data are exclusive to each others. Also each layers improves the quality of streaming data.

However, in layered scheme, the hierarchical relationship between layers decreases the efficiency in streaming service. When referred layers are not transmitted successfully, referring layers not used for decoding even though the referring layers are transmitted. So, it is critical to understand the relationship between layers in layered streaming service.

37.2 Related Works

37.2.1 Background

Wireless connection is not stable and guaranteed the bandwidth as wired connection. Usually, streaming data are transmitted through wired connection and wireless connection in a row. In wired connection, there are not happened as losing and delaying multimedia data. But, wireless connection is error-prone network, so data transmission is not guaranteed. So, streaming service through wireless connection required the scheme to protect media data.

Scalable streaming service is well designed for streaming data through wireless connection. In scalable streaming, one original stream is divided into several streams. It is good to transmit media data through congested network through providing only the basic layer to be played. However, the relationship between layers increases the possibilities to discard the transmitted layers, even though they

are transmitted successfully. When the referred layers are not transmitted, the referring layers cannot be decoded.

In this paper, we introduce an *ActiveSTB* that is located between wired connection and wireless connection. *ActiveSTB* receives the media data from data source, then selects media data to be transmitted. Also, *ActiveSTB* transformed media data to enhance the adaptation of wireless connection.

37.2.2 Previous Works

Intermediate Node for Streaming Video There are many studies for active intermediate nodes for enhancing service quality of streaming service [1–4]. Intermediate node decreases the quality of cached stream through selecting user's request to the low quality multimedia stream for improving the hit rate of cached streams [1, 5].

References [2, 6] discards the multimedia packets with large jitter at cache storage when they detects the large jitter. The intermediate node improves the quality of multimedia streaming scheme through retransmission based on presentation time [3]. When clients are joined the ongoing broadcasting, intermediate node provide the missing part to clients [4]. To the best of our knowledge, converting scalable streams based on relationship between layers, have not been studied.

An Set Top Box (STB) is one of intermediate node through streaming service. There are four components including Media Codec, Network Interface, Presentation Module and Graphic Module [7]. IPTV STBs requires additional functions that includes personal video recorders, digital media adapters, and more [8]. However, we should understand the relationship between layers in scalable streams to enhance the quality of streaming service. Telchemy provides the intermediate node for detecting packet loss and jitter [9].

37.3 Active Set Top Box

37.3.1 MP4 Scalable Streaming Versus H.264 Scalable Streaming

When H.264 standard provides the scalable streaming scheme, they provide the complicated hierarchical relationship. This improves the structure of layered streams. But, it will also increase the relationship between layers. In H.264 scalable streaming, even B frames has the hierarchical relationship between layers. During streaming service, the loss of one B frame drops other related B frames. In the figure, when B_3 is lost during the scalable streaming, H.264 streaming also

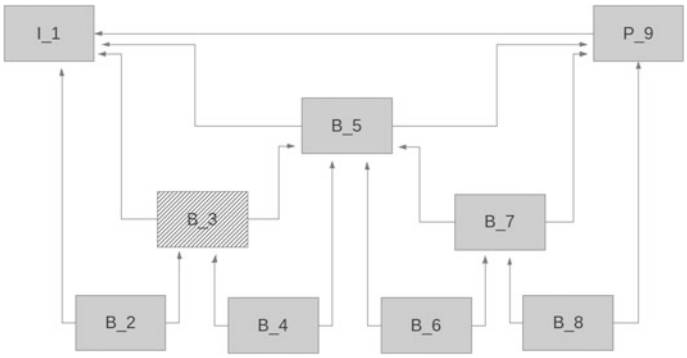


Fig. 37.1 H.264 scalable streaming scheme

didn't use the B_2 frames and B_4 frames, even though they are transmitted successfully. However, MP4 scalable streaming B_2 and B_4 can be decoded without B_3 frames (Fig. 37.1).

37.3.2 ActiveSTB

ActiveSTB is located between wired network and wireless network. It is gateway between two different networks. Scalable stream in wired connection is transmitted safely. But, transmission over wireless connection is not guaranteed. Therefore, *ActiveSTB* converts the received layers into the forwarded layers to wireless connection.

ActiveSTB changes the layered streams to decrease the dependency between layer streams. During the H.264 streaming service, *ActiveSTB* changes the H.264 scalable streaming into MP4 scalable streaming. MP4 scalable streaming has the simple relationship between scalable streams, so it decreases the possibility of indirect loss through relationship between layer streams.

37.4 Experimental Result

Through NS-2 simulator, we evaluate that our *ActiveSTB* outperforms other scalable streams. As shown in Fig. 37.2, there are three wired nodes, three wireless nodes and AP for simulation. Wired nodes and AP are connected through 100 Mbps wire link, while wireless nodes are connected through 10 Mbps wireless link.

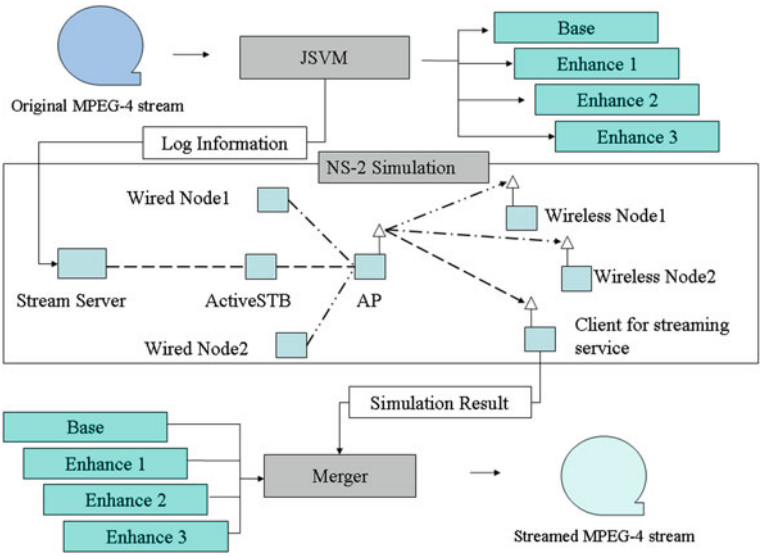


Fig. 37.2 Simulation environment

Table 37.1 Scalable Stream

	QP	Frame rates	Frame size
Layer 0	38.0	15	320 × 240
Layer 1	32.0	30	320 × 240
Layer 2	30.0	30	320 × 240
Layer 3	28.0	30	640 × 480
Layer 4	26.0	30	640 × 480

37.4.1 Scalable Multimedia

In the simulation, we creates two streams for proving the efficiency of scalable streaming scheme. Each stream is divided into six sub streams using Joint Video Model (JSVM) codecs [10]. We follow the below configuration of JSVM for creating the sub layers from each stream (Table 37.1).

We divide one stream into five scalable layers streams. Base stream contains 320 × 240 frames with 15 frame per seconds (FPS) and 38 quantization value. It can enhance the quality of stream up to 640 × 480 frames with 30 frame per seconds and 26 quantization value (Table 37.2).

Table 37.2 Scalable streams

Frame name	Number of frames
Bourne Altimatum	2,125
I am legend	2,397

37.4.2 Simulation Environments

Figure 37.2 shows that AP provide the three connections between three wired nodes and three wireless nodes. The Wired Node 1 is connected to Wireless Node 2, and the Wired Node 2 is connected to Wireless Node 1. Streaming Server is connected to Client for streaming service. The Wired Nodes create the cross traffic through wireless connection. We evaluate the estimation scheme including *MP4 Scalable Streaming* and *H.264 Scalable Streaming*.

We make three steps in the simulation for scalable streaming service through JSVM, merger and NS-2 simulator. In first step, JSVM creates several layers from original stream, and then collects the information of each frame in scalable streams. At second step, we conduct the simulation of the streaming environment based on the information from step 1. At last step, our merger decides whether packets in the simulation can be decoded through comparing the arrival time of packets with decoding time of media data in each packet. Multimedia server sends the multimedia data into *ActiveSTB*, and then *ActiveSTB* selects the forwarding media data to client.

37.5 Simulation Result

Figure 37.3 shows the indirect loss that cannot be decoded, even though media data are transmitted successfully. It is happened because of the dependency between layers in scalable streams. For example, B frames are required the reference frame including I or P frame. They are trashed out if the reference frames are not available. This indirect loss wastes the bandwidth and streaming service. In Fig. 37.3 x axis shows the error rate of wireless link and y axis shows indirect loss of the transmitted media data. In two cases, *MP4 Scalable Streaming* service is better than *H.264 Scalable Streaming* service. The complexity of *MP4 Scalable Streaming* is simple, so indirect loss is not happened frequently.

Figure 37.4 shows the ratio between transmitted data from an *ActiveSTB* and available media data at client. In Fig. 37.4, X-axis shows the error rate, while y-axis shows for ratio between transmitted data and available data. The ratio stands for the performance of scalable streaming scheme. The high ratio shows the good efficient management of scalable streaming service over wireless network. While the low ratio is for bad efficient management of scalable scheme. When the error

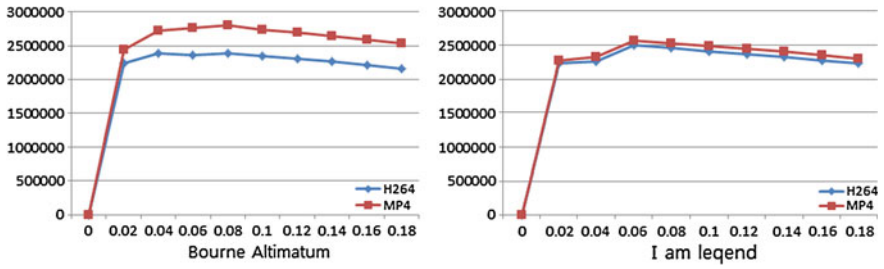


Fig. 37.3 Indirect loss

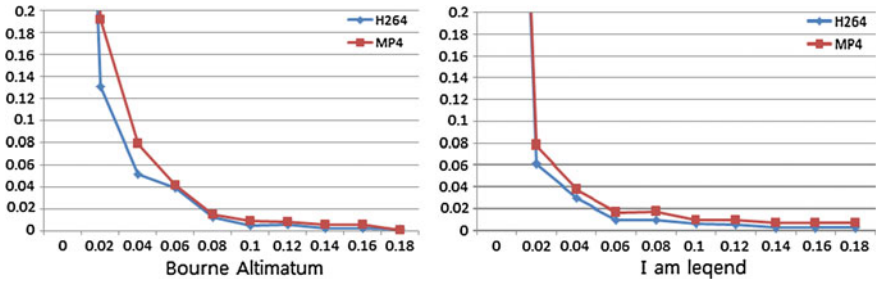


Fig. 37.4 Ratio between sent and decoded

rate is zero, *ActiveSTB* can send whole scalable streaming media data into client. When the error rate is increased, *MP4 Scalable Streaming* scheme outperforms *H.264 Scalable Streaming* scheme.

37.6 Conclusion

In this paper, *ActiveSTB* is suggested as enhancing scalable streaming scheme in Wireless Networks through converting *H.264 Scalable Streaming* data into *MP4 Scalable Streaming* data. *ActiveSTB* releases the overhead of relationship between layers in scalable stream during the scalable streaming in wireless network. Therefore, *ActiveSTB* decreases the indirect loss over wireless networks in order to improves the quality of scalable streams. We prove the performance of our scheme through real streams and codec. Future work includes searching the simulation in different environments with different link speed and various stream data.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2063006).

References

1. Shen B, Lee S, Basu S (2004) Caching strategies in transcoding-enabled proxy systems for streaming media distribution networks. In: Proceedings of IEEE transactions on multimedia, vol 6, Special Issue on Streaming Media
2. Chan S, Kok CW, Wong AK (2003) Multimedia streaming gateway with jitter detection. In: Proceedings of international conference on communication, 2003
3. Huang THC, Chang C (2002) A proxy-based adaptive flow control scheme for media streaming. In: Proceedings of the 2002 ACM symposium on applied computing, pp 750–754
4. Gae L, Zhang Z, Towsley DF (2003) Proxy-assisted techniques for delivering continuous multimedia streams. Proc IEEE/ACM Trans Netw 11:884–894
5. Schojer P, Boszormenyi L, Hellwagner H, Penx B, Podlipnig S (2003) Architecture of a quality based intelligent proxy (qbix) for mpeg-4 videos. In: Proceedings of world wide web, 2003
6. Chan S, Kok CW, Wong AK (2005) Multimedia streaming gateway with jitter detection. Proc IEEE Trans Multimedia 7:585–592
7. Laursen A, Olkin J, Porter M (1994) Oracle media server: providing consumer based interactive access to multimedia data. In: Proceedings of the ACM SIG on management of data, vol 23, p 47077
8. Robertson S, Rivin R, Analog devices: designing IPTV set-top boxes without getting boxed in. Analog Devices Inc
9. Telchemy Inc., Analog devices: designing IPTV set-top boxes without getting boxed in. Telchemy Inc
10. I. J. S. 11 and I.-T. S. Q.6, Joint scalable video model (JSVM) software manual. Included in JVT-X203. ISO/IEC JTC 1/SC29/WG 11 and ITU-T SG16 Q.6., 2007

Chapter 38
Rhythm-Based Authentication Scheme on Smart Devices

Jae Dong Lee, Won Min Kang, Chang Hyun Moon, Hwa-Young Jeong, Young-Sik Jeong, Jai-Jin Jung and Jong Hyuk Park

Abstract Due to computing power and advances of smart devices, recently uses of smart-phone have been rapidly increasing. Smart devices (smartphones and tablet PC etc.) threats such as a personal information leakage and fraudulent use of smart devices have also increased. So, effective measures are needed to address security issues of smart devices—authentication, access control, prevention from the attackers. In this paper, we proposed a rhythm-based user authentication scheme on smart devices. The proposed scheme complements the vulnerabilities of the existing schemes, and provides authentication that ensures availability. In additional, it provides stronger security and greater convenience.

J. D. Lee · W. M. Kang · C. H. Moon · J. H. Park (✉)
Seoul National University of Science and Technology, 172 Gongreung 2-dong, Nowon-gu, Seoul 139-743, South Korea
e-mail: jhpark1@seoultech.ac.kr

J. D. Lee
e-mail: jdlee731@seoultech.ac.kr

W. M. Kang
e-mail: wkaqhdsk0@seoultech.ac.kr

C. H. Moon
e-mail: ckdgus2482@seoultech.ac.kr

H.-Y. Jeong
Humanitas College, Kyung Hee University, 1 Hoegi-dong, Dongdaemun-gu, Seoul, South Korea
e-mail: hyjeong@khu.ac.kr

Y.-S. Jeong
Department of Multimedia Engineering, Dongguk University, 30, Pildong-ro 1-gil, Jung-gu, Seoul, South Korea
e-mail: ysjeong@dongguk.edu

J.-J. Jung
Department of Multimedia Engineering, Dankook University, 152, Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do 448-701, South Korea
e-mail: dothan@dankook.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_38, © Springer Science+Business Media Dordrecht 2014

Keywords User authentication · Rhythm · Smart devices · Tablet PC · Android phone

38.1 Introduction

Recently, the distribution of convergent devices has been made prevalent by the lead taken by smart devices such as smartphones, tablet PCs and mobile screen devices in the modern IT market. These smart devices, store basic personal information, such as that found in call lists, contact lists, and address books. In additional, services on the web, SNSs, mobile applications, account synchronization, PC synchronization, and banking are being provided, which require the storage of account information and authentication certificates, which further increases the risk of data leakage. Therefore, smart device authentication systems require ample security [1, 2].

There are a variety of security authentication methods used in systems and applications: biometric-based identification (something you are) authenticates by recognizing a user's biometric information. Knowledge-based identification is that in which users are recognized by their knowledge. In token-based identification (something you have), users are identified by an object in their possession [3].

Android systems provide pattern lock, face recognition, password, and PIN methods for user authentication. However, these basic authentication methods lack availability and do not provide sufficient confidentiality. Therefore, this paper explores the vulnerabilities and limitations of the existing authentication methods, and proposes a new method that compensates for these limitations.

This paper is organized as follows. In Sect. 38.2, we discuss related works including the relevant security threats and analyses of authentication schemes. We explain the detailed proposed scheme: rhythm-based authentication scheme for android phone in Sect. 38.3. Finally, the conclusion should be provided in Sect. 38.4.

38.2 Related Works

In this section, we discuss user authentication schemes and their security threats, and analyses of authentication scheme for android phone.

38.2.1 Security Threats in User Authentication in Smart Devices

Possible security threats in user authentication in smart devices include shoulder-surfing attacks, smudge attacks, and brute-force attacks. A brute-force attack is

when all possible values are tried to decode a specific password. A shoulder-surfing attack involves creeping over another person's shoulder to obtain a password, security code, or PIN. Finally, smudge attack refers to the lock pattern identification method used on touch screen devices, such as tablet PCs [4–7].

38.2.2 Smart Phone User Authentication Methods

The following paragraph addresses the user authentication methods used on Android phones.

- Pattern Lock: Pattern lock depends on users to remember the drag pattern created by nine points for authentication. This is a knowledge-based as well as a behavior-based authentication, as it depends upon the kinesthetic memory of the user's fingers. The maximum number of possible key combinations is 388,912, which is rather small. Furthermore, since users tend to use less points for convenience, this method is vulnerable to brute-force attacks. Due to high visual memory capacity and static authentication keys, it is also vulnerable to shoulder-surfing attacks. Lastly, it is vulnerable to smudge attacks, because patterns can be inferred by using traces of fingerprints on the screen, using drag patterns as the key [4–6].
- Face recognition: Face recognition is a biometric-based authentication, in which the user's face is recognized. Using an individual's own biological characteristics, it has the advantage of not having to separately rely upon any key or pattern, but due to technical limitations, misidentifications are possible. In addition, it is possible to complete authentication using a photograph or video, exposing the same kind of limitations as token-based authentication [4, 6, 7].
- Password: This method uses a virtual keypad in a touch screen that looks similar to those in mobile or PC environments, and takes advantage of sufficient key space using the alphabet, numbers, and special characters. However, using a virtual keypad in touch screen devices is unnatural and lacks availability because the keys are more complicated. Since the keys that users easily remember are often composed of users' personal information, passwords are also vulnerable to the attacks mentioned previously [4–6].
- PIN: In order to compensate for the lack of availability of password authentication methods, a PIN (using digits from 0 through 9) is used as a key. However, this method is vulnerable to brute-force and shoulder-surfing attacks due to limited key space, which allows only 10,000 combinations [4, 6, 7].

38.2.3 Analyses of User Authentication Methods for Smart Phones

Because the basic authentication method provided by Android operating systems lacks sufficient security, many studies have proposed substitutions. The main

system is a graphical password system, which is based on graphics and keystroke dynamics-based user authentication (KDA), which depends on the keystroke dynamics of a habitual rhythm when entering a PIN [4, 5]. In the case of a graphical password system, the user is not expected to use all the pixels of the image. Only those distinct parts of the image that the user can easily remember will be utilized, and this will not provide enough entropy [4, 6, 7]. The KDA method is limited in that it requires many trials at entering the PIN to clearly distinguish a naturally generated rhythm. In addition, the length of the PIN is only four letters, which exacerbates the problem of insufficient entropy. In order to compensate for such a limitation, the user must intentionally generate an artificial rhythm or a longer, augmented PIN. However, remembering a 10-character PIN and artificially creating an artificial rhythm decreases the method's availability [4, 6, 7].

38.3 Rhythm-Based Authentication Scheme

A new authentication scheme based on rhythm recognition is proposed in order to overcome the limitations of existing authentication schemes in this section. The proposed scheme is secure against brute-force attacks, shoulder-surfing attacks, and smudge attacks, and offers greater availability than the pattern lock scheme.

38.3.1 Key Generation Mechanism

The proposed authentication scheme uses touch rhythm as a key, depending on auditory memory and behavior memory. The input component and the recorded touch rhythm is defined as a track. The interface of each track is realized with one button, and through a touch listener and timer, the button input status per word is checked. It is then attached to the string and a single button for that track is generated. In order to make the key more complicated, multiple tracks are used to combine keys from single tracks, generating a multiple key.

If a, b, c, and d are the notes that compose the four single keys A, B, C, and D, then from the first index, a synthetic note is generated along the table, as shown in Fig. 38.1. These synthetic notes continue to string and are then combined to make a multiple key. Figure 38.2 represents a timing map.

38.3.2 Key Space

People are able to easily remember things related to music. Additionally, auditory memory has less availability than visual memory. Therefore, the proposed authentication scheme is advantageous in that the user easily remembers the key,

키조합	Union note
non	0
a	1
b	2
c	3
d	4
ab	5
ac	6
ad	7
bc	8
bd	9
cd	a
abc	b
abd	c
acd	d
bcd	e
abcd	f

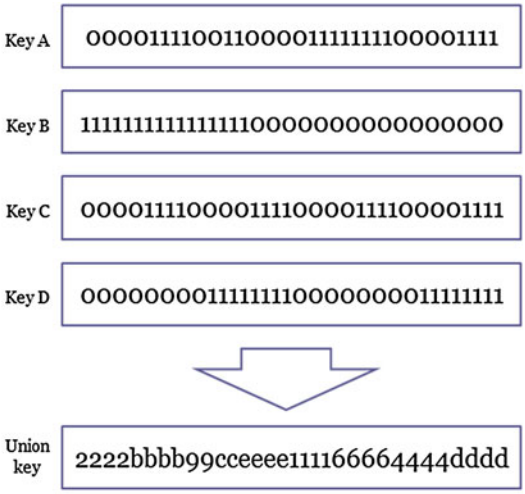


Fig. 38.1 Generation of a multiple key in four tracks

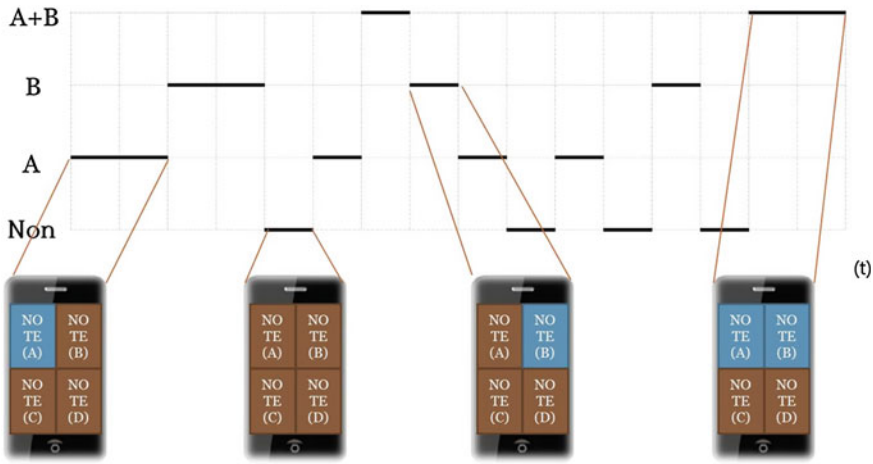


Fig. 38.2 Note A and B's key generation timing map

while it is impossible for others to memorize. For this reason, the scheme provides adequate user availability, and protection from shoulder-surfing attacks. It is safe from smudge attacks because it is impossible to infer the key with fingerprints. The key space is also very large, as shown below.

$$\sum_{k=1000a/m}^{1000b/m} 2^{nk} = \frac{2^{\frac{1000an}{m}}\left(2^{\frac{n(1000(b-a)+m)}{m}}-1\right)}{2^n-1}$$

The variables in the above equation indicate that the note check cycle is m (ms), the track number is n, and the allowed key entering time is greater than a second and less than b seconds. For example, m = 10, n = 4, a = 1, b = 3, and the key space is $2^{200}*(2^{201}-1)/(2^4-1)$, which is about 2^{397}. This is only a theoretical value, so it would be smaller in reality, considering the error, but nonetheless, adequate key space is provided considering the expected number of users.

38.3.3 Authentication Process

The authentication process progresses in the order of single key generation, multiple key generation, noise reduction, and key authentication. The single key generation module receives a touch value from each track, and ten touch values for a single key are generated in a single track. When a single key is generated in the module, the single key is combined with a synthetic note at an appropriate time. Based on these steps, a string is generated and then a multiple key. The created multiple keys incorporate noise, so a noise reduction module reduces each type of noise and generates a key. The final key is codified and then stored in the Smartphone, where it is compared with the user's entered value upon authentication request, and the appropriate responses are made, depending on the result of the comparison.

38.4 Conclusion

Recently, researches in user authentication scheme have been given a great deal of attention due to the increased possibility of personal information leakage through smart devices and users' interests in this matter. Smart devices using android operating systems have been increased and a variety of authentication methods have been proposed in order to strengthen the security offered by these devices.

In this paper, the vulnerabilities and limitations in the existing user authentication schemes employed in smart devices were analyzed. A rhythm-based authentication scheme was proposed. The proposed scheme improved the confidentiality and availability issues of the existing authentication schemes. In order to increase the accuracy of rhythm recognition, the False Acceptance Rate (FAR) and the False Rejection Rate (FRR) need to be considered in the design for such methods. Moreover, further studies need to be conducted on authentication by user location, situation, App, and role.

Acknowledgments This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the CPRC (Communications Policy Research Center) support program supervised by the KCA (Korea Communications Agency) (KCA-2013-003).

References

1. Jaegal B (2010) Trend of smartphone market and mobile OS. Samsung Wireless Division, 6.D
2. Lee A, Bok D (2008) Interface technology for improving production value. Samsung Economics Research Institute (SERI) Economic Focus, No. 216, Oct 2008
3. Mohammed MM, Elsadig M (2013) A multi-layer of multi factors authentication model for online banking services. In: ICCEEE 2013, IEEE, pp 220–224, Aug 2013
4. Android Security Overview, Android open source project, http://source.android.com/devices/tech/security/
5. Aviv AJ, Gibson K, Mossop E, Blaze M, Smith JM (2010) Smudge attacks on smartphone touch screens. In: 4th USENIX workshop on offensive technologies, Aug 2010
6. Lashkari AH, Farmand S, Zakaria OB, Saleh R (2009) Shoulder surfing attack in graphical password authentication. Int J Comput Sci Inf Secur 6(2):145–154
7. Yan J, Blackwell A, Anderson R, Grant A (2004) Password memorability and security: empirical results. Secur Privacy IEEE 2(5):25–31
8. Wiedenbeck S, Waters J, Birget JC, Brodskiy A, Memon N (2005) PassPoints: design and longitudinal evaluation of a graphical password system. Int J Human-Comput Stud 63(1–2):102–127
9. Kang P, Park S, Hwang S, Lee H, Cho S (2008) Improvement of keystroke data quality through artificial rhythms and cues. Comput Secur 27(1–2):3–11

Chapter 39
Cloud Services Aided e-Tourism: In the Case of Low-Cost Airlines for Backpacking

Qingguo Zhou, Jason C. Hung, Jun Hu, Huaming Chen, Rui Zhou, Ji Qi, Lei Yang and Xuan Wang

Abstract The emergence of Cloud Services and Mobile Internet has influenced the society a lot, including the tourism industry. This paper proposes a design of backpacking service, not only aiming at the travelling routines, but also focusing on the low-cost airlines. This kind of service aids backpackers with an effective travelling and satisfies the price requirement. With this low-cost airline system, the backpackers can experience real e-Tourism and enjoy a better travel aided by the

Q. Zhou · J. Hu · H. Chen · R. Zhou (✉)
School of Information Science and Engineering, Lanzhou University, Lanzhou, China
e-mail: zr@lzu.edu.cn

Q. Zhou
e-mail: zhouqg@lzu.edu.cn

J. Hu
e-mail: hj7345506@sina.com

H. Chen
e-mail: chenhm12@lzu.edu.cn

J. C. Hung
Department of Information Management, Overseas Chinese University, Taichung, Taiwan
e-mail: jhungc.hung@gmail.com

J. Qi · L. Yang
ADS Spallation Target Laboratory, Institute of Modern Physics, Chinese Academy of Science, Beijing, China
e-mail: qijiin21c@impcas.ac.cn

L. Yang
e-mail: lyang@impcas.ac.cn

X. Wang
School of Science, Lanzhou University of Technology, Lanzhou, China
e-mail: wangxuan2010@lut.cn

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_39, © Springer Science+Business Media Dordrecht 2014

web information. According to the basic principles of low-cost airlines, this paper provides the backpackers with a definitely efficient way to check out a suitable flight itinerary under an acceptable price.

Keywords e-Tourism · Backpacking · Cloud services · Mobile internet

39.1 Introduction

The emergence of Cloud Services and Mobile Internet helps to support all the activities such as economy, culture etc. indispensable for information, and has influenced the society a lot, including the tourism industry. As an information intensive industry, tourism is being transferred from a traditional industry into synchronizing with current information society. It can be referred to as e-Tourism. E-Tourism is a classic business field that can best describe the future of cloud service usage [1]. The informationized environment has enhanced tourism services such as travel bookings, itinerary planning, destination marketing, and information sharing. Originally, information technologies help tourism in Computer Reservations Systems (CRS) or Global Distribution System (GDS) of airlines, incorporating the broader tourism sector. Information technologies are becoming a key determinant of organizational competitiveness and a wide range of technological developments propelling tourism [2].

In tourism industry, there is a form of travelling, usually called backpacking, widely accepted and recommended by a large number of tourists, particularly by those with low budgets. They are expense sensitive and always try to achieve high cost performance transport, accommodation etc. during the travel. Backpacking has been popular for years, and has become a culture symbolic of the increasingly mobile world [3]. The innovations of information technologies, such as Cloud Services and Mobile Internet etc., have helped a lot to achieve the freedom and flexibility of backpacking.

For backpackers, they are always seeking the transport with high cost performance. As one of the traditional ways of transport, airplane is usually concerned by backpackers, especially those who are travelling to long-distanced places. In conventional opinion, airplane fares seem to be the most expensive option of transport. However, the competition of the single market for air transport has lead to significant changes in the airline industry. One of the most visible demonstrations of these changes is the establishment of a number of low fare, no or low frill airlines, called low-cost airlines (LCA), whose appearance has changed the reality a lot [4]. A low-cost airline, also known as a low-cost carrier (LCC), is an airline that offers generally low fares in exchange for eliminating many traditional services the target groups of passengers do not typically demand [5]. It emerged as a byproduct of the deregulation of the air transportation sector initiated by the US in the 1970s [6]. Especially, the liberalization of the airline industry in Europe in 1990s brought higher quality and drastic price reduction offered by airlines, which

stimulated competition and development in the industry of low-cost airlines [7]. Meanwhile, the emergence of e-Tourism provides much more technologies to utilize low-cost airlines in more convenient ways, such as various mobile applications for searching and booking flights.

The remainder of this paper is organized as follows. In Sect. 39.2, there is related work about Cloud Services in e-Tourism and backpackers. Section 39.3 presents the main design and implementation of Low-Cost Airline Service System. Finally, Sect. 39.4 concludes the paper and brings some issues as future work.

39.2 Related Work

As a rapidly developing technology, Cloud Computing focuses on the goal of "information all the time everywhere". It describes intelligent front-ends for existing and upcoming information technology applications. Also it extends the information service infrastructure, where the back-end systems and service contents must be prepared to react to the greatly increasing backpackers demand for network access or service requests [8]. According to the definition of The National Institute of Standards and Technology (NIST 2011): "Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction".

Based on Cloud Services, there are many systems emerged with the service-oriented technology [1]. OWL technology serves service selection and BPEL serves service execution. Thus Feng Zhou successfully provides a travel service recommendation system for e-Tourism to use travel services. And with the prototype being developed, there can be much more services to take charge of. The Cloud Services are supposed to connect backpackers to the information infrastructures in a persistent and productive way [9].

As the smart travel mentioned in [10], Cloud Service has allowed smart phones to provide a variety of movable services. In [10], Jason C. Hung proposed a design of service including human-centered recommendation service and travel-gaming service. The system called Smart-Travel System aid e-Tourism with personal requirements for tour by smart phones. It provides personalized travelling real-time information, and automatically tells e-Tourism when they show up around the task. With the cloud-based services, this system is a new way to look up the travel information.

Even though Internet has become a new approach for backpackers to search information about attractions and tourism destination before making a purchase decision, it is still difficult for backpackers to collect every piece of information from the Internet. There is a case study about backpacker Internet information search behavior in China [11]. By comparing different kinds of perceived benefit and risk of backpacking, it shows that how they influence the Internet information search behavior of backpackers. There also has been an information system of i3

Fig. 39.1 System architecture and mobile application interface

Travel is planned. Based on the concepts of "information integration" and "initiative service", real-time and location-based transportation, travel information and seamless public transportation quality are provided [12].

Backpacking and low-cost airlines are two topics closely related in traditional tourism industry. The trend of Cloud Computing has provided the basis for the backpacking culture to gain the cohesiveness with e-Tourism. This paper proposes a system aimed at facilitating the backpacker's travelling, especially about low-cost airlines. It can help backpackers to save costs in airlines. Also, backpackers can enjoy a different kind of travelling from the novelties of technologies, not only Mobile Internet, but also Cloud Service.

39.3 Low-Cost Airline Service System

Low-Cost Airline Service System is an application including Cloud Services which can help backpackers make decision about how they travel cheaper through low-cost airlines. With Cloud Computing technology, it tries to deliver most services like computation, software, data access and storage for backpackers who do not need the end-backpackers knowledge and system configuration. The whole system contains client (mobile applications) and servers (cloud services). Cloud Services mainly function as transforming real-time information, computing engine and ticket services. Mobile applications contain task manager and Web services. The system architecture is shown as Fig. 39.1.

And as shown in Fig. 39.1, the normal login interface contains Name, ID, Phone, Email and so on. With the Personal Manager, backpackers can edit the user information by themselves and generate a travel plan with the corresponding item. Once the system is told by the departure place, destination and departure time, it will calculate the best plan for backpackers.

39.3.1 Task Manager

Task manager mainly deals with the demand of backpackers and manage backpackers' personal information, including analyzing backpackers' instructions, submitting the instruction to Web services, storing and modifying backpackers' information.

39.3.2 Web Services

Web services sub-system has the duty to transfer the information which backpackers deal with and the feedback which backpackers need between mobile applications and cloud.

39.3.3 Real-Time Information Transformation

The cloud provides real-time information about related cities or towns provided by backpackers. This paper will introduce how to pick out the related cities or towns below. The real-time information can be retrieved from websites of airlines or other organizations.

39.3.4 Computing Engine

Computing engine can produce a recommendation flight itinerary by comparing and computing the information received from real-time information transformation sub-system.

Here the main idea of shortest path is utilized to generate the least-cost algorithm to find out the cheapest route. In this case, the cost takes the place of distance, and Dijkstra algorithm [13] is deployed in the system for the reasons as follows:

(1) The place of departure and destination are both certain. Dijkstra algorithm deals with the problem which starting vertex is certain.
(2) Global shortest path information is not needed. Dijkstra algorithm only computes the shortest path from starting point to each of other vertices.
(3) Many vertices should be computed. Dijkstra algorithm runs in the complexity of $O(n^2)$ (where n is the number of vertices).
(4) The cost between the two places is not negative. Dijkstra algorithm only works for positive cost.

Here the heap is utilized to store the shortest cost from start to each other node (set Q) instead of array. By this way, it will decrease the complexity of $O(n^2)$ to $O(nlog(n))$ [14].

39.3.5 Booking Services

Booking services sub-system can help backpackers to book Low-Cost Airline flight which is selected by the backpacker from the results of computing engine. Also backpackers can choose and book other flights they like.

39.4 Conclusion

The purpose of backpacking aims for the most travel amusement with the least cost. Low-cost airline is one of feasible methods to achieve that goal. Real travel fans focus more on the thinking and feeling during the travel than other issues. In current information society, the emergence of Cloud Computing provides powerful technology support for e-Tourism. With this kind of evolution, backpackers can enjoy more from travel and low-cost airlines can serve in a more flexible way. The combination promotes the cost performance for backpacking and guarantees the profits of low-cost airlines. Though there is discrimination between different low-cost airlines, for backpacking, this kind of transport has been widely accepted and is worthy of being widely spread and utilized in current e-Tourism.

Acknowledgments We would like to thank the anonymous reviewers for their constructive comments. This work is supported in part by National Natural Science Foundation of China under Grant No. 60973137, Program for New Century Excellent Talents in University under Grant No. NCET-12-0250, Gansu Sci.&Tech. Program under Grant No. 1104GKCA049, 1204GKCA061 and 1212RJYA003, Strategic Priority Research Program of the Chinese Academy of Sciences with Grant No. XDA03030100, The Fundamental Research Funds for the Central Universities under Grant No. lzujbky-2013-k05, lzujbky-2013-43, lzujbky-2013-44 and lzujbky-2012-44, Google Research Awards, and Google Faculty Award.

References

1. Zhou F, Fang Y, Chen H (2012) Personalized travel service discovery and usage in cloud environment. In: IEEE ninth international conference on e-Business engineering (ICEBE), 2012. IEEE, pp 333–337
2. Buhalis D, O'Connor P (2005) Information communication technology revolutionizing tourism. Tourism Recreation Res 30(3):7–16
3. Paris CM (2009) The virtualization of backpacker culture. In: Proceedings of information and communication technologies in tourism, Amsterdam, Jan 2009, pp 25–35

4. Mason KJ (2000) The propensity of business travelers to use low cost airlines. J Transp Geogr 8(2):107–119
5. Knorr A (2007) Southwest airlines: the low cost pioneer at 35. In: Handbook of low cost airlines: strategies, business processes and market environment. Erich Schmidt Verlag, Berlin, pp 77–109
6. Evangelho F, Huse C, Linhares A (2005) Market entry of a low cost airline and impacts on the Brazilian business travelers. J Air Transp Manage 11(2):99–105
7. Piga CA, Filippi N, Bachis E (2002) Booking and flying with low cost airlines. Int J Tourism Res 4(3):237–249
8. Hansmann U, Merk L, Nicklous MS, Stober T (2003) Pervasive computing: the mobile world. Springer, Berlin, pp 11–24
9. Bask J (2001) Pervasive computing: travel and business services, www.tml.tkk.fi/Studies/Tik-111.590/2001s/papers/joni_bask.pdf. Retrieved on 06 Aug 2013
10. Hung JC (2012) The smart–travel system: utilising cloud services to aid traveller with personalised requirement. Int J Web Grid Serv 8(3):279–303
11. Luo F, Li J (2009) Backpacker internet information search behavior: a case study of university students in China. In: 2009 International conference on information management, innovation management and industrial engineering, IEEE, vol 3, pp 547–550
12. Chen CH, Wu TL, Lee MT, et al (2012) A love of travel—demonstration project of Sun-Moon Lake importing ITS/telematics. In: 12th International conference on ITS telecommunications (ITST), IEEE, pp 129–134
13. Dijkstra's algorithm reference webpage address: http://en.wikipedia.org/wiki/Dijkstra's_algorithm
14. Cormen TH, Leiserson CE, Rivest RL, Stein C (2009) Introduction to algorithms, 3rd edn. The MIT press, Cambridge

Chapter 40
Modified Cat Swarm Optimization Algorithm for Feature Selection of Support Vector Machines

Kuan-Cheng Lin, Yi-Hung Huang, Jason C. Hung and Yung-Tso Lin

Abstract Cat swarm optimization (CSO) is a novel meta-heuristic for evolutionary optimization algorithms based on swarm intelligence. CSO imitates the behavior of cats through two sub-modes: seeking and tracing. Previous studies have indicated that CSO algorithms outperform other well-known meta-heuristics, such as genetic algorithms and particle swarm optimization. This study presents a modified version of cat swarm optimization (MCSO), capable of improving search efficiency within the problem space. The basic CSO algorithm was integrated with a local search procedure as well as the feature selection of support vector machines (SVMs). Experimental results demonstrate that the proposed MCSO algorithm provides better results in less time than basic CSO algorithms.

Keywords Swarm intelligence · Cat swarm optimization · Feature selection · Support vector machine

40.1 Introduction

A wide range of bio-inspired algorithms, such as genetic algorithms (GA) [1] and particle swarm optimization (PSO) [2], have been developed to solve optimization problems in numerous domains including project scheduling [3] and intrusion

K.-C. Lin (✉) · Y.-T. Lin
Department of Management Information Systems, National Chung Hsing University, Taichung, Taiwan

Y.-H. Huang
Department of Mathematics Education, National Taichung University of Education, Taichung, Taiwan

J. C. Hung
Department of Information Management, Overseas Chinese University, Taichung, Taiwan

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_40, © Springer Science+Business Media Dordrecht 2014

detection [4, 5]. Cat swarm optimization (CSO) [6] is a recent development based on the behavior of cats, comprising two search modes: seeking and tracing. Seeking mode is modeled on a cat's ability to remain alert to its surroundings, even while at rest; tracing mode emulates the way that cats trace and catch their targets. Experimental results have demonstrated that CSO outperforms PSO in functional optimization problems.

Optimization problems can be divided into functional and combination problems. Most functional optimization problems, such as the determination of extreme values, can be solved through calculus; however, combination optimization cannot be dealt with so efficiently. Feature selection is particularly important in domains such as bioinformatics [7] and pattern recognition [8] for its ability to reduce computing time and enhance classification accuracy.

Classification is a supervised learning technology, in which input data comprises features classified by support vector machines (SVMs) [9], neural networks [10], or decision trees [11]. The classifier is trained by the input data to build a classification model applicable to unknown class data. However, input data often contains redundant or irrelevant features, which increase computing time and may even jeopardize classification accuracy. Selecting features prior to classification is an effective means of enhancing the efficiency and classification accuracy of a classifier.

Two feature selection models are commonly used: filter and wrapper models [12]. Filter models are used to evaluate feature subsets by calculating a number of defined criteria, while the latter evaluates feature subsets through the assembly of classification models followed by accuracy testing. Filter models are faster but provide lower classification accuracy, while wrapper models tend to be slower but provide high classification accuracy. Advancements in computing speed, however, have largely overcome the disadvantages of wrapper models, which has led to its widespread adoption for feature selection. In 2006, Tu, Chang, Chang, and Yang proposed PSO-SVM [13], which obtains good results through the integration of feature subset optimization and parameter optimization by PSO. Researchers have demonstrated the superior performance of CSO-SVM over PSO-SVM [14]; however, its searching ability remains weak. This paper proposes a modified CSO, called MCSO, capable of enhancing the searching ability of CSO through the integration of feature subset optimization and parameter optimization for SVMs.

40.2 Related Works

40.2.1 Support Vector Machine (SVM)

Vapnik [15] first proposed the SVM method based on structural risk minimization theory in 1995. Since that time, SVMs have been widely applied in the solving of problems related to classification and regression. The underlying principle of SVM

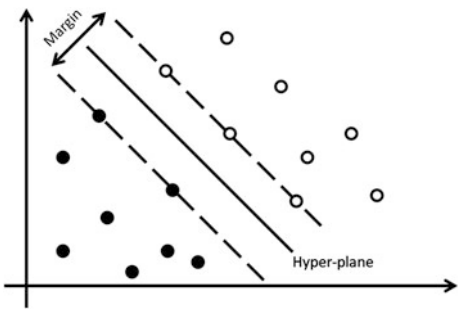


Fig. 40.1 Optimal hyper-plane

theory involves establishing an optimal hyper-plane with which to separate data obtained from different classes. Although more than one hyper-plane may exist, just one hyperplane maximizes the distance between two classes. Figure 40.1 presents an optimal hyper-plane for the separation of two classes of data.

In many cases, data is not linearly separable. Thus, a kernel function is applied to map data into a Vapnik-Chervonekis dimensional space, within which a hyper-plane is identified for the separation of classes. The most common kernel functions is radial basis functions (RBFs), as shown in Eq. (40.1).

$$\Phi(x_i - \mathrm{x}_j) = \exp(-\gamma \left\| x_i - \mathrm{x}_j \right\|) \tag{40.1}$$

This study employed LIBSVM [16, 17] as a classifier and an RBF as its kernel function.

40.2.2 Cat Swarm Optimization (CSO)

CSO [6] process is outlined as follows:

(1) Randomly initialize *N* cats with position and velocities within a D-dimensional space.
(2) Distribute cats to tracing or seeking mode; the number of cats in the two modes determines *MR*.
(3) Measure the fitness value of each cat.
(4) Search for each cat according to its mode in each iteration. The processes involved in these two modes are described in the following sub-section.
(5) Stop the algorithm if terminal criteria are satisfied, otherwise return to (2) for the following iteration.

Seeking Mode

Seeking mode has four operators: seeking memory pool (*SMP*), self position consideration (*SPC*), seeking range of the selected dimension (*SRD*) and counts of dimension to change (*CDC*). *SMP* defines the pool size of seeking memory. For example, an *SMP* value of 5 indicates that each cat is capable of storing 5 solution

sets as candidates. *SPC* is a Boolean value, such that if *SPC* is true, one position within the memory will retain the current solution set and not be changed. *SRD* defines the maximum and minimum values of the seeking range, and *CDC* represents the number of dimensions to be changed in the seeking process. The process of seeking is outlined as follows:

(1) Generate *SMP* copies of the position of the current cat. If *SPC* is true, one of the copies retains the position of the current cat and immediately becomes a candidate. The other cats must be changed before becoming candidates. Otherwise, all SMP copies will perform searching resulting in changes in their positions.
(2) Each copy to be changed randomly changes position by altering the *CDC* percent of dimensions. First, select the *CDC* percent of dimensions. Every selected dimension will randomly change increase or decrease the current value of the *SRD* percent. After being changed, the copies become candidates.
(3) Calculate the fitness value of each candidate via the fitness function.
(4) Calculate the probability of each candidate being selected. If all candidate fitness values are the same, then the selected probability (P_i) is equal to one. Otherwise, P_i is calculated via Eq. (40.2). Variable i is between 0 and SMP; P_i is the probability that this candidate will be selected; $FS_{\max}$ and $FS_{\min}$ represent the maximum and minimum overall fitness values; and FS_i is the fitness value of this candidate. If the goal is to find the solution set with the maximum fitness value, then $FS_b = FS_{\min}$; otherwise $FS_b = FS_{\max}$. Calculating probability using this function gives the better candidate a higher chance of being selected, and vice versa:

$$P_i = \frac{|FS_i - FS_b|}{FS_{\max} - FS_{\min}} \tag{40.2}$$

(5) Randomly select one candidate according to the selected probability (P_i). Once the candidate has been selected, move the current cat to this position.

Tracing Mode

Tracing mode represents cats tracing a target, as follows:

(1) Update velocities using Eq. (40.3).
(2) Update the position of the current cat, according to Eq. (40.4).

$$\begin{aligned} v_{k,d}^{t+1} &= v_{k,d}^{t} + r_1 \times c_1 \times \left(x_{best,d}^{t} - x_{k,d}^{t}\right), \\ d &= 1.2\ldots, D. \end{aligned} \tag{40.3}$$

$$x_{k,d}^{t+1} = x_{k,d}^{t} + v_{k,d}^{t} \tag{40.4}$$

$x^t_{k,d}$ and $v^t_{k,d}$ are the position and velocities of current cat k at iteration t. $x^t_{best,d}$ denotes the best solution set from cat_k in the population. c_1 is a constant and r_1 is a random number between 0 and 1.

40.3 Proposed MCSO-SVM

Neither seeking nor tracing mode is capable of retaining the best cats; however, performing a local search near the best solution sets can help in the selection of better solution sets. This paper proposes a modified cat swarm optimization method (MCSO) to improve searching ability in the vicinity of the best cats. Before examining the process of the algorithm, we must examine a number of relevant issues.

Classifiers

Classifiers are algorithms used to train data for the construction of models used to assign unknown data to the categories in which they belong. This paper adopted an SVM as a classifier. SVM theory is derived from statistical learning theory, based on structural risk minimization. SVMs are used to find a hyper-plane for the separation of two groups of data. This paper regards the SVM as a black box that receives training data for the classification model.

Solution Set Design

Figure 40.2 presents a solution set containing data in two parts: SVM parameters (C and γ), and feature subset variables ($F_1 \sim F_n$), where n is the number of features. The range of the variables in each feature subset falls between 0 and 1. If F_i is greater than 0.5, its corresponding feature is selected; otherwise, the corresponding feature is not chosen.

Mutation Operation

Mutation operators are used to locate new solution sets in the vicinity of other solution sets. When a solution set mutates, every feature has a chance to change. Figure 40.3 presents an example of mutation. Before mutation, the first, fourth, and fifth features of the solution set were assigned for mutation; therefore, changed these features from unselected (0) to selected (1) or selected (1) to unselected (0). This operation is used only for the best solution sets; therefore, it is not necessary to maintain the actual position—recording the selected features is sufficient.

In addition, C and γ must be changed to binary, so they can select a mutation operation for the following search.

Evaluating Fitness

This study employed k-fold cross validation [18] to test the search ability of the algorithms. k was set to 5, indicating that 80 % of the original data was randomly selected as training data and the remainder was used as testing data. And then retain features up to a solution set that we want to evaluate for training data and testing data. This training data is input into an SVM to build a classification model

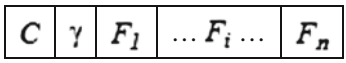


Fig. 40.2 Representation of a solution set

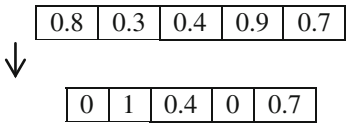


Fig. 40.3 An example of mutation

used for the prediction of testing data. The prediction accuracy of this model represents its fitness.

In order to compare solution sets with the same fitness, we considered the number of selected features. If prediction accuracy were the same, the solution set with fewer selected features would be considered superior.

Proposed MCSO-SVM Approach

The steps of the proposed MCSO-SVM are presented as follows:

(1) Randomly generate N solution sets and velocities with D-dimensional space, represented as cats. Define the following parameters: seeking memory pool (SMP), seeking range of the selected dimension (SRD), counts of dimension to change (CDC), mixture ratio (MR), number of best solution sets (NBS), mutation rate for best solution sets (MR_Best), and number of trying mutation (NTM).
(2) Evaluate the fitness of every solution set using SVM.
(3) Copy the NBS best cats into best solution set (BSS).
(4) Assign cats to seeking mode or tracing mode based on MR.
(5) Perform search operations corresponding to the mode (seeking/tracing) mode assigned to each cat.
(6) Update the BSS. For every cat after the searching process, if it is better than the worst solution set in BSS, then replace the worst solution set with the better solution set.
(7) For each solution set in BSS, search by mutation operation for NTM times. If it is better than the worst solution set in BBS, then replace the worse solution set with the better solution set.

If terminal criteria are satisfied, output the best subset; otherwise, return to (4).

40.4 Experimental Results

This study followed the convention of adopting UCI datasets [19] to measure the classification accuracy of the proposed MCSO-SVM method. Table 40.1 compares MCSO-SVM and CSO-SVM with regard to classification accuracy and the number of selected features. In Australian, Ionosphere, Vehicle, and Vowel datasets, the

Table 40.1 Comparison results for CSO-SVM and MCSO-SVM

Datasets	No. of original features	CSO-SVM		MCSO-SVM	
		No. of selected features	Average accuracy rate (%)	No. of selected features	Average accuracy rate (%)
Australian	14	5.0	90.87	5.0	91.45[b]
Bupa	6	3.8	79.13	4	79.42[b]
German	24	6.6	80.00	8.2	82.50[b]
Glass	9	4.4	83.65	4.2	84.60[b]
Ionosphere	34	7.2	99.72	7.2	100[b]
Pima	8	4.6	81.25	3.6	81.78[b]
Vehicle	18	7.8	86.41	10.2	90.55[b]
Vowel	10	7.8	100	7	100[a]
Wine	13	3	100	2.8	100[a]

[a] Accuracy is equal to the other method; however, fewer features are selected
[b] Higher accuracy

MCSO-SVM had higher classification accuracy and fewer selected features than CSO-SVM. In Bupa, German, Glass, and Pima datasets, MCSO-SVM had higher classification accuracy but more selected features. Due to the advantages afforded by searching in the vicinity of the best solution sets, MCSO-SVM clearly outperformed CSO-SVM.

40.5 Conclusion

This study developed a modified version of CSO (MCSO) to improve searching ability by concentrating searches in the vicinity of the best solution sets. We then combined this with an SVM to produce the MCSO-SVM method of feature selection and SVM parameter optimization to improve classification accuracy. Evaluation using UCI datasets demonstrated that the MCSO-SVM method requires less time than CSO-SVM to obtain classification results of superior accuracy.

References

1. Holland JH (1975) Adaptation in natural and artificial systems. The University Michigan Press, Ann Arbor
2. Kennedy J, Eberhart RC (1995) Particle swarm optimization. In: IEEE international conference on neural networks. Perth, Australia
3. Kim K, Gen M, Kim M (2006) Adaptive genetic algorithms for multi-resource constrained project scheduling problem with multiple modes. Int J Innovative Comput Inf Control 2(1):41–49

4. Li Y et al (2008) A lightweight web server anomaly detection method based on transductive scheme and genetic algorithms. Comput Commun 31(17):4018–4025
5. Su MY (2011) Real-time anomaly detection systems for Denial-of-Service attacks by weighted k-nearest-neighbor classifiers. Expert Syst Appl 38(4):3492–3498
6. Chu SC, Tsai PW (2007) Computational intelligence based on the behavior of cats. Int J Innovative Comput Inf Control 3(1):163–173
7. Stevens R et al (2001) A classification of tasks in bioinformatics. Bioinformatics 17(2):180–188
8. Kishore JK et al (2000) Application of genetic programming for multicategory pattern classification. IEEE Trans Evol Comput 4(3):242–258
9. Furey TS et al (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 16(10):906–914
10. Zhang GQP (2000) Neural networks for classification: a survey. IEEE Trans Syst Man Cybern Part C-Appl Rev 30(4):451–462
11. Quinlan JR (1993) C4.5: programs for machine learning. The Morgan Kaufmann series in machine learning. Morgan Kaufmann Publishers, San Mateo, x, 302 p
12. Liu H, Motoda H (1998) Feature selection for knowledge discovery and data mining. Kluwer international series in engineering and computer science. Kluwer Academic Publishers, Boston, p 214
13. Tu CJ et al (2006) Feature selection using PSO-SVM. In: IMECS 2006: international multiconference of engineers and computer scientists, pp 138–143
14. Lin KC, Chien HY (2009) CSO-based feature selection and parameter optimization for support vector machine. In: Joint conference on pervasive computing, pp 783–788
15. Vapnik VN (1995) The nature of statistical learning theory. Springer, New York
16. Chang C-C, Lin C-J, LIBSVM: a library for support vector machines. Available from: http://www.csie.ntu.edu.tw/∼cjlin/libsvm/
17. Hsu CW, Chang CC, Lin CJ (2003) A practical guide to support vector classification. Available from: http://www.csie.ntu.edu.tw/∼cjlin/papers/guide/guide.pdf
18. Salzberg SL (1997) On comparing classifiers: pitfalls to avoid and a recommended approach. Data Min Knowl Disc 1(3):317–328
19. Hettich S, Blake CL, Merz CJ (1998) UCI repository of machine learning databases. Available from: http//www.ics.uci.edu/∼mlearn/MLRepository.html

Chapter 41
Using Affective Computing to Detect Emotions with Color

Min-Feng Lee, Guey-Shya Chen and Jen-Chieh Wang

Abstract This research uses an innovative way to detect and recognize user's emotion with RGB color system. The RGB color system that is widely used in diverse computer system is an additive color model in which red, green, and blue light are added together showing various color. This study was based on Thayer's emotion model which describes the emotion with two vectors, valence and arousal, and gathers the emotion color with RGB as input of neural network for calculating and forecasting user's emotion. In this experiment, using 320 data translate to quarter into emotion groups to train the weight in the neural network and uses 160 data to proof the accuracy. The result reveals that this model can be valid estimated the emotion by reply color response from examinee. In other hand, this experiment found that trend of the different element of color on Cartesian coordinate system figures out the distinguishing intensity in RGB color system. Via the foregoing detect emotion model is going to design an affective computing intelligence framework try to embed the emotion component in it.

Keywords Affective computing · Neural network · Detecting emotion · Color · Emotion classification

M.-F. Lee (✉) · G.-S. Chen
Graduate Institute of Educational Measurement and Statistics, National Taichung University of Education, Taichung, Taiwan
e-mail: antonio6561@gmail.com

G.-S. Chen
e-mail: grace@mail.ntcu.edu.tw

J.-C. Wang
Department of Business Administration, Overseas Chinese University, Taichung, Taiwan
e-mail: friend22895586@yahoo.com.tw

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_41, © Springer Science+Business Media Dordrecht 2014

41.1 Introduction

Fuzzy cognitive map (FCM) [1] is an effective knowledge presentation and inference model that is based on mankind knowledge and experience. FCM is also a tool for design intelligent systems, which integrate components of neural network and fuzzy set. Moreover, FCM can be showed with the graph form, in which notions are showed as nodes and weighted edges represent concept relations about associations between the notions [2]. Affective computing was included in Artificial intelligence domain; it is computing that related to, cause, or obviously affection emotion [3], which is firstly proposed by Professor Picard at MIT in 1997. Affective computing is a kind of newly research domain and focuses the emotion activities and processes from human. The real class environment seems centralize on the cognitive capacities of the human and makes a fitness knowledge system for themselves [4–6]. The recent researches of psychologists and neurologists point out the major key of the motivation and the affectivity in cognitive activities, especially learning [7–9].

In this paper, the related work is summarized and the frame construction of artificial emotional model process is proposed according to the FCM framework simulating the emotional response. Then the model of emotion of intelligent system is given which is learned in its interaction with intuitive color feedback. The structure of this model was designed according to the Thayer's emotion model [5] one of the emotion classification. In the end, experimental research is used to test the effectiveness of the model.

41.2 Research Related Literatures

41.2.1 Affective Computing

Affective computing is an interdisciplinary field spanning computer sciences, psychology, and cognitive science [12]. And Picard in 1997 described that an affective computing system must have a few of the following capacities: (1) recognize, (2) express, or (3) Possess emotions. It focuses on creating personal computing systems having ability to sense, recognize and understand significantly after the positive interventions than the conditions with no intervention [13]. At present there are multiform ways of current study to detect for emotion status.

- Automatic detection of cognitive-affective states by computers [14, 15].
- Recent progress in real time affects detection [16].
- Progress on affect detection through body movement and gestures [17].
- Through acoustic-prosodic cues [18], lexical features [19] and physiological features [20].

41.2.2 Emotion

Emotions are naturally part of our internal activity and perform a crucial role decision-making and cognitive relation processes. The definition of oxford dictionary online mentioned that "emotion is a strong feeling deriving from one's circumstances, mood, or relationships with others" [21]. Emotion was defined in "Taxonomy of Learning" that is a complex psychophysiological experience of an individual's state of mind as interacting with biochemical (internal) and environmental (external) influences [22]. Emotion is the complex inside processing of an individual's state of mind. Leland Beaumont states that every emotion includes mood, motivation, temperament, personality, and disposition [24].

41.2.3 Emotion Classification

Basic emotions are claimed to be biologically fixed, innate and as a result universal to all humans and many animals as well. Complex emotions are then either refined versions of basic emotions, culturally specific or idiosyncratic. [25] The basic categories of emotion are joy, angry, sad, and happy. The categories of complex emotions can be combined or associated from basic emotions in emotions conditioning. In Table 41.1, Emotion classifications collected from early scholars. Since ancient times, many scholars proposed their emotion classifications. The researchers used their aspect to define the emotion expression making many varieties in emotion classification.

In 1972, Ekman devised a list of basic emotions that are not culturally determined, but universal. The six basic emotions are: (1) Anger, (2) Disgust, (3) Fear, (4) Happiness, (5) Sadness, and (6) Surprise. The newly included emotions are: Amusement, Contempt, Contentment, Embarrassment, Excitement, Guilt, Pride in achievement, Relief, Satisfaction, Sensory pleasure, Shame [26].

41.2.4 Thayer's Emotion Model

Thayer's emotion model was defined from the book, Biopsychology of Mood and Arousal, that is based on mood analysis as biopsychological notion [28]. Thayer's emotion model is often adopt to avert ambiguous description [29]. This emotion model clarifies the emotion classification in a two-dimensional way according to valence quantity (positive to negative) and arousal quantity (exciting to sleepy). The emotion classifications are allotted into the four quadrants of a two-dimensional coordinate system, x axis is valence; y axis is arousal as shown in Fig. 41.1.

Table 41.1 Different emotional classification from scholar

Scholar	Emotion classification
Rene Descartes (1956–1690)	Love, hate, desire, joy, sadness and admiration
Baruch de Spinoza (1632–1677)	Joy, sorrow and desire
Thomas Hobbes (1588–1679)	Appetite, desire, love, aversion, hate, joy, and grief
A. Jorgensen (1894–1977)	Fear, happiness, sorrow, want, and angry shyness
Silvan Solomon Tomkins (1911–1991)	Interest, surprise, joy, anguish, fear, shame, disgust, and rage
Paul Ekman (1934–Now)	Anger, disgust, fear, happiness, sadness and surprise

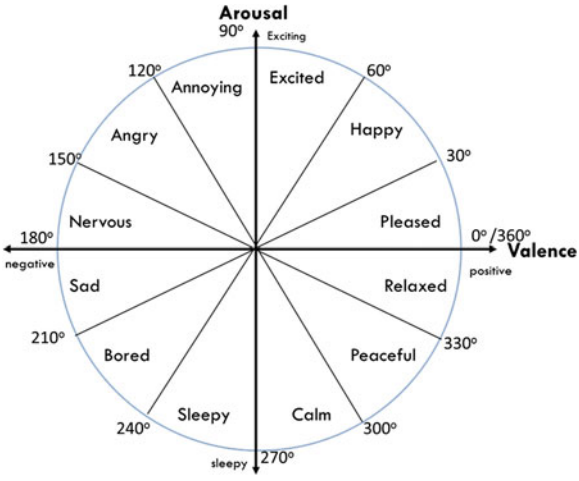


Fig. 41.1 Thayer's emotion model

41.2.5 Fuzzy Cognitive Maps

The traditional FCM decision support technologies suppose that decision making can build a precise indication of a decision issue. The real-world problems, there are usually random event components with complex relations in the solving case. Hence, the application with FCM needs the new properties which are flexible, dynamic and swift. The FCM is a network, showed in Fig. 41.2, that every node describes the event, condition or status and every line between nodes draw the relation, weight, value and sequence.

41.2.6 Backpropagation Neural Network Model

The backpropagation neural network (BPNN) one kind of neural network model is the most widely applied neural network technique to classify or predict [30]. BPNN is layered feed-forward supervised network and the popular tool to solve

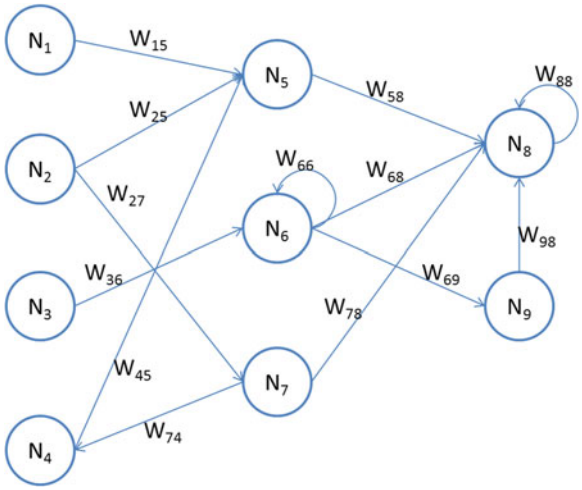


Fig. 41.2 Fuzzy cognitive map network

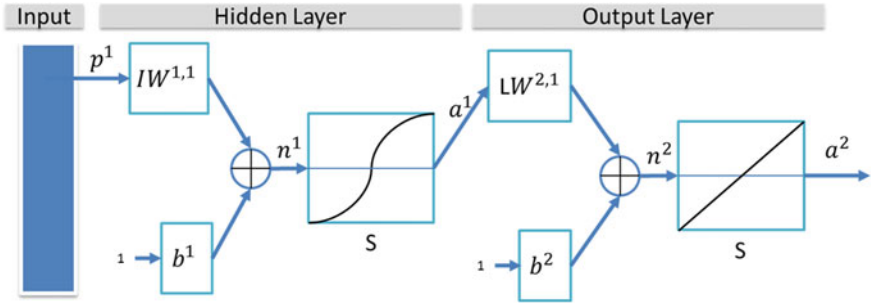


Fig. 41.3 Backpropagation neural network

wide range of problems like classification, recognition and control. The architecture of BPNN is a multilayer neural network with one layer of hidden units is shown in Fig. 41.3.

41.3 Method

This detecting emotion model with color in backpropagation model is an innovative framework for emotion status forecast. This experiment model adopts the two emotion models' (Thayer's model) merit to approach the human's emotion status by RGB color from user's choice.

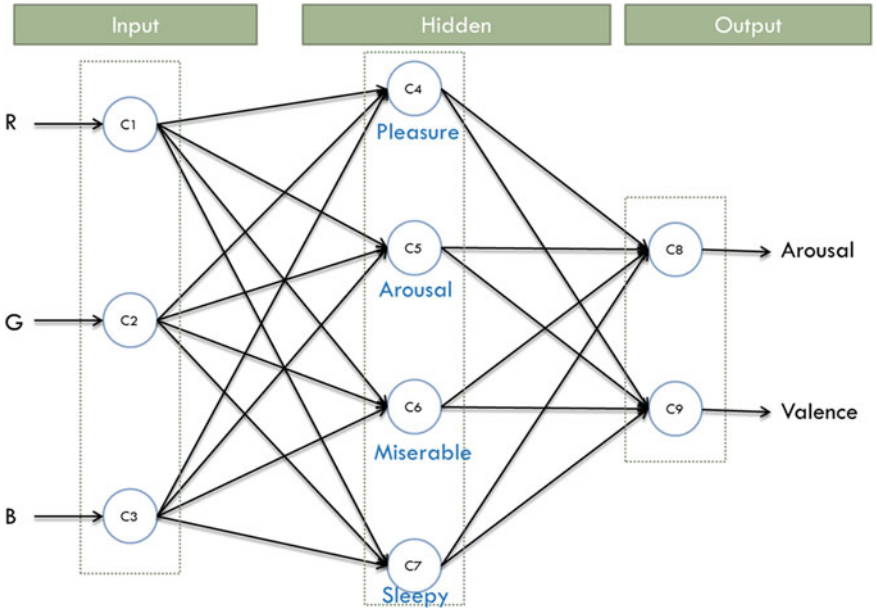


Fig. 41.4 Emotion colors detects with FCM model

41.3.1 Experiment Flow

The experiment flow in the beginning, experiment subjects input their basic data (name, age, gender) and key in the four colors about the emotion (arousal, pleasant, miserable and sleepy). This experiment system imports the subject's data to emotion database and calculates the color record to draw the personal emotion circle.

41.3.2 BPNN Process

This experiment uses the subject's emotion colors (input layer: RGB color), each emotion colors include pleasant, arousal, miserable and sleepy, four emotion status (Hidden layer). Gathering the every set emotion colors importing to the BPNN to train and simulate showed in Fig. 41.4.

41.4 Result

The result of emotion color detecting shows a splendid solution. The result lists below:

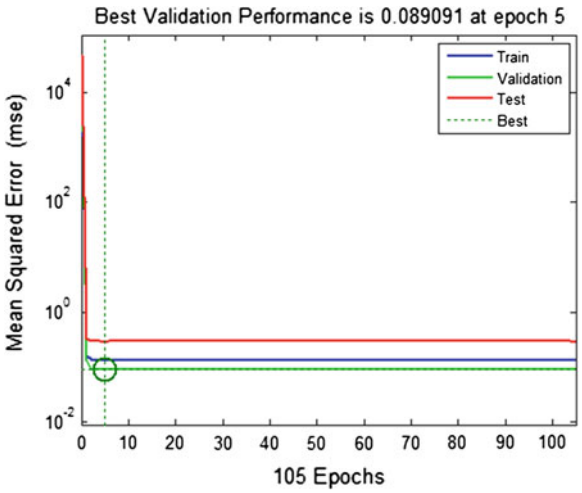


Fig. 41.5 MSE during the training, validation, test and simulation process

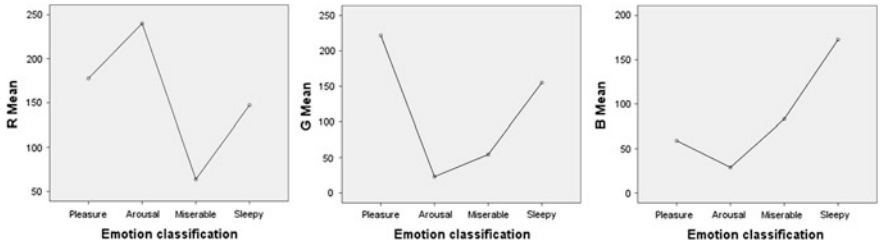


Fig. 41.6 RGB color mean figure

Input: 240 samples from 60 users' RGB color
Output: 120 values with two vectors: arousal and valence
Simulation Best Validation MSE = 0.089091
Simulation MSE = 0.1057, Simulation MAE = 0.1574
Correction zone rate = 78 % (less than 30 degrees).

In the result shows use the FCM model can gradually converge into the stable small value the best validation value is 0.089891 and shown as Fig. 41.5. In Fig. 41.6 shows the RGB color data in four emotion classifications present different trend. The R color in the arousal emotion is the highest value the mean is 240.03 and the second is pleasure emotion. The G color in the pleasure emotion is the highest value the mean is 221.84 and the second is sleepy emotion. The B color in the sleepy emotion is the highest value the mean is 172.79 and the second is miserable emotion.

References

1. Kosko B (1986) Fuzzy cognitive maps. Int J Man Mach Stud 24:65–75
2. Stach W, Kurgan L, Pedrycz W (2010) A divide and conquer method for learning large fuzzy cognitive maps. Fuzzy Sets Syst 161:2515–2532
3. Picard RW (2000) Affective computing: MIT press, Cambridge
4. Cheng BHC, De Lemos R, Giese H, Inverardi P, Magee J (2009) Software engineering for self-adaptive systems. Springer, Berlin, pp 48–70
5. Thayer RE (1989) The biopsychology of mood and arousal. Oxford University Press, Oxford
6. Jaques PA, Vicari RM (2007) A BDI approach to infer student's emotions in an intelligent learning environment. Comput Educ 49:360–384
7. Damasio A (2008) Descartes' error: emotion, reason and the human brain. Random House, New York
8. Izard CE (1984) Emotion-cognition relationships and human development. Emotions Cogn Behav, pp 17–37
9. Goleman D (2006) Emotional intelligence: why it can matter more than IQ. Random House Digital Inc, New York
10. Minsky M (1889) The society of mind. 1988. Simon & Schuser, NY
11. Minsky M (2007) The emotion machine: commonsense thinking, artificial intelligence, and the future of the human mind. SimonandSchuster. com
12. Tao J, Tan T (2005) Affective computing: a review. In: Affective computing and intelligent interaction. Springer, Berlin, pp 981–995
13. Rezazadeh IM, Wang X, Firoozabadi M, Hashemi Golpayegani MR (2011) Using affective human–machine interface to increase the operation performance in virtual construction crane training system: a novel approach. Autom Constr 20:289–298
14. D'Mello S, Chipman P, Graesser A (2007) Posture as a predictor of learner's affective engagement. In: Proceedings of the 29th annual cognitive science society, pp 905–910
15. Kapoor A, Burleson W, Picard RW (2007) Automatic prediction of frustration. Int J Hum Comput Stud 65:724–736
16. Pantic M, Rothkrantz LJ (2003) Toward an affect-sensitive multimodal human-computer interaction. Proc IEEE 91:1370–1390
17. Castellano G, Villalba SD, Camurri A (2007) Recognising human emotions from body movement and gesture dynamics. In: Affective computing and intelligent interaction. Springer, Berlin, pp 71–82
18. Grimm M, Kroschel K, Harris H, Nass C, Schuller B, Rigoll G, Moosmayr T (2007) On the necessity and feasibility of detecting a driver's emotional state while driving. In: Affective computing and intelligent interaction. Springer, Berlin, pp 126–138
19. Wagner J, Vogt T, André E (2007) A systematic comparison of different HMM designs for emotion recognition from acted and spontaneous speech. In: Affective computing and intelligent interaction. Springer, Berlin, pp 114–125
20. Komatsu T, Ueda K, Komeda T (2007) Comprehension of users' subjective interaction states during their interaction with an artificial agent by means of heart rate variability index. In: Affective computing and intelligent interaction. Springer, Berlin, pp 266–277
21. Holden R, Rubery J (2013) Emotion. In: Pearsall J (ed) Oxford dictionaries. Oxford University Press, Oxford
22. Pittman RE, Taxonomy of learning. PediaPress, Germany
23. Ellsworth PC (1994) William James and emotion: is a century of fame worth a century of misunderstanding? Psychol Rev 101:222
24. Hofmann M, Beaumont LR (2005) Content networking: architecture, protocols, and practice. Elsevier, pp 179–216
25. Ekman P (2000) In: Dalgleish T, Power M (eds) Handbook of cognition and emotion. Wiley, pp 45–60

26. Ekman P (2000) In: Dalgleish T, Power M (eds) Handbook of cognition and emotion. Wiley, pp 51–53
27. Russell JA (1980) A circumplex model of affect. J Pers Soc Psychol 39:1161–1178
28. Acampora G, Loia V, Vitiello A (2011) Distributing emotional services in ambient intelligence through cognitive agents. SOCA 5:17–35
29. Yang Y-H, Lin Y-C, Su Y-F, Chen HH (2007) Music emotion classification: a regression approach. In: IEEE international conference on multimedia and expo, 2007, pp 208–211
30. Nielsen RH (1990) Neurocomputing. Addison-Wesley, Reading

Chapter 42
Bucket Index Ordering Problem in Range Queries

Taek-Young Youn, Nam-Su Jho, Ku-Young Chang and Dowon Hong

Abstract In this paper, we first point out a weakness in range queries, which we have named the index ordering problem, where an adversary can recover actual randomized bucket indexes. At first, we define the index ordering problem and examine the security of existing bucket-based encrypted DB management techniques for range queries. We also provide simple countermeasures that enhance the security of bucketing techniques against the bucket index ordering problem.

42.1 Introduction

In recent years, rather than maintaining their own database management system (DBMS), some organizations have managed their data by outsourcing to a trusted third-party (TTP) [1–7]. It is not easy to fully trust such third parties, however, and thus data are submitted to the TTP in an encrypted manner. In this case, servers cannot search specific data from their database since information is stored in an encrypted manner. To solve this problem, random indexes are assigned to ciphertexts to permit the server to search specific data without decryption. When a client wants to retrieve data, it sends certain indexes to the server. The server then searches all ciphertexts that have been assigned by the indexes and gives the retrieved ciphertexts to the client.

T.-Y. Youn (✉) · N.-S. Jho · K.-Y. Chang · D. Hong
Electronics and Telecommunications Research Institute Daejeon, Daejeon, Korea
e-mail: taekyoung@etri.re.kr

T.-Y. Youn · N.-S. Jho · K.-Y. Chang · D. Hong
Kongju National University, Kongju, South Korea

D. Hong
e-mail: dwhong@kongju.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_42, © Springer Science+Business Media Dordrecht 2014

In practice, important information includes numerical data such as social security numbers and various medical records. We can control such data efficiently by permitting users to ask for information through range queries [1, 5–7]. Unfortunately, range queries reveal more information than basic bucket-based database management techniques since adjacent buckets would also be asked for a query, and buckets include numerical data in a sequential order. Some techniques have been proposed to remove this insecurity [5, 7]. They improve the security of range queries by exchanging ciphertexts between two or more buckets. As a result, in both schemes, non-adjacent buckets can be sent for a range query. The feature improves the security of bucketing, but it also decreases the efficiency.

For range queries, bucket boundaries are considered secret data [7]. In other words, bucketing schemes are not secure against adversaries who can guess the boundary values of the buckets. In this paper, we first define the bucket index ordering problem and show that an adversary can find boundary values by solving this problem. We also discuss the insecurity of range queries in existing bucket indexing techniques [5, 7]. We then propose simple countermeasures that can improve the security of range queries.

42.2 Bucket-Based Index for Range Queries

Before defining the bucket index ordering problem inherent in range queries, we describe the basic bucket-based database management technique for range queries using a simple example.

42.2.1 Bucket Construction and Data Query

For secure range queries, we have to adopt an equi-depth partition as indicated in [5, 7]. If we adopt an equi-width partition, the bucketing scheme loses its security as can be seen in the following example:

Example 1 In this example, we consider the salary of employees. The original plaintexts are listed In Table 42.1 in the form of (Employee ID, Salary). The third component is the ciphertext of each salary data. Let us assume that we apply an equi-width partition, [3,000, 4,000) and [4,000, 5,000), and each bucket is identified by random indexes r_1 and r_2, respectively A client requests data by sending a set of indexes. The server then responds to the request by sending all buckets that are assigned to the indexes. To receive data within the range [3,000, 4,000), the client may send a random index r_1 to the server. The server then finds all ciphertexts indexed by r_1 and gives them to the client.

Although random indexes are used to indicate buckets, an adversary can guess a bucket that includes higher salaries since the equi-width partition is used. If the

Table 42.1 Salary data

ID number	Salary	Encrypted message	Bucket index
0001	4,100	B6F548C4AD52E81F	r_2
0002	4,000	C4AD5238B6254AD2	r_2
0003	3,500	548C4ADC4AD523F2	r_1
0004	4,400	ADC4423AD5548CD4	r_2
0005	4,800	1D244D2AD583C515	r_2
0006	3,800	423ADC4AD53AD55A	r_1

adversary knows that more than half of the salaries are greater than 4,000, it is easy to determine that r_1 and r_2 correspond to ranges [3,000, 4,000) and [4,000, 5,000), respectively. □

If we adopt an equi-depth partition, each bucket may include three ciphertexts. One bucket includes ciphertexts for 0,002, 0,003 and 0,006, and the other bucket includes the remainder. The adversary cannot then determine the bucket that includes the higher salary. Hence, the equi-depth partition should be used.

42.2.2 Adversarial Model

Note that most of the numerical data has a probability distribution, and thus it is reasonable to assume that an adversary knows the probability distribution of encrypted data since the description of the target data is not secret information. For the same reason, it was assumed in [5] that an adversary knows the entire bucketing scheme and the exact probability distribution of the values within each bucket. In this paper, we adopt the most suitable adversarial model presented in [7], which makes the following assumptions about the adversary's knowledge: full knowledge of ciphertexts, and approximate knowledge of plaintext distribution. Here, approximate knowledge means an adversary does not learn each distinct plaintext value or the corresponding frequency. However, he learns the plaintext domain range and its probability distribution. We also assume that each range query is a single range query, as in [5].

42.3 Bucket Index Ordering Problem

In this section, we first describe the bucket index ordering (BIO) problem and examine the security of existing schemes.

42.3.1 BIO Problem in a Range Query

To describe the bucket index ordering problem, we consider a normal distribution, where m is the mean value and v is the variance. Let B_i be the set of all elements in the interval $[I_{i-1}, I_i]$. Note that we construct buckets such that $|B_i| = |B_j|$ for all i and j since an equi-depth partition should be adopted for range queries (Fig. 42.1).

As we briefly explained in Sect. 42.2.1, random values are used as indexes for hiding the bucket order. For example, we can assign a random index, r_i, for each bucket B_i, as in Table 42.2.

Throughout the paper, we use the following notations:

- $AI(r)$: actual order of index r
- $r_i \overset{\mapsto}{} r_j : AI(r_i) + 1 = AI(r_j)$
- $||r_i, \ldots, r_j|| : r_i \overset{\mapsto}{} \ldots \overset{\mapsto}{} r_j$
- $|[r_i, \ldots, r_j]| : r_i \overset{\mapsto}{} \ldots \overset{\mapsto}{} r_j$ or $r_j \overset{\mapsto}{} \ldots \overset{\mapsto}{} r_i$
- $(r_i, \ldots, r_j) : ||r_{ai}, \ldots, r_{aj}||$ holds, where $r_{ai} = r_k$ for some k.

For example, in Table 42.2, the followings notations hold: $1 \mapsto 5$, ||2, 10, 3||, |[3, 10, 2]|, and (2, 3, 10).

We will now describe the BIO problem based on the above example. The BIO problem is how to recover the bucket order from random indexes. In practice, an adversary can use the range queries requested by users for recovering the bucket order. The adversary can use the following facts:

Fact 1. A set of buckets returned for a range query are close to each other. In other words, for a given n of indexes r_i, …, r_{i+n}, we can conclude that (r_i, …, r_{i+n}) if the indexes are sent for a query.

When a user makes a range query for $[i, j]$ such that $I_1 < i < I_2$ and $I_3 < j < I_4$, the user sends three random index values 2, 7, and 10 to the server. The server then retrieves three buckets indexed by 2, 7, and 10, and gives them to the user. Based on Fact 1, that the following relation holds: (2, 7, 10). Note that three buckets indexed by 2, 7, and 10 are close to each other, as seen in Table 42.2.

Fact 2. For two given range queries $\{S_1, S_2\}$ and $\{S_2, S_3\}$, we can conclude that $|[S_1, S_2, S_3]|$, where S_1, S_2, and S_3 are three distinct sets of indexes such that $S_1 \cap S_2 = S_2 \cap S_3 = S_1 \cap S_3 = 0$.

When an adversary obtains two range queries {2, 7, 10} and {1, 3, 10}, he can conclude that the following relations hold: (2, 7, 10) and (1, 3, 10), based on Fact 1. Since 10 is included in both queries, the adversary can conclude that |[(1, 3), 10, (2, 7)]| based on Fact 2. Based on Facts 1 and 2, we can recover the order of random indexes as described above. As seen in Example 2, we can obtain the order of indexes when $n - 1$ distinct range queries of size 2 are given.

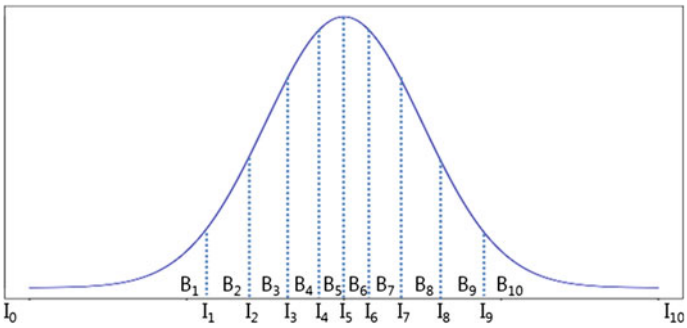


Fig. 42.1 Bucket construction with a normal distribution

Table 42.2 Random index assignment

Bucket	Random index	Bucket	Random index
B_1	$r_1 = 4$	B_6	$r_6 = 1$
B_2	$r_2 = 7$	B_7	$r_7 = 5$
B_3	$r_3 = 2$	B_8	$r_8 = 6$
B_4	$r_4 = 10$	B_9	$r_9 = 9$
B_5	$r_5 = 3$	B_{10}	$r_{10} = 8$

Example 2 We consider the following range queries of size 2: {1, 3}, {1, 5}, {2, 7}, {2, 10}, {3, 10}, {4, 7}, {5, 6}, {6, 9}, and {8, 9}. By Fact 1, we have (1, 3), (1, 5), (2, 7), (2, 10), (3, 10), (4, 7), (5, 6), (6, 9), and (8, 9). By applying Fact 2, we can recover the order of indexes as follows:

Step 1: (1, 3), (3, 10) → I[1, 3, 10]I
Step 2: I[1, 3, 10]I, (2, 10) → I[1, 3, 10, 2]I
Step 3: I[1, 3, 10, 2]I, (2, 7) → I[1, 3, 10, 2, 7]I
Step 4: I[1, 3, 10, 2, 7]I, (4, 7) → I[1, 3, 10, 2, 7, 4]I
Step 5: (1, 5), I[1, 3, 10, 2, 7, 4]I → I[5, 1, 3, 10, 2, 7, 4]I
Step 6: (5, 6), I[5, 1, 3, 10, 2, 7, 4]I → I[6, 5, 1, 3, 10, 2, 7, 4]I
Step 7: (6, 9), I[6, 5, 1, 3, 10, 2, 7, 4]I → I[9, 6, 5, 1, 3, 10, 2, 7, 4]I
Step 8: (8, 9), I[9, 6, 5, 1, 3, 10, 2, 7, 4]I → I[8, 9, 6, 5, 1, 3, 10, 2, 7, 4]I

Note that, while we cannot determine the correct order of indexes among II8, 9, 6, 5, 1, 3, 10, 2, 7, 4II and II4, 7, 2, 10, 3, 1, 5, 6, 9, 8II, the information is still useful to the adversary. Without the order of indexes, the adversary can correctly find the correct bucket order with probability $1/n!$ ($n! = 1*\ldots*n$). However, using the index order, the adversary can guess the bucket order with probability 1/2. The range queries described in Example 2 are a special case, but we use the example in order to emphasis that range queries can reveal the order of randomized indexes.

42.3.2 Determine the Bucket Boundaries

Let I_L and I_R be the left and right end points of bucket B, respectively. Note that two points I_L and I_R are very significant information since we can conclude that all messages in B exist in $[I_L, I_R]$. Here, we show that an adversary can break the security of a bucketing scheme by determining the boundaries if the adversary can solve the BIO problem. Note that we assumed that the adversary knows the probability distribution of the target data and that each equi-depth partition is adopted for bucket construction.

Let p be a probability distribution function of the target data and B_i the i-th bucket that includes all elements in $[I_{i-1}, I_i]$, where $[I_{i-1}, I_i]$ is the i-th sub-interval. We define $\mathbf{INT}(f, i, j)$ as the integral of f from i to j. Since we assume that the adversary knows p, he can guess the end points of B_i by searching two values I_{i-1} and I_i that satisfy the following equations:

$$\mathbf{INT}(p, 0, I_{i-1}) = (i-1)/n \text{ and } \mathbf{INT}(p, 0, I_i) = i/n.$$

It is clear that I_{i-1} and I_i are approximate quantities, which, while not exactly the same in terms of the actual end points, are still useful to the adversary since he can break the privacy of certain data such as medical records and accounts, where the approximate value is also significant. The insecurity is increased if the number of buckets is large and the range size is narrow.

42.3.3 BIO Problem in Existing Bucketing Schemes

Here, we examine the security of existing bucketing schemes proposed for secure range queries [5, 7]. The main idea of the scheme in [5] is to scatter consecutive data to several buckets, and thus the number of transmitted buckets is increased for a range query. Although the scheme can solve the insecurity of a bucketing scheme regarding the BIO problem, its inefficiency is a burden to the database management system. Hence, we exclude it from those of our interest. The other scheme is the local overlapping bucket (LOB) scheme [7]. In the LOB scheme, ciphertexts are exchanged between two adjacent buckets until all buckets seem uniform. Thus, for a uniform distribution, the bucket construction in the LOB scheme is identical with the original bucketing scheme. For the same reason, the security of the LOB scheme is identical with the basic bucketing scheme with uniform probability distribution.

We will now discuss the insecurity of the bucketing scheme for uniformly distributed data. The range I can be divided into n sub-intervals $[0, I_1] \ldots [I_{n-2}, I_{n-1}]$, and $[I_{n-1}, I_n]$, where $I_i = iI/n$. Since an adversary can find the order of buckets, he can guess (with high probability) that the plaintexts for the ciphertexts in B are chosen from the interval $[i|D|/n, (i+1)|D|/n]$ for an integer i. Since the

approximate value is revealed to the adversary, the privacy of the data is compromised. The insecurity is increased when the size of I is small and the number of buckets is large.

42.4 Counter Measure

In this section, we provide simple countermeasures against the BIO problem. Until now, two bucketing schemes have been proposed for securing range queries, and the schemes adopt almost the same idea in which consecutive data is scattered to several buckets. Unfortunately, the scheme in [5] is inefficient, and the other scheme in [7] is not secure against the problem of uniformly distributed data. Therefore, the design of an efficient countermeasure that can enhance the security of the bucketing scheme against this problem is significant.

We can consider two kinds of efficiency. One is computation cost and the other is communication overhead. The scheme in [5] is inefficient in both types due to an increase in false positives. Note that users may not want to receive all ciphertexts assigned to the same index, and unintended ciphertexts are called false positives. We cannot avoid the existence of false positives, but it is desirable to reduce their number to as few as possible, which is particularly important when designing efficient countermeasures.

42.4.1 Random Query

Adversaries can recover the actual order of indexes by using the relation between range queries and the closeness of buckets, and thus we can improve the security of the bucketing scheme by breaking this relationship. One simple way to do this is to make a random query that includes random indexes in the range queries. Let us consider the example described in Table 42.2. In the original bucketing scheme, when a user wants to receive $\{B_2, B_3\}$ and $\{B_3, B_4\}$, two sets of indexes $\{2, 7\}$ and $\{2, 10\}$ are transmitted to the user, respectively. In this case, we can conclude that $|[7, 2, 10]|$ as described in Sect. 42.3.1. However, if we insert random indexes for range queries, adversaries cannot use both Facts 1 and 2. Suppose that a user makes two queries $\{2, 5, 7\}$ and $\{3, 5, 10\}$ for $\{B_2, B_3\}$ and $\{B_4, B_5\}$, respectively. In this case, 5 is a useless index, and thus Fact 2 is not valid. Therefore, we can enhance the security of the bucketing scheme by inserting random indexes in the range queries. By increasing the number of redundant random indexes, we can expect higher security. Though useless buckets are still sent to the requestor as in [5], the computational cost is not increased since the user does not decrypt unexpected buckets. Hence, the random query method is computationally more efficient than the technique in [5].

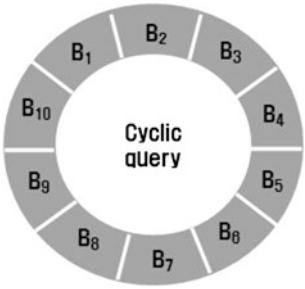


Fig. 42.2 Cyclic structure of cyclic bucket query

42.4.2 Cyclic Query

When the order of indexes is discovered, an adversary can guess two end points as discussed in Sect. 42.3.2. Even if the order of indexes is recovered, we can improve the security of the bucketing scheme by making it hard to determine the correct position of a bucket. We achieve this security feature by performing cyclic queries. The basic idea of a cyclic query is to append additional indexes of adjacent buckets under the assumption that the indexes form a cyclic structure in the sense that the first and last buckets are close to each other. For example, when we want to receive two buckets B_1 and B_2 in Table 42.2, we may send three indexes $\{4, 7, 8\}$. Note that 8 in the index of B_{10} can be included in the query since we assumed that the two end buckets are close to each other. If we perform a cyclic query, the manipulated query makes the adversary decide that B_1 and B_{10} are adjacent buckets, and thus the adversary may construct the following structure using the collected cyclic queries (Fig. 42.2).

Since the adversary cannot determine the first and last bucket from the above cyclic structure, we can expect that the adversary guesses the original order of indexes with probability $1/2n$, since he will have to guess one condition among the following n relation: $|[r_1, \ldots r_{n-1}, r_n]|$, $|[r_2, \ldots r_n, r_1]|$, …, and $|[r_n, \ldots r_{n-2}, r_{n-1}]|$, and each condition implies two possible cases. Note that the adversary can guess the order of indexes with probability 1/2 for the basic bucketing scheme. Therefore, we can enhance the security of a simple bucketing scheme even though the order of indexes is revealed to the adversary.

42.5 Conclusion

In this paper, we defined the bucket index ordering problem in range queries and discussed the security of the existing bucketing schemes. We also provided two simple countermeasures.

Acknowledgments This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the

Ministry of Science, ICT and Future Planning (Grant No. 2011-0029925). This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (Grant No. 2011-0029926).

References

1. Agrawal R, Kiernan J, Srikant R et al (2004) Order preserving encryption for numeric data. In: Proceedings of SIGMOD'04, pp 563–574
2. Ceselli A, Damiani E, Vimercati S et al (2005) Modeling and assessing inference exposure in encrypted databases. ACM Trans Inf Syst Secur 8(1):119–152
3. Damiani E, Vimercati S, Jajodia S et al (2003) Balancing confidentiality and efficiency in untrusted relational DBMSs. In: Proceedings of CCS'03, pp 93–102
4. Hacigumus H, Iyer B, Li C et al (2002) Executing SQL over encrypted data in the database-service-provider model. In: Proceedings of ACM SIGMOD'02, pp 216–227
5. Hore B, Mehrotra S, Tsudik G (2004) A privacy-preserving index for range queries. In: Proceedings of VLDB'04, pp 720–731
6. Jang SM, Song SI, Yoo JS (2007) An efficient PAB-based query indexing for processing continuous queries on moving objects. ETRI J 29(05):691–693
7. Wang J, Du X (2008) LOB: bucket based index for range queries. In: Proceedings of WAIM'08, pp 86–92

Chapter 43
Signal Interference of Ubiquitous Wireless Networks on Data Throughput

Jie Zhang, Goo Yeon Lee, Hwa Jong Kim and Yong Lee

Abstract Ubiquitous wireless networks consist of many wireless access points and devices, which share and contend for wireless medium. Generally, wireless devices are trying to generate high throughput and have longer signal coverage which also bring unnecessary signal interference to neighboring wireless networks, and result in decreased network throughput. Signal interference is an inevitable problem because of the broadcast nature of wireless transmission. However it could be minimized by reducing signal coverage of wireless devices. On the other hand, small signal coverage means low transmission power and low data throughput. In the paper, we analyze the relationship among signal strength, coverage, interference and network throughput in ubiquitous wireless networks by simulation on various network topologies.

Keywords Ubiquitous wireless networks · Signal interference · Throughput · Signal coverage

J. Zhang · G. Y. Lee (✉) · H. J. Kim · Y. Lee
Department of Computer and Science Engineering, Kangwon National University, Chuncheon, South Korea
e-mail: leegyeon@kangwon.ac.kr

J. Zhang
e-mail: zargggg@gmail.com

H. J. Kim
e-mail: hjkim@kangwon.ac.kr

Y. Lee
e-mail: yleehyun@gmail.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_43, © Springer Science+Business Media Dordrecht 2014

43.1 Introduction

Ubiquitous computing is an advanced computing concept where computing is made to appear everywhere and anywhere, which can be achieved with ubiquitous wireless networks deployed all around the area.

In the ubiquitous computing area, many wireless networks can be used, and they usually need high power for larger transmit range and higher data throughput. However, with rapid increasing number of wireless users such as smart phones, large signal coverage of wireless transmission causes higher signal interference among the wireless devices, which causes network quality deterioration especially in densely populated areas.

Signal interference is a common problem in wireless network, which may decrease throughput and cause security problems. The signal interference is inevitable due to wireless characteristics but could be reduced by controlling network configuration. Burkhart et al. [1] and Karagiorgas et al. [2] suggested topology control methods with changes of transmission power in order to reduce signal interference in wireless ad-hoc network.

In wireless networks, the interference can be managed by controlling the sender's transmission power. However, small signal power results in low throughput. There is a trade-off between signal interference and network throughput. In the paper, we analyze the relationship among signal transmission power, signal interference and network throughput with simulations to find a better way to improve quality of wireless network.

43.2 Related Work

Signal interference deteriorates network throughput and also wastes power. Reducing signal interference is widely researched in wireless ad-hoc network in order to minimize power consumption. Burkhart et al. [1] compared several topology control methods which claim to resolve interference, and proposed an interference-minimal method in wireless ad-hoc network with connectivity-preserving and spanner construction. Karagiorgas et al. [2] introduced a multicost routing that constructs route with variable transmission power to reduce interference in ad-hoc network.

Bianchi [3] made a study on performance analysis of CSMA/CD protocol based on IEEE 802.11 RTS/CTS exchange procedure. He showed that due to the characteristics of CSMA/CD protocol, high transmission power causes more adjacent devices to sense carriers, resulting in decrement of network efficiency from signal interference. Tongngam [4] proposed a reducible transmission range approach for wireless network, which optimizes broadcasting latency. Tinnirello and Bianchi [5] analyzed the interference effects in WiFi networks. Kashyap et al. [6] presented a passive monitoring of wireless traffic to estimate interference in WiFi networks.

Above researches show that in any of wireless network like ad-hoc network or broadcasting network, performance in data throughput takes high influence from signal interference. Wireless interference cannot be removed because of broadcasting properties but can be optimized by routing algorithms or topology control algorithms.

43.3 Signal Interference in WiFi Network

Wireless networks are composed of Basic Station Set (BSS), where each BSS contains an AP as a relay station to the Internet for the Wireless Local Area Network (WLAN). The most widely used scheme for a wireless network station to choose the appropriate AP is measuring the received signal strength from AP, known as RSSI (Received Signal Strength Indication). However, many researches showed that AP selection scheme based on the RSSI does not coincide with that of throughput based algorithm [7–9]. This is because the wireless communication quality depends on signal interference, fading and many other effect besides the signal strength itself. In the paper, we first simulate an interference-free environment and measure data throughput with different signal strength to find out the relationship between RSSI level and network throughput. We then extend the experiment with various signal interferences.

43.3.1 Network Throughput in Interference-Free Environment

Table 43.1 shows the specifications of the wireless device we used in the experiment to measure the RSSI and throughput under interference-free conditions. We chose a time of 3:00–4:00 a.m. that few wireless user around to avoid signal interference, experiment space of no objects between AP and clients.

We measured the upload and download throughput varying distances between AP and wireless device to adjust receiving signal strength for every 5 dBm of RSSI from −35 to −70 dBm, and the result is shown in Fig. 43.1.

Maximum bandwidth of the wireless channel was 54 Mbps, however, the actual maximum throughput was about 22 Mbps because of the protocol overhead, such as control traffic and Ack frames etc.

Table 43.1 Experiment device parameters used to measure the RSSI and throughput

Wireless protocol	802.11g
Transmission power (AP)	18 dBm
Antenna gain (AP)	4 dBi
Receive sensitivity	−74 dBm
Maximum signal range	70 m
Maximum throughput	54 Mbps

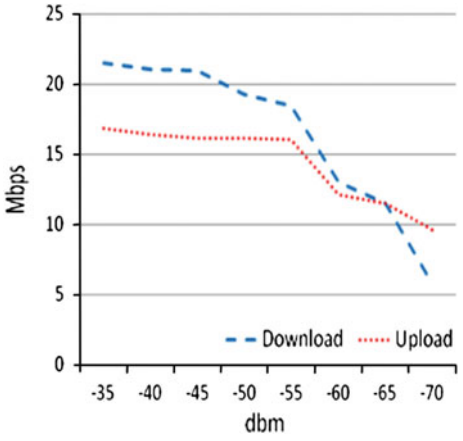


Fig. 43.1 Throughput in a wireless network without signal interference as a function of RSSI

43.3.2 Affects to Network Throughput from Signal Interference

In order to analyze interference effects in wireless network, we estimate the number of active wireless networks in the signal range and evaluate the average throughput. Signal range is expressed by the path loss model [10] as

$$PL = PL_{1\,\mathrm{m}} + 10\log(d^n) + s \tag{43.1}$$

$$RSSI = TxPower + AntennaGain - PL. \tag{43.2}$$

Variables used in the formula are:

- PL: Total path loss experienced between the receiver and sender in dB
- $PL_{1\ \mathrm{m}}$: Reference path loss in dB for the desired frequency when the receiver-to-transmitter distance is 1 m
- d: Distance between the transmitter and receiver in meters
- n: Path loss exponent for the environment, 2 in free space, 3.5–4.5 in indoor environment
- s: Standard deviation associated with the degree of shadow fading in dB (3–7 dB).

Table 43.2 Network throughput and signal transmission distance for different RSSI values

RSSI (dB)	Distance (m)	Throughput (downlink) (Mbps)
−35	1	21.49
−40	1.778279	21.03
−45	3.162278	20.94
−50	5.623413	19.25
−55	10	18.42
−60	17.78279	13.04
−65	31.62278	11.47
−70	56.23413	5.79

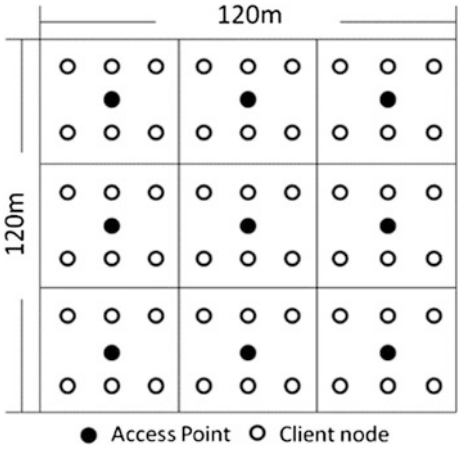


Fig. 43.2 Simulation topology with 9 cells and 54 clients

From (43.1) and (43.2), signal distance between transmitter and receiver is given by

$$d = 10^{(TxPower + AntennaGain - RSSI - PL_{1m} - s)/10\,n} \tag{43.3}$$

Signal distance between AP and client devices obtained from the experiment of Sect. 43.3.1 and Eq. (43.3) is shown in Table 43.2. In the derivation, $PL_{1\ m}$ is set to be 54 dB, path loss exponent n is 2 assuming free space, and shadow fading s is 3 dB. Exact distance while experimenting was a little different due to path loss, etc., however the comparative distribution is calculated very similar.

We performed simulations to investigate the interference effects in ubiquitous wireless networks. The simulation topology is illustrated in Fig. 43.2. A 120 m × 120 m area is divided into 9 cells, each cell contains one AP at the center and 6 devices are evenly located in each cell. Network throughput for downlink (from AP to clients) is measured for different transmission powers of 18, 13 and 8 dBm.

In the simulation, we assumed all clients use same channel and channel association time is equally shared by all client devices no matter how the actual

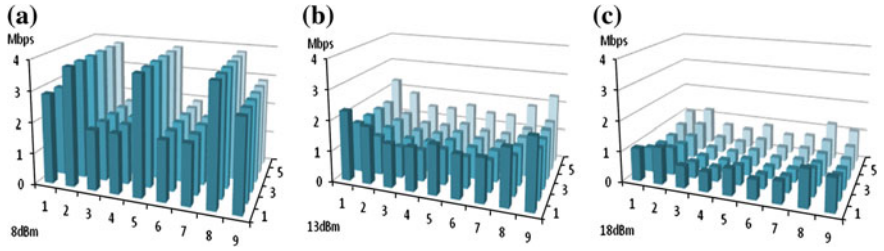


Fig. 43.3 Throughput distribution for 6 × 9 clients with transmission power **a** 8 dBm, **b** 13 dBm and **c** 18 dBm

throughput according to distance is. Note that any client in a wireless coverage with k − 1 other activated clients, data throughput would be decreased by 1/k. Figure 43.3 shows network throughput for each clients with each having transmission power of 8, 13 and 18 dBm.

Network throughput with transmission power of 8 dBm showed the best performance because of the minimum interference. In Fig. 43.3a, clients at [2, y], [5, y], [8, y] (where [x, y] denotes client's position (x, y) in Fig. 43.3) show best throughput because it is close to AP, which means high received signal strength. Edge clients at [1, y] and [9, y] show better throughput than clients in [3, y], [4, y], [6, y] and [7, y] which have same distance to APs but have less interference from other APs. Figure 43.3b and c also show that clients have higher throughput at the corners of the cell area.

High transmission power increases throughput in interference-free environments, however, today's wireless devices rarely show non-interference cases. For instance, from our smart phone's WiFi setting menu, the number of reachable APs just means the interference we may take from. Moreover, the number does not even include interference of devices such as neighboring smart phones.

Low transmission power gives small signal coverage and less signal interference, resulting in higher throughput as shown in the above simulations. However, smaller signal coverage may also make shadow zone where clients may not reach APs. The simulation results in Fig. 43.4 shows this situation. The simulation environment in Fig. 43.4 is the same as the former simulation except the number of nodes in cells which increased from 54(6 × 9) devices to 81(9 × 9) devices.

Figure 43.4b and c show similar throughput distributions as Fig. 43.3b and c. However, in Fig. 43.4a, we can find that some nodes have zero throughputs. This is because the nodes are located in shadow zone where they do not reach APs because of smaller signal coverage. At the expenses of shadow zone nodes, the other nodes can have higher throughput. Figure 43.5 shows the average throughput comparisons of different transmission powers.

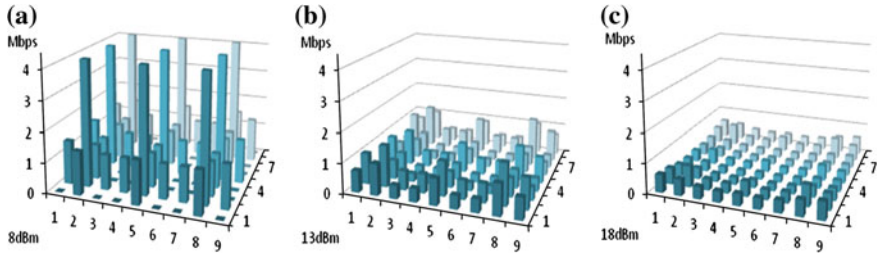


Fig. 43.4 Throughput distribution for 9 × 9 clients with transmission power **a** 8 dBm, **b** 13 dBm and **c** 18 dBm

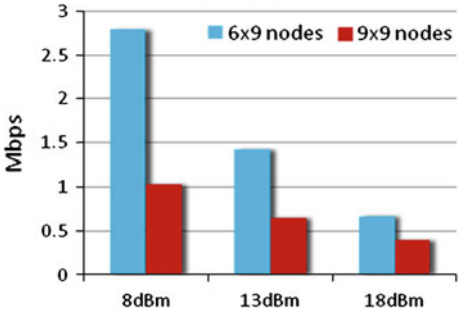


Fig. 43.5 Average throughput comparisons of different transmission powers

43.4 Conclusion

Performance of ubiquitous wireless networks does not only depend on signal strength but also on the interference from neighbor wireless networks and devices. With the motivation, we analyzed the relationship between signal power, signal interference, and network throughput. From the results of the analysis, we see that higher transmission power from AP gives higher signal interference to other wireless networks and, on other hand, if transmission power is too small it can reduce signal interference but may cause shadow zones where client devices could not connect to the network. Also, we found that network throughput could be maximized if transmission power of APs in wireless networks are properly controlled.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. NRF-2011-0013951). And this research was also supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the CITRC (Convergence Information Technology Research Center) support program (NIPA-2013-H0401-13-1002) supervised by the NIPA (National IT Industry Promotion Agency).

References

1. Burkhart M, von Rickenbach P, Wattenhofer R, Zollinger A (2004) Does topology control reduce interference? In: Proceedings of the 5th ACM international symposium on mobile ad hoc networking and computing
2. Karagiorgas NM, Kokkinos PC, Papageorgiou CA, Varvarigos EA (2007) Multicost routing in wireless ad-hoc networks with variable transmission power. In: IEEE 18th international symposium on personal, indoor and mobile radio communications, pp 1–5
3. Bianchi G (2000) Performance analysis of the IEEE 802.11 distributed coordination function. Sel Areas Commun 18:535–547
4. Tongngam S (2011) A reducible transmission range approach for interference-aware broadcasting in wireless networks. In: International conference on future information technology, pp 144–148
5. Tinnirello I, Bianchi G (2010) Interference estimation in IEEE 802.11 networks. Control Syst 30:30–43
6. Kashyap A, Paul U, Das SR (2010) Deconstructing interference relations in Wi-Fi networks. In: 7th Annual IEEE communications society conference on sensor mesh and Ad Hoc communications and networks, pp 1–9
7. Fukuda Y, Honjo M, Oie Y (2006) Development of access point selection architecture with avoiding interference for WLANs. In: IEEE 17th international symposium on personal, indoor and mobile radio communications
8. Lee H, Kim S, Lee O, Choi S, Lee S-J (2008) Available bandwidth-based association in IEEE 802.11 wireless LANs. In: Proceedings of the 11th international symposium on modeling, analysis and simulation of wireless and mobile systems
9. Wu D, Djukic P, Mohapatra P (2008) Determining 802.11 link quality with passive measurements. In: IEEE international symposium wireless communication systems, pp 728–732
10. Cisco (2008) Wi-Fi location-based services 4.1 design guide—location tracking approaches

Chapter 44
Privacy-Preserving Data Collection with Self-Awareness Protection

Kok-Seng Wong and Myung Ho Kim

Abstract Data privacy protection is an emerging issue in data collection due to increasing concerns related to security and privacy. In the current data collection approaches, data collector is a dominant player who enforces the secure protocol. In other words, privacy protection is only defined by the data collector without the participation of any respondents. Furthermore, the privacy protection becomes more crucial when the raw data analysis is performed by the data collector itself. In view of this, some of the respondents might refuse to contribute their personal data or submit inaccurate data. In this paper, we study a self-awareness protocol to raise the confidence of the respondents when submitting their personal data to the data collector. Our self-awareness protocol requires each respondent to help others in preserving his privacy. At the end of the protocol execution, respondents can verify the protection level (i.e., *k*-anonymity) they will receive from the data collector.

Keywords Privacy-preserving data collection · Self-awareness privacy protection · *k*-anonymity

44.1 Introduction

Data collection is a process which involves collaboration between a trusted party (data collector) and a number of data owners (respondents). Due to concerns regarding privacy, respondents might refuse to contribute their personal data or

K.-S. Wong · M. H. Kim (✉)
School of Computer Science and Engineering, Soongsil University, Information Science Building, Sangdo-Dong Dongjak-Gu, Seoul 156-743, South Korea
e-mail: kmh@ssu.ac.kr

K.-S. Wong
e-mail: kswong@ssu.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_44, © Springer Science+Business Media Dordrecht 2014

submit inaccurate data to the data collector. Therefore, the data collector needs to ensure the privacy of data submitted through a series of secure mechanisms. However, the protection level provided by the data collector is hard to be verified by the respondents.

Often, the data submitted by the respondents will be released for research or data analysis. There are two settings can be observed when the data is released to the data recipient. If the data recipient is a third party, data must be released in an anonymous form without compromising the privacy of the respondents. Let's consider a scenario where a hospital (data collector) wishes to publish patients' records to a research institute (data recipient) for data analysis. In a common practice, all the explicit personal identity information (*PII*) such as name and social security number will be removed from the original dataset before it is release to the data recipient. However, removing *PII* does not preserve privacy.

Data anonymization is an interesting solution to protect the privacy of the respondents for this setting. Sweeney and Samarati proposed *k*-anonymity model to address the linking attack [1]. The concept of *k*-anonymity [2] is such that each released data is indistinct from at least $(k - 1)$ other data. However, *k*-anonymity is found vulnerable against background knowledge attacks by Machanavajjhala et al. [3]. Another privacy model called *l*-diversity model was proposed in [3] to complement the *k*-anonymity model. This model requires the representation of sensitive attributes in the released dataset with at least *l* "well-represented" values. A survey of recent attacks and privacy models in data publishing can be found in [4].

In this paper, we consider the second setting where the data analysis is performed by the data collector. This scenario is more complex to deal with because the data collector has the full access to all raw data from the respondents. Therefore, we need to design a protocol to increase the confidence of the respondents before they submit their records to the data collector. In other words, respondents are aware about the protection level they received from the data collector after the data submission.

44.2 Background and Definitions

44.2.1 Homomorphic Encryption Scheme

We use homomorphic encryption scheme (i.e., Paillier [5]) as our primary cryptographic tool. Let $Enc_{pk}(m)$ denote the encryption of m with the public key, pk. Given two ciphertexts $Enc_{pk}(m_1)$ and $Enc_{pk}(m_2)$, there exists an efficient algorithm $+_h$ to compute $Enc_{pk}(m_1 + m_2)$. This additive property can be performed without the decryption key.

44.2.2 Preliminary Definitions

Let us assume there are n respondents $\mathcal{R} = \{\mathcal{R}^1, \mathcal{R}^2, \ldots, \mathcal{R}^n\}$ and a data collector $\mathcal{C}$. Each respondent i has a database $\mathcal{D}_i$ with m records. We denote T as the dataset collected by the data collector. Also, the dataset T consists of d quasi-identifier $QID = \{QI_1, QI_2, \ldots QI_d\}$ and a sensitive attribute. A quasi-identifier (QI) is a minimal set of attributes in T that can be joined with external information to uniquely distinguish individual records [1]. Note that the quasi-identifier can be either categorical or continuous data while the sensitive attribute is a categorical data from its domain.

Definition 1 (*Self-awareness privacy*) Each respondent i is said to achieve self-awareness privacy if he learns the protection level (e.g., k-anonymity) provides by the data collector. At the end of the protocol execution, each respondent is remaining anonymous to others and the data collector is not able to identify any of the respondents with probability more than 0.5.

44.3 Related Works

The self-awareness privacy idea presented in this paper is motivated by the co-privacy concept in [6, 7]. Co-privacy (or co-operative privacy) considers the best option for a party to achieve his privacy protection is to help another party in achieving her privacy. The formal definition of co-privacy and its generalizations can be found in [6].

Various self-oriented privacy protections have been proposed in the literature. Self-enforcing privacy (SEP) for e-polling was proposed in [8]. The idea of SEP is to enforce the pollster to protect the respondents' privacy by allowing the respondents to trace their data after the submission. If the pollster releases the poll results, the respondents can indict the pollster by using the evidence they obtained during the data collection process. A fair indictment scheme for SEP can be found in [9].

The most related research to our work in this paper is the respondent-defined privacy protection (RDPP) for anonymous data collection proposed in [10]. The basic idea of RDPP is to allow the respondents to specify the level of protection they require before providing any data to the data collector. For instance, a number of respondents (minimum threshold) must satisfy the constraint chosen by the respondent i before he agrees to submit the data. In their protocol, respondents are aware of the minimum level of privacy protection they will receive before submit their dataset to the data collector. Instead of relying on the data collector to guarantee the privacy protection, the respondents are free to define their preferred protection level.

In this paper, we do not consider indictment for our protocol because the data analysis is done by the data collector. Instead of allowing the respondents to freely define their own privacies, we assume that respondents are willing to submit their data if the protection level offered by the data collector can be verified by them.

44.4 Our Algorithm

The basic idea of our protocol is to allow the respondents to know the protection level they will receive from the data collector before the data submission process. In our design, the data collector releases a set of quasi-identifier $QID = \{QI_1, QI_2, \ldots, QI_n\}$ for T and defines a protection level it wants to provide to the respondents (e.g., a threshold k). Note that a larger k will make the respondents feel more comfortable to submit their records. We also require the respondents to collaborate together to find the number of records in $(\mathcal{D}_1 \cup \mathcal{D}_2 \cup \ldots \cup \mathcal{D}_n)$ which met the quasi-identifier determined by the data collector. We assume the communication between the data collector and the respondents is via a mixture network such as Tor [11].

In the preliminary phase, the data collector releases QID and k for T. Each respondent i first generates a Paillier [5] cryptographic key pair (pk_i, pr_i) and then encrypts his personal identifiable information (PII) such as name or social security number by using pk_i. The encrypted PII will be used as the public identity $\mathcal{I}_i$ of the respondent i. Each respondent i sends $(\mathcal{J}_i, pk_i)$ to the data collector via a Tor network.

Next, the data collector randomly assigns a public key pk_i for each QI_i. Let us assume the number of public keys is same as the size of the quasi-identifier and $\ell = \{(pk_1, QI_1), (pk_2, QI_2), \ldots, (pk_n, QI_n)\}$. Next, the data collector publishes ℓ and $\mathcal{I}_i$ of each respondent i into a share location (e.g., a webpage) as follows:

$$\{(\mathcal{I}_1, (pk_1, QI_1)), (\mathcal{I}_2, (pk_2, QI_2)), \ldots, (\mathcal{I}_n, (pk_n, QI_n))\} \tag{44.1}$$

Based on the public identity $\mathcal{I}_i$ from Eq. (44.1), each respondent i retrieves ℓ and examines if his records in $\mathcal{D}_i$ match any of the quasi-identifier, $QI_i \in QID$. At this step, each respondent i maintains a list $\{s_i^1, s_i^2, \ldots, s_i^n\}$ for QID. We denote s_i^j as the score determines by the respondent i for QI_j. The respondent raises this score by 1 when a record in $\mathcal{D}_i$ matches the quasi-identifier. Next, the respondent encrypts each score by using its corresponding public key. For example, s_i^j will be encrypted by using pk_j to produce $\alpha_i = \{Enc_{pk_1}(s_i^1), Enc_{pk_2}(s_i^2), \ldots, Enc_{pk_n}(s_i^n)\}$. All the respondents send α_i to the shared location and the data collector, respectively.

Upon receiving α_i from all the respondents, the data collector aggregates the scores for each QI_i by using additive property of the Paillier cryptosystem to produce $Enc_{pk_j}(\mathcal{S}_i) = Enc_{pk_j}(s_1^j) +_h Enc_{pk_j}(s_2^j) +_h \ldots +_h Enc_{pk_j}(s_n^j)$. The data

Table 44.1 Outcome table

	v_1	v_2	$\cdots$	v_n
$\mathcal{U}_1$	$Enc_{pk_1}(s_1^1)$	$Enc_{pk_2}(s_1^2)$	$\cdots$	$Enc_{pk_n}(s_1^n)$
$\mathcal{U}_2$	$Enc_{pk_1}(s_2^1)$	$Enc_{pk_2}(s_2^2)$	$\cdots$	$Enc_{pk_n}(s_2^n)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathcal{U}_n$	$Enc_{pk_1}(s_n^1)$	$Enc_{pk_2}(s_n^2)$	$\cdots$	$Enc_{pk_n}(s_n^n)$
SUM	$Enc_{pk_1}(\mathcal{S}_1)$	$Enc_{pk_2}(\mathcal{S}_2)$	$\cdots$	$Enc_{pk_n}(\mathcal{S}_n)$

collector then publishes all the computed results in an outcome table as shown in Table 44.1.

In Table 44.1, each row $(\mathcal{U}_i)$ represents the encrypted scores received from each respondent while the column (v_i) shows the encrypted scores for each of the quasi-identifier. Note that all the data in v_i are encrypted by using the same respondent's public key pk_i. Therefore, only the respondent who has been assigned the QI_i can learn the number of matched records $(\mathcal{S}_i)$for the quasi-identifier.

After the release of information in Table 44.1, each respondent i needs to verify that the contents of $\mathcal{U}_i$ are identical as those stored in the shared location. If any of the data released by the data collector is differ, the respondent then issues a decision message m_i with a random value. Let us assume all the information in Table 44.1 have been verified by every respondents. Next, each respondent i retrieves v_i and decrypts its contents by using the private key pr_i. After the decryption, the respondent i computes $\mathcal{S}_i = \sum_{j=1}^{n} \left(s_j^i\right)$. At this stage, each respondent verifies his computed score with the SUM value computed by the data collector in Table 44.1. Lastly, the respondent compares $\mathcal{S}_i$ with the threshold k. If the number of matched records $\mathcal{S}_i$ is greater than the threshold (e.g., $\mathcal{S}_i > k$), we assume that the respondent will submit his records to the data collector. Otherwise, the respondent will abort from the data collection process.

At the final phase, each respondent i sends a decision message m_i to the shared location. If the decision message m_i is set to 1, this indicates that $\mathcal{S}_i > k$. Therefore, the respondents should submit their records to the data collector. Otherwise, if m_i is set to 0, the respondents should not reveal any record to the data collector.

44.5 Analysis and Discussion

In our protocol design, we utilize Tor network to prevent direct communication between the data collector and the respondents. This approach will not allow the data collector to track the identity of any respondent. Also, we assume that each respondent has no knowledge about the profile of other respondents, but the number of respondents in the protocol is known publicly. During the protocol execution, each respondent is required to verify (1) the encrypted scores and

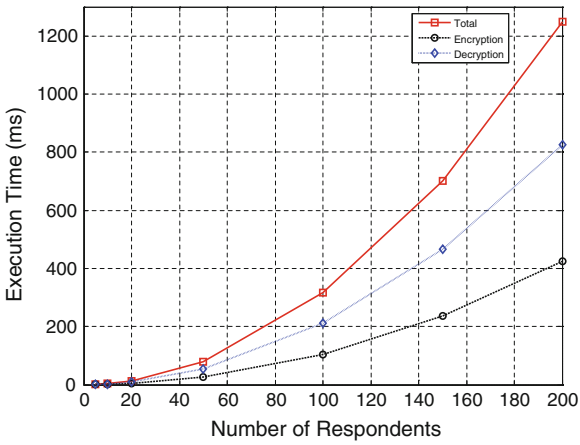


Fig. 44.1 Performance of the proposed solution

(2) the SUM value in Table 44.1. The first verification is to ensure that the data collector has received all the data correctly while the second verification is used to detect malicious activities by the adversary. Note that the data collector can define difference anonymous levels for each QI_i. The respondents then perform the same steps to verify each value of k_i.

In this paper, we assume that the size of the public keys (or the number of respondents) and the quasi-identifier is equal (e.g., $|\mathcal{R}| = |QID| = d$). However, our protocol works correctly for unequal cases. The owner of the public key only performs the decryption and computes $\mathcal{S}_i$ at the end of the protocol execution. A respondent may not be involved in the final phase if his public key is not selected by the data collector ($|\mathcal{R}| > d$). Otherwise, a respondent needs to repeat final phase for several times if his public key is assigned to more than one QI.

The complexity of our protocol is dominated by the cryptographic operations (encryption and decryption) performed by respondents. We implement our protocol in Java and ran it on a single computer with a 2 GHz CPU and a 2 GB RAM. The performance evaluation is shown in Fig. 44.1. Each respondent performs the same amount of cryptographic operations in our experiment.

44.6 Conclusion

In this paper, we presented a self-awareness protocol for data collection process. Since the release of raw data to the data collector has a high risk to compromise privacy of the respondents, we aim to increase confidence of the respondents before they submit their records to the data collector. Our self-awareness protocol allows each respondent to help others in order to preserve his own privacy. Also, our solution can be extended to support indictment scheme (when the data is

released to a third party) because the respondents have evidence (e.g., value of k) to indict a malicious data collector.

References

1. Sweeney L (2002) k-Anonymity: a model for protecting privacy. Int J Uncertainty Fuzziness Knowl Based Syst 10:557–570
2. Samarati P, Sweeney L (1998) Generalizing data to provide anonymity when disclosing information (abstract). In: Proceedings of the 7th ACM SIGACT-SIGMOD-SIGART symposium on principles of database systems. ACM, Seattle, Washington, United States, pp 188
3. Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M (2007) l-diversity: privacy beyond k-anonymity. ACM Trans Knowl Discov Data 1:3
4. Fung BCM, Wang K, Chen R, Yu PS (2010) Privacy-preserving data publishing: a survey of recent developments. ACM Comput Surv 42:1–53
5. Paillier P (1999) Public-key cryptosystems based on composite degree residuosity classes. In: Proceedings of the 17th international conference on theory and application of cryptographic techniques. Springer, Prague, Czech Republic, pp 223–238
6. Domingo-Ferrer J (2010) Coprivacy: towards a theory of sustainable privacy. In: Proceedings of the 2010 international conference on Privacy in statistical databases. Springer, Corfu, Greece pp 258–268
7. Domingo-Ferrer J (2011) Coprivacy: an introduction to the theory and applications of co-operative privacy. SORT Stat Oper Res Trans 35:25–40
8. Golle P, McSherry F, Mironov I (2006) Data collection with self-enforcing privacy. In: Proceedings of the 13th ACM conference on computer and communications security. ACM, Alexandria, Virginia, USA, pp 69–78
9. Stegelmann M (2010) Towards fair indictment for data collection with self-enforcing privacy. In: Rannenberg K, Varadharajan V, Weber C (eds) Security and privacy—silver linings in the cloud, vol 330. Springer, Berlin Heidelberg, pp 265–276
10. Kumar R, Gopal RD, Garfinkel RS (2010) Freedom of privacy: anonymous data collection with respondent-defined privacy protection. INFORMS J Comput 22:471–481
11. Dingledine R, Mathewson N, Syverson P (2004) Tor: the second-generation onion router. In: Proceedings of the 13th conference on USENIX security symposium, vol 13. USENIX Association, San Diego, CA, pp 21–21

Chapter 45
User-Centric Security Service Model for Cloud Storage

SeongMin Yoo, PyungKoo Park, JinSeop Shin, HoYong Ryu and JaeCheol Ryou

Abstract For the Personal Cloud Storage, security is an important issue. There is latent threat when it comes to the data loss or leakage which may be committed by malicious Cloud Service Provider (CSP) employee. The most basic solution is to encrypt user's data. However, when the encryption key is directly managed by CSP or user's device, then there is latent threat, too. In this paper, we propose the User-Centric Security Service Model (USSM). This security service model enables user to store mandatory key fragment and enables only user to use the encryption key since optional key fragments are stored in a dispersed manner. Meanwhile, even when mandatory key fragment is lost, this prevents leakage of encryption key.

Keywords Key management · Cloud security · Personal cloud · Cloud service

S. Yoo (✉) · J. Shin · J. Ryou
Department of Computer Engineering, Chungnam National University, Daejeon, South Korea
e-mail: mingoon@home.cnu.ac.kr

J. Shin
e-mail: whiter1cka@home.cnu.ac.kr

J. Ryou
e-mail: jcryou@home.cnu.ac.kr

P. Park · H. Ryu
Network SW Research Section, ETRI, Daejeon, South Korea
e-mail: parkpk@etri.re.kr

H. Ryu
e-mail: hyryou@etri.re.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_45, © Springer Science+Business Media Dordrecht 2014

45.1 Introduction

For the Cloud service, security is an important issue since user does not own IT resources, and since parts or all of them are secured through outsourcing. Cloud environment's security threat is not very different from the security threat of the existing computing environment, and it includes Personal Cloud or Cloud security threat.

Not-for-profit organization, CSA (Cloud Security Alliance) that shares Cloud security issue analyzed Cloud computing's security threat by dividing it into seven categories. Among these security threats, the security threats that this research paper addresses are the 'Threat #3 Malicious Insiders' and 'Threat #5 Data Loss or Leakage' that are related to the data that is stored in the Cloud Storage [1, 2].

User data in case of Cloud service has a structure in which data is stored in the storage of a third party that is provided by the Cloud Service Provider (CSP). Thus, it is characterized by the following:

- User has limited authority over user's data.
- CSP has excessive authority over user's data.

Accordingly, there is latent threat when it comes to the data loss or leakage which may be committed by malicious CSP employee. This in turn leads to very low reliability from the viewpoint of user when using Cloud service. Accordingly, security mechanism that can guarantee integrity and confidentiality to user's data must be provided. The most basic solution measure is to encrypt user's data that is stored. However, when the encryption key that is used for encryption is managed by CSP, then latent security threat by malicious insiders continues to exist. On the contrary, when encryption key is managed on user's personal device (PC, mobile device, etc.), then there is the threat resulting from the loss of the device.

The purpose of this work is to present security service model that can minimize threat of loss while storing encryption key in user's personal device. The general idea behind the security service model that is proposed, entails dividing up the encryption key into various fragments, to store them at different locations. In particular, one among various fragments is designated as the fragment mandatory for encryption key recovery, and applicable fragment is stored in user's personal device, ensuring that user will have the final authority for the use of the encryption key. This fundamentally blocks CSP's approach to encryption key, which in turn delegates all the authorities pertaining to the data encrypted and stored in the Cloud Storage to user, who are the owners of the data.

This research paper is composed as follows. First and foremost, Sect. 45.2 examines the security service model's scenario and security threat to resolve. Section 45.3 proposes Weighted Information Fragmenting Scheme (WIFS) that divides up the confidential information into various fragments and that can designate one fragment as the mandatory fragment. Section 45.4 proposes User-centric Security Service Model (USSM) that is based on the WIFS. Section 45.5

analyses the security of the proposed protocol. Lastly, Sect. 45.6 concludes by explaining about the limitations and contributions of this work.

45.2 Objective of User-Centric Security Service Model

This chapter explains the Model (USSM)'s scenario that this research paper proposes, and the security requirements needed to resolve this problem.

45.2.1 Scenario for USSM

In case of Cloud service, CSP provides Cloud Storage to user. Users store their data in the Cloud Storage. When needed, they bring their data from Cloud Storage. Among the data that is stored as such, there is regular data and private data that require security such as personal photo or personal information. In this case, user wants to get confidentiality and integrity guaranteed while using encryption service.

When it is assumed that user's data was encrypted with safe encryption algorithm, applicable data's safety relies on encryption key's safe management. If encryption key is managed by CSP, this assumes the same level of threat as that of the instance in which data is not encrypted due to malicious CSP employee. On the contrary, if user manage encryption key directly in person, then there is the threat of losing the encryption key. Accordingly, user wants to ensure that no one else would know the encryption key while they want safety in case of loss as well.

45.2.2 Definition for Requirements of USSM

Security requirements that need to be addressed to provide security service model mentioned above can be summarized as follows.

- Only user, who is the owner of the data, should be able to conduct decryption on the encrypted data.
- Only user should be able to use the encryption key used for encryption.
- Encryption key should be safe even when user loses it.

When these security requirements are satisfied, decryption of the encrypted data that is stored in the Cloud Storage takes place, focusing on user. Accordingly, it is possible to prevent data loss or leakage by malicious CSP employee.

45.3 Weighted Information Fragmenting Scheme

This chapter explains about the Weighted Information Fragmenting Scheme (WIFS). WIFS is the scheme that divides confidential information into n number of fragments and where confidential information is recovered if there are at least $t + 1$ fragments ($n >= t + 1$). In particular, by designating one fragment mandatory for the confidential information recovery among number of fragments, it can block recovery of confidential information without that mandatory fragment. This scheme is used to manage encryption key safely when it comes to User-centric Security Service Model to be explained on Sect. 45.4.

WIFS's confidential information division and recovery process are based on Shamir's Secret Sharing Scheme [3, 4].

45.3.1 Fragmentation Process of the Secret Information

In the WIFS, the process for diving up the confidential information into n number of fragments is as follows. First and foremost, confidential information S is assumed as the point that passes the two-dimensional plane's y-axis, and the following definition can be concluded when the mandatory fragment M is assumed as the arbitrary point that passes the two-dimensional plane's x-axis.

$$\begin{aligned} S &= \{0, y_s\} \\ M &= \{x_M, 0\} \end{aligned} \tag{45.1}$$

Defined S and M are used to determine arbitrary polynomial $f(x)$. Process for deciding polynomial $f(x)$ is as follows. First, arbitrary polynomial $q(x)$, which is the $t - 1$ degree that passes the confidential information S as shown on Eq. (45.2) is decided.

$$q(x) = \sum_{i=1}^{t-1} a_i x^i + a_0 \tag{45.2}$$

where, $a_0 = y_s$, $(a_1, \ldots, a_{t-1})$ is random values

Next, when M, S and polynomial $q(x)$ are used, it is possible to determine the arbitrary polynomial $f(x)$, which is the $t-1$ degree, that passes the two-dimensional plane's point $\{x_M, y_s\}$ as shown on Eq. (45.3). Finally, optional fragments O_i of $n - 1$ are generated at the $f(x)$ as shown on Eq. (45.4).

$$f(x) = q(x) - q(x_M) + a_0 \quad \text{where, } a_0 = y_s \tag{45.3}$$

$$O_i = \{i, f(i)\} \quad \text{where, } 1 \le i \le n-1 \tag{45.4}$$

The scheme that divides up the above mentioned confidential information into n number of fragments is used for dividing, dispersing and storing the encryption key in the data storage protocol explained on Clause 4.1.

45.3.2 Defragmentation Process of the Secret Information

The original information recovery process is as follows. First, M and O_i that is at least t is collected. O_i that is at least t collected among these is used to determine polynomial $f(x)$. When Lagrange interpolation is used as shown on Eq. (45.5), then it is possible to determine the polynomial $f(x)$ that is the only $t - 1$ degree [5]. In the polynomial $f(x)$ determined as such, S' is calculated in the end by using M as shown on Eq. (45.6).

$$f(x) = \sum_{j=1}^{n-1} \left[y_j \times \prod_{\substack{i=1 \\ (i \neq j)}}^{n-1} \left(\frac{x - x_i}{x_j - x_i} \right) \right] \tag{45.5}$$

$$S' = \{0, f(x_M)\} \tag{45.6}$$

The scheme that recovers the above mentioned confidential information is used for recovering user's encryption key when it comes to the data use protocol explained on Clause 4.2.

45.3.3 Algorithm Proof

Polynomial $f(x)$ decided on Eq. (45.5) in accordance to Lagrange interpolation is the same as the Eq. (45.3)'s polynomial $f(x)$. Accordingly, this $f(x)$ passes through two-dimensional plane's point $\{x_M, y_s\}$. Thus, S' calculated ultimately on Eq. (45.6) is the same as the confidential information S. This is proven as follows in accordance to Eq. (45.3).

$$\begin{aligned} f(x_M) &= q(x) - q(x_M) + a_0 \\ &= q(x_M) - q(x_M) + a_0 \\ &= a_0 = y_s \end{aligned} \tag{45.7}$$

In other words, S' is the $\{0, y_s\}$, and this is the same as the Eq. (45.1)'s confidential information S.

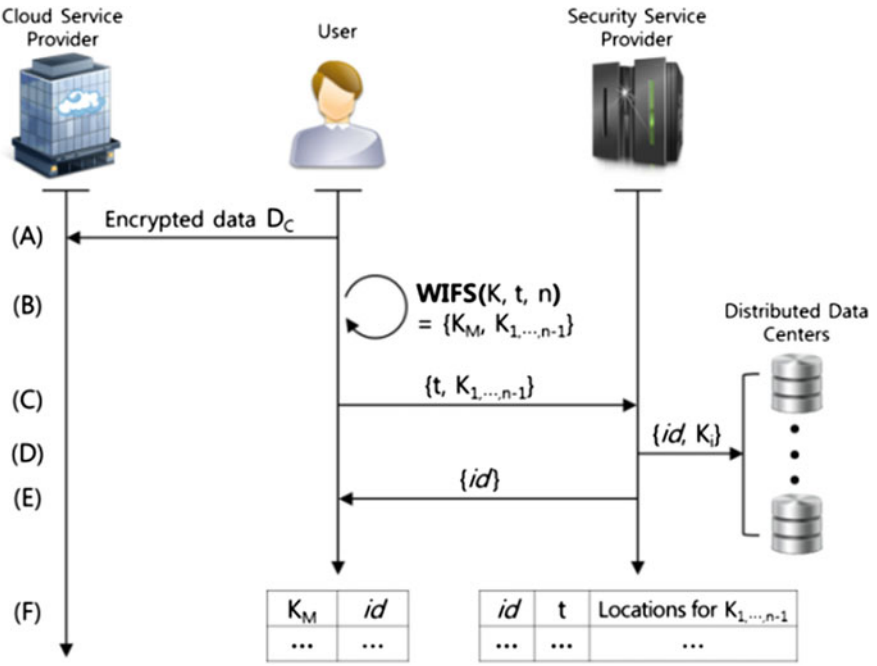


Fig. 45.1 Data storage protocol

45.4 USSM: User-Centric Security Service Model

This chapter explains User-centric Security Service Model (USSM) protocol needed to delegate all the authorities over the encrypted data that is stored in the Cloud Storage to user. This protocol is based on the WIFS scheme proposed on Sect. 45.3.

45.4.1 Data Storage Protocol

Procedure with which user store data in the Cloud Storage can be explained using total of six steps as shown on Fig. 45.1.

In the Step (A), user encrypts data and stores it in the Cloud Storage. In the Step (B), WIFS scheme is used to divide encryption key into n number of fragments. In the Step (C), Optional Key Fragments are sent to Security Service Provider (SSP). In the Step (D), SSP stores Key Fragments in the Distributed Data Centers (DDCs). Finally, user stores the *id* for finding Mandatory Key Fragment and Optional Key Fragments while SSP stores the information on the location where Optional Key Fragments are stored during Steps (E) and (F). Detailed process for

using WIFS scheme to divide encryption key to store in a dispersed manner is as follows:

1. User selects n number of Key Fragments and minimum number of Key Fragments, t ($t <= n - 1$) needed for recovery.
2. When encryption key K and t, n are input in the WIFS scheme, one Mandatory Key Fragment K_M and $n - 1$ number of Optional Key Fragments $K_{1, \ldots, n-1}$ are generated.

$$\text{WIFS } (K, t, n) = \{K_M, K_{1,\ldots,n-1}\} \tag{45.8}$$

3. Generated $K_{1,\ldots, n-1}$ and t are transmitted to SSP.

$$\text{User} \rightarrow \text{SSP: } \{t, K_{1,\ldots,n-1}\} \tag{45.9}$$

4. SSP disperses $K_{1,\ldots, n-1}$ in the registered DDCs. At this time, *id* that is needed to issue Key Fragments later on is stored as well.

$$\text{SSP} \rightarrow \text{DDCs:} \{id, K_i\} \tag{45.10}$$

5. SSP stores the information on the location here *id*, t and Key Fragment are dispersed, and returns *id* to user.

id	t	Locations for $K_{1,\ldots,n-1}$
...	...	...

6. Finally, user stores Mandatory Key Fragments and applicable *id*.

K_M	*id*
...	...

45.4.2 Data Use Protocol

Procedure for user to bring the data stored in the Cloud Storage can be explained with a total of six steps as shown on Fig. 45.2.

In Step (A), user brings encrypted data from Cloud Storage. In Step (B), request is made to the SSP for the applicable data's Key Fragments. In Step (C), SSP collects t number of Key Fragments that is dispersed in the DDCs and delivers it to user during Step (D). In Step (E), user utilizes WIFS scheme to recover encryption key. Finally, user decrypts the encrypted data with the recovered encryption key during Step (F). Detailed process for collecting Key Fragments and for recovering encryption key by using WIFS scheme is as follows:

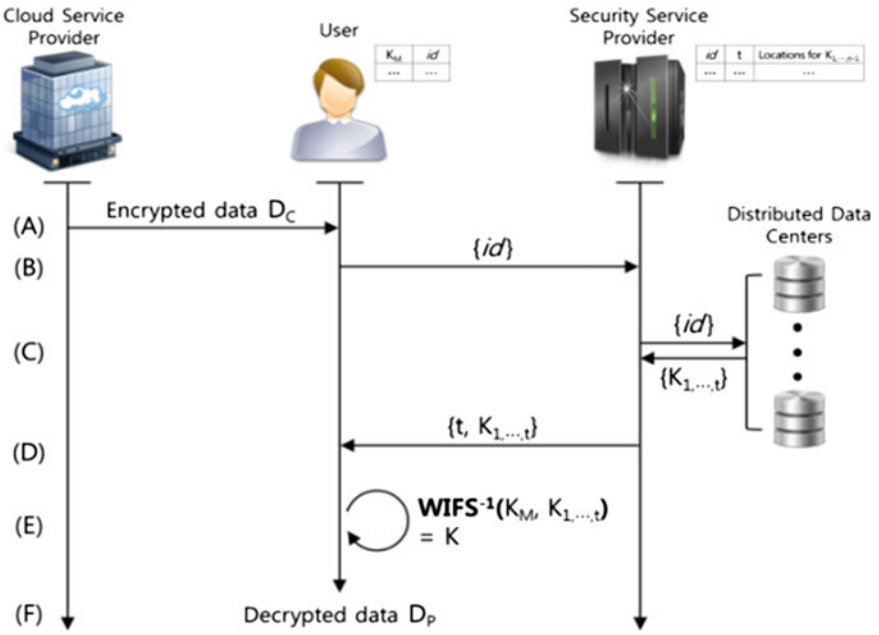


Fig. 45.2 Data use protocol

1. User delivers the *id* of the encryption key to recover to the SSP.
2. SSP collects t number of Key Fragments from DDCs with applicable *id*. Collected $K_{1,\ldots,t}$ and t are delivered to user.

$$\text{SSP} \rightarrow \text{User}: \{t, K_{1,\ldots,t}\} \tag{45.11}$$

3. Encryption key K is recovered when user inputs the Mandatory Key Fragment K_M and t number of Optional Key Fragments $K_{1,\ldots,t}$ that was stored into the WIFS^{-1} scheme.

$$\text{WIFS}^{-1}(K_M, K_{1,\ldots,t-1}) = K \tag{45.12}$$

4. Finally, user decrypts the encrypted data with recovered encryption key K.

45.5 Security Assessments

This chapter analyses how well the security service model proposed on Sect. 45.4 satisfies the security requirements identified on Sect. 45.2.

Confidential information S pertaining to the Weighted Information Fragmenting Scheme proposed on Sect. 45.3 depends on mandatory fragment $M = \{x_M, 0\}$ and

polynomial $f(x)$ because S is $\{0, f(x_M)\}$. However, there is no correlation between M and polynomial $f(x)$ because M is a point that exists on the two-dimensional plane's x-axis regardless of the polynomial $f(x)$, and because polynomial $f(x)$ is determined only by at least t number of Optional Fragment O_i based on polynomial interpolation.

Accordingly, to find out the original confidential information S, M and at least t number of Optional Fragment O_i are needed. These characteristics need to satisfy the following security requirements.

CSP shouldn't be able to get hold of the encryption key

When the above explained security service model's scenario is examined, there is no point of contact between CSP and encryption key during data storage and use. Accordingly, encryption key cannot be obtained when CSP tries to get hold of the encryption key since it is necessary to collect all these; mandatory key fragment K_M that user has and at least t number of optional key fragments O_i that was stored after getting dispersed by SSP.

SSP should not be able to get hold of the encryption key

SSP knows the location of the optional key fragments O_i that was stored after getting dispersed. Accordingly, SSP can collect at least t number of optional key fragments O_i. However, in order to recover encryption key K, it is necessary to get hold of the mandatory key fragment K_M that user has. Thus, it is impossible to get hold of the encryption key.

Should be safe when user loses device

When user loses personal device, mandatory key fragment K_M can be leaked. However, in order to recover encryption key K, it is necessary to collect at least t number of optional key fragments O_i first and foremost, and the only polynomial $f(x)$ needs to be determined. Accordingly, loss of the personal device alone won't leak the encryption key.

45.6 Conclusions

This research paper proposed security service model for Cloud Storage based on Weighted Information Fragmenting Scheme. This security service model has applied the WIFS scheme to the management of the encryption key for the data that is encrypted and stored in the Cloud Storage. This scheme enables user to store mandatory key fragment and enables only user to use the encryption key since optional key fragments are stored in a dispersed manner. Meanwhile, even when mandatory key fragment is lost, this prevents leakage of encryption key.

Finally, overall security was analyzed when it comes to the proposed protocol. Through this, proposed security service model appeared to satisfy the requirements identified on Sect. 45.2.

This work is not without limitations. This research assumes that there is a user certification mechanism between user and SSP when it comes to the Step (B)'s data use protocol that can be trusted. In case this assumption should not exist, there is a latent threat that encryption key may be leaked by the malicious user who acquired the personal device lost by user since it is possible to collect at least t number of optional key fragments O_i. However, if and when certification mechanism exists, then it is possible for user to take appropriate measures prior to the leakage of encryption key or to prevent this threat depending on the strength level of the certification mechanism.

This research's future work entails conducting research on this certification mechanism. In particular, lightweight certification mechanism that is suitable for mobile Cloud environment is needed. Another important issue is the lack of performance resulting from limited hardware specs of mobile device due to its nature. Accordingly, research on the performance of the proposed scheme will be conducted in the near future.

Acknowledgments This work was partly supported by the IT R&D program of MSIP/KEIT. [10041579, Development the Personal Information Security service solution using tokenization technology] and the R&D program of MKE (Ministry of Knowledge Economy) [Project No. 10043380].

References

1. Drago I, Mellia M, Munafo MM, Sperotto A, Sadre R, Pras A (2012) Inside dropbox: understanding personal cloud storage services. In: Proceedings of the 2012 ACM conference on internet measurement conference, pp 481–494 (Nov 2012)
2. Archer J, Boehm A (2009) Security guidance for critical areas of focus in cloud computing. Cloud Secur Alliance
3. Shamir A (1979) How to share a secret. Commun ACM 22(11):612–613
4. Blakley GR (1979) Safeguarding cryptographic keys. In: Proceedings of the national computer conference, pp 313–317
5. Berrut JP, Lloyd NT (2004) Barycentric lagrange interpolation. Siam Rev 46(3):501–517

Chapter 46
Continuity Labeling Technique of Multiple Face in Multiple Frame

Hyun-jung Kim, Il-Yong Weon and Chang-hun Lee

Abstract Research on recognizing and tracking objects have recently been carried out actively. There are especially high numbers of application fields using face recognition and tracking. Existing methods for recognizing and tracking objects have many difficulties when the target is multiple and of the same type. This study is on the continuous labeling method of the same face between frames in videos in which the faces of multiple people are included. The core of the algorithm is divided into detecting the face regions in one frame and recognizing faces using the previous frame, while applying the suitable methods. The usefulness of the proposed method was proven through experimentation and somewhat achievements were acquired through the test results.

Keywords FLA · Multiple frame · Multiple face tracking

46.1 Introduction

Research has been continuously made on methods for recognizing and tracking objects in not only still images, but also in videos. Research on recognizing and tracking faces has been applied in various fields such as biometric authentication

H. Kim (✉) · C. Lee
Department of Computer Science and Engineering, Konkuk University, Seoul, South Korea
e-mail: nygirl@konkuk.ac.kr

C. Lee
e-mail: chlee@konkuk.ac.kr

I.-Y. Weon
Cyber Hacking Security Seoul Hoseo Technical College, Seoul, South Korea
e-mail: clccclcc@shoseo.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_46, © Springer Science+Business Media Dordrecht 2014

systems, secure entrance systems, security surveillance systems, military systems, and interaction between humans and robots [1, 2].

Existing methods for recognizing and tracking objects in videos are commonly made up of methods that recognize the position of the object via the different video of the standard frame and test frame to connect the continuity of the previous object in the following video using the directional features of videos [2].

Such method may have some achievement in a restricted environment, but it has the difficulty of having to classify each and every object in the condition that that target being tracked is multiple, but of the same type [3]. In particular, methods that only use the directional features of object are difficult to use when the object in the previous frame severely changes its position in the surveillance frame. For the previous frame, if the object in the following frame is a human face, traditional supervision-learning based facial recognition learning algorithms can be applied to procure the continuity of the object. However, such method becomes obsolete when there is a lot to compute and when unspecified faces not registered through advanced learning appears.

In this paper, we discussed labeling methods of the same person by detecting multiple faces region in continuing image frames that include the faces of many people and finding the same face in the next frame. The proposed algorithm is divided into the part that detects the face region in one frame and the part that recognizes the face of the same person. In particular, this was an attempt to overcome the restrictions of existing still learning methods by applying the most typical facial recognition methods dynamically. The effectiveness of the proposed method was verified through experimentation.

This paper is composed as the following. Section 46.2 mentions the methods for face recognition. Section 46.3 explains the facial recognition methods that we use and also algorithms proposed in this paper and Sect. 46.4 is on the testing and analysis to verify the proposed algorithm. Lastly, Sect. 46.5 is on the conclusion and future projects.

46.2 Face Region Detection and Face Learning

46.2.1 Adaboost and Haar-Like Algorithm

One of the most widespread methods for detecting objects such as faces is the detection method using the Haar-like feature. This method is an object classification method based on Adaboost [4, 5]. The Haar-like feature method was proposed by Viola and Jones. The features of the object detected by the haar-like feature is learned to create a week classifier, and then a strong classifier is made through the weak classifier. The feature value calculated through haar-like applies

the Adaboost algorithm to detect the object from the inputted video [5]. Using this, various classifiers and learners are combined to create one meta learner, which can overcome the weaknesses that occur due to learning methods. However, because multiple learners and classifiers are used there is a large time overhead and it is therefore difficult to apply in areas that require real-time face detection. Adaboost is normally used for generating one final classifier by combining simple but multiple weak classifiers quickly [6, 7].

46.2.2 Principle Component Analysis (PCA)

PCA algorithm, which is one of the most commonly used methods for face detection, is effectively used for reducing dimensions and extracting features through multi-dimensional feature vector and low-dimension feature vector [8, 9]. Using this, features between frames are extracted to recognize the same face. Face recognition that uses Eigenface acquired from each image tracks multiple faces in a certain interval of frames from a video.

46.2.3 Linear Discriminant Analysis (LDA)

Unlike the dimension reduction based PCA, LDA [10–14] is a method made so that objects with different classes can be easily separated. However, based on PCA the classification of classes is applied to the maximum to reduce the dimensions. Fisher-face [13], which is a common method of LDA, is a method to maximize sorting capacities for extraction. As the pre-treating process PCA is used and dispersion to different classes are made to be the maximum, while minimizing dispersion of the constituents within the same class in the spatial composition to determine the class by projecting the input image.

46.2.4 Local Binary Pattern Histogram (LBPH)

LBPH [15] was created in 2004 by a research team Oulu University as a recognition technology that can be successfully applied to face recognition and it uses histograms featured by the texture of the image. LBP is allotted by comparing the center of pixel value with pixel values. When the pixel value is larger than the center of pixel value, it is given 1 and if it is smaller, 0. The face image is partitioned into small regions and from each region, LBPH is extracted.

46.3 Face Labeling Algorithm (FLA)

46.3.1 Candidate Face Regions

There are a few steps to acquire face regions from the input frame. First of all, aforementioned Haar-like and Adaboost are used to extract candidate face regions from the input frame and expressed in C. At this time, R_i means the m by n matrix.

$$C = \{R_1, R_2, R_3, \ldots, R_n\} \tag{46.1}$$

The scope of elements need to be limited to the ones that are within a certain size because it is not sure if the elements in C are all within the face regions. When the set that meets this condition is defined as C_1 and radius of R_i is defined as r_i, C_1 should meets the following conditions.

$$C_1 = \{R_i | r_{min} \leq r_i \leq r_{max}\}, \quad C_1 \subseteq C \tag{46.2}$$

The dot rate which is the skin color in each element of R_i that belongs to C_1 is defined as s_i and the set of these values that meet some conditions is defined as C_2, and C_2 can be defined as follows.

$$C_2 = \{R_i | s_i \geq \text{Threshold}\}, \quad C_2 \subseteq C_1 \tag{46.3}$$

R_i in C_2 has some noisy elements including the face regions. Therefore, R_i' that excluded noisy regions from R_i can be defined as follows. In a matrix R_i, an element in row i and column j is represented a_{ij}, and a_{center} means the center element of R_i.

$$R_i' = \begin{cases} a_{ij}, & \text{if } |a_{center} - a_{ij}| < \textit{Threshold} \\ a_{ij}, & \text{if } |a_{center} - a_{ij}| \geq \text{Threshold} \cap a_{ij} \in \{\text{skin color}\} \\ \text{empty}, & \text{others} \end{cases} \tag{46.4}$$

The set C_3 that excluded the noisy region from C_2 can be expressed as follows and it is the final values for the labeling step.

$$C_3 = \{R_1', R_2', R_3', \ldots, R_n'\} \tag{46.5}$$

46.3.2 Continuous Labeling Between Frames

To say the faces learned in the previous frame are set D, D can be defined as follows. At this time, the subscript of d_i means label of the face.

$$D = \{d_1, d_2, d_3, \ldots, d_m\} \tag{46.6}$$

The faces observed from the input frame are defined as set F. It is expressed as follows.

$$F = \{f_1, f_2, f_3, \ldots, f_n\} \tag{46.7}$$

The similarity between the observed face f_i and d_j is defined as $S(f_i, d_j)$ and $S(f_i, d_j)$ is expressed as follows. At this time, the calculation methods for V_i are depending on the learning algorithm that are used.

$$S(f_i, d_i) = \{v_1, v_2, v_3, \ldots, v_m\} \tag{46.8}$$

The set of f_i that has more than a threshold maximum similarity with d_j is defined as $B(d_i)$, and $B(d_i)$ needs to meet the following conditions.

$$B(d_j) = \{f_i | \forall i, \mathrm{Max}(S(f_i, d_j)) \geq \mathrm{Threshold}\}, \quad B(d_j) \subseteq F \tag{46.9}$$

The label value of face f_i is defined as $L(f_i)$ and $L(f_i)$ needs to meet the following conditions.

$$L(f_i) = \begin{cases} d_{m+1}(\text{new label}), & \text{if } \forall j,\ B(d_j) = \emptyset \\ d_j, & \text{if } \forall j,\ \mathrm{Max}(B(d_j) = f_i \end{cases} \tag{46.10}$$

When a new label is provided to D, it performs $D = D \cup \{f_i\}$ operation and updates the value of D. In particular, if $B(d_i) = \emptyset$, it means that d_j is disappeared from the observed frame.

46.3.3 FLA Procedure

The FLA treatment method that continuously labels the face of the same person between frames in videos that include the faces of several people is divided into two main stages. The first stage is for determining the candidate region of the face in the input frame and the second stage is finding the facial similarity in the previous frame and that in the input frame to recognize the sameness. In this stage, for the face candidates that maintain sameness, it is labeled with the value of the previous frame, and for faces that are not, a new label is given and it is saved in the label database.

```
Procedure FLA
 Initialization: labeled_face_list=0
 Input: multi-frame image
 For each multi-frame image
   Initialization: upate_face_list=0,rank_face_list=0
   detect all face region
   For each all face region
     If labeld_face_list is empty
      add input face region to update_face_list
     Else
      get best similarity and label using face recognizer
    If best similarity < threshold
           add input face region to update_face_list
       Else
           Add input face region to rank_face_list
          Endif
       Endif
    Endfor
  If rank_face_list is not tempty
     sort rank_face_list using similarity and label
     do label to input face region with best label
   Endif
   If update_face_list is not tempty
     do label to input face region with new label
     Initialization: face recognizer
     add update_face_list to labeled_face_list
     learning labeled_face_list using face recognizer
   Endif
 Endfor
```

In other words, for the input frame (n), all possible face region candidates in frame (n) are detected. Also, based on the learned patterns of the face labeled after being recognized as a new face in frame (n − 1), which is the frame in the previous stage, this method tracks continuity by measuring the facial similarity in frame (n). Methods for calculating the similarity of two different faces differ according to the face recognizer being used.

46.4 Experimental Results

Dataset was collected for experiments. A camera was installed on a specific location that can view from top to bottom at the end of a long space like a hallway and the people passing by were filmed for 1–2 min. The number of people passing by were one to five people with five categories, and each person as a different category were collected five times for a total of 25 videos.

For the algorithm for detecting face area, the widely known Adaboost Harr-like method was used and skin color filters were applied to enhance accuracy. The face recognizers used were Eigenfaces, Fisherfaces, and LBPH. FLA was applied every five frames in the video and Fig. 46.1 shows part of the labeling results from the each surveillance frame.

Fig. 46.1 Frame face labeling result

Table 46.1 Labeling experimaental results

Face number	Test frame size	Accuracy rate (%)			Relearning time (ms)
		Eigenfaces	Fisherfaces	LBPH	
1	65	82	87	78	2.45
2	87	79	80	73	6.48
3	80	74	76	64	7.57
4	52	72	73	57	10.17
5	38	70	72	54	15.0
…	…	…	…	…	…

Table 46.1 is part of the test results that applied FLA in each video. The number of faces that applied simultaneously in the video and number of frames that applied FLA, as well as the total treating time and accuracy per face recognizer were observed. Figure 46.2 displays the average accuracy per face recognizer according to the face number in a graph format.

Labeling accuracy for N number of frames in FLA is defined as seen in the following formula:

$$\begin{aligned} \mathbf{L}_{\mathbf{accu}} &= \frac{1}{\boldsymbol{N}} \sum\nolimits_{f=1}^{n} \frac{\boldsymbol{Correct\ labeling\ face}}{\sum \boldsymbol{face\ of\ num}} \times \boldsymbol{100} \\ \mathbf{N} &= \sum\nolimits_{f=1}^{n} \boldsymbol{face} \end{aligned} \tag{46.11}$$

The experimental result was showed that LDA accuracy was most excellent. As to PCA, the error rate was high but the processing time was a little advantageous side in comparison with LDA. In the case of LBPH, when the face was one, the good performance was shown. However, the rapid performance degradation was shown with the number in which the number of face will be increased. The reason is that the recognition of LBPH about the front face is high but inclining angle of

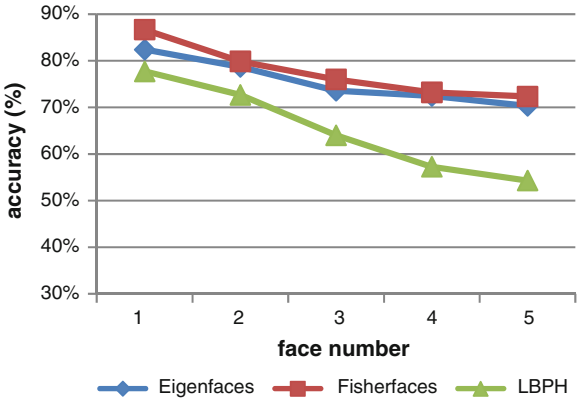


Fig. 46.2 Average labeling accuracy

the face is low. The main cause of the error detection is predicted as two in the whole view. Firstly it is the difference including the inclining angle of the face image, face size, light, and etc. Secondly it is the performance difference of the algorithm of detecting face region.

46.5 Summary and Conclusions

In this study, we proposed a continuous labeling method for the same face in between frames in a video that includes the faces of multiple people. Various face recognizers in use were applied in the experiment to check the usefulness of the pro-posed algorithm. Test results showed some significant results, but there are still some improvements needed.

When considering the FLA performance from the aspect of processing time and accuracy, the two factors have contradicting features, and therefore, a new face recognizer that can satisfy all is necessary. Moreover, research that applies FLA in various application fields is also necessary in the future.

References

1. Zhou F, Duh HB, Billinghurst M (2008) Trends in augmented reality tracking, interaction and display: a review of ten years of ISMAR. In: IEEE international symposium on mixed and augmented reality, pp 193–202
2. Bakowski A, Jones GA (1999) Video surveillance tracking using color region adjacency graphs. Image processing and its applications. Conference Publication No. 465, pp 794–798
3. Lee S-Y, Jung T-R, Hur C-W, Ryu K-R (2006) A study on the position tracking of moving image for surveillance system. In: KIMICS integrated conference, pp 205–208
4. Froba B, Ernst A (2004) Face detection with the modified census transform. In: IEEE international conference on automatic face and gesture recognition, pp 91–96

5. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of CVPR, pp 511–518
6. Schapire RE, Singer Y (1999) Improved boosting algorithms using confidencerated predictions. Mach Learn 37(3):297–336
7. Freund Y, Schapire RE (1999) A short introduction to boosting. J Japan Soc Artif Intell 14(5):771–780
8. Viola P, Jones M (2001) Robust real-time object detection. Compaq Cambridge research Laboratory (2001)
9. Turk M, Pentland A (1991) Eigenfaces for recognition. J Cogn Neurosci 3(1):71–86
10. Sayeed AM, Jones DL (1995) Optimal detection using bilinear time-frequency and time-scale representations. IEEE Trans Signal Process 43:2872–2883
11. Swets DL, Weng J (1996) Using discriminant eigenfeatures for image retrieval. IEEE Trans Pattern Anal Mach Intell 18:831–836
12. Belhumer PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces versus fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 9(7):711–720
13. Yeom S (2012) Multi-classifier decision-level fusion for face recognition. J Inst Electron Eng Korea 49-SP(4)
14. Lu J et al (2003) Face recognition using LDA-based algorithms. IEEE Trans Neural Networks 14(1):195–200
15. Lee K, Lee C (2013) Content-based image retrieval using LBP and HSV color histogram. JBE (J Broadcast Eng) 18(3):372–379

Chapter 47
Audio Classification Using GA-Based Fuzzy C-Means

Myeongsu Kang and Jong-Myon Kim

Abstract The purpose of automatic audio classification is to meet the rising need for efficient multimedia content management. This paper proposes a robust audio classification approach that classifies audio streams into one of five categories (speech, music, speech with music, speech with noise, and silence). The proposed method is composed of two steps: efficient audio feature extraction and audio classification using genetic algorithm-based fuzzy c-means. Experimental result indicates that the proposed classification approach achieves higher than 96.16 % in terms of classification accuracy.

Keywords Audio classification · Fuzzy c-means · Genetic algorithm · Multimedia content management

47.1 Introduction

Typical multimedia databases often contain large numbers of audio signals that require automatic audio retrieval for efficient production and management [1], and the efficacy of audio content analysis depends on extracting appropriate audio features and using an effective classifier to precisely classify the audio stream.

In literatures, many researchers have tried to utilize classifiers with effective audio features such as mel-frequency cepstral coefficients, and zero-crossing rate for audio classification. In addition, various classifiers such as the Gaussian mixture model and support vector machine have been utilized for audio

M. Kang · J.-M. Kim (✉)
School of Electrical Engineering, University of Ulsan, Ulsan, South Korea
e-mail: jongmyon.kim@gmail.com

M. Kang
e-mail: ilmareboy@gmail.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_47, © Springer Science+Business Media Dordrecht 2014

classification. Among many classifiers, a number of methods based on fuzzy c-means (FCM) have been proposed. Park et al. proposed different fuzzy methods in order to classify audio signals into different musical genres [2, 3]. In spite of the fact that FCM is an efficient for audio classification, the FCM-based classifiers exhibit performance degradation since FCM requires initialization [4]. To address this problem, this paper integrates FCM clustering with a genetic algorithm (GA) to globally optimize the objective function of FCM and offer better classification performance.

The rest of this paper is organized as follows. Section 47.2 presents audio features extraction based on principal component analysis and Sect. 47.3 introduces the proposed audio classification scheme. Section 47.4 analyzes experimental results and Sect. 47.5 concludes this paper.

47.2 Audio Features Extraction

The purpose of audio features extraction is to obtain as much information as possible about the audio streams. After rigorous studies of broad categories of audio features used in the previous studies, this paper extracts the following features to obtain accurate audio classification. Feature extraction is a frame-based process, and thus audio features are calculated in the processing window whose length is set to 0.02 s.

- Low Root-Mean-Square Ratio

Low root-mean-square ratio (L_R) is defined as the ratio of number of frames with root-mean-square (RMS) values below the 0.5-fold average RMS in the processing windows, as expressed in (47.1):

$$L_R = \frac{1}{2N}\sum_{m=0}^{N-1}\left[sgn\left(0.5 \times \overline{RMS} - RMS(m)\right) + 1\right], \tag{47.1}$$

where N is the total number of frames, m is the frame index, $RMS(m)$ is the RMS at the mth frame, $\overline{RMS}$ is the average RMS in the processing window, and $sgn(\cdot)$ is 1 for positive arguments and 0 for negative arguments. L_R is suitable for discriminating between speech and music because L_R is generally high for a speech signal that consists of words mixed with silence, whereas L_R for music is low.

- Spectrum Spread

Spectrum spread (SS) is effective for determining whether the power spectrum is concentrated around the centroids or is spread out over the spectrum. Music is composed of a broad mixture of frequencies, whereas speech consists of a limited range of frequencies. Consequently, the spectrum spread efficiently determines between speech and music. Its mathematical definition is given by

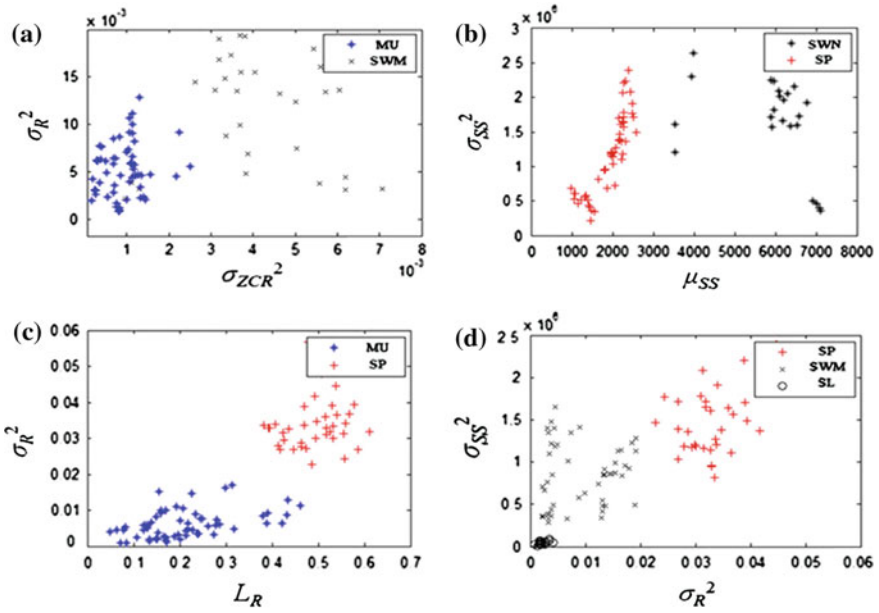


Fig. 47.1 Feature distribution based on $[L_R, \sigma_R^2, \mu_{SS}, \sigma_{SS}^2, \sigma_{ZCR}^2]$. Classification, **a** between music and speech with music, **b** speech and speech with noise, **c** music and speech, and **d** among speech, speech with music, and silence

$$SS(n) = \sqrt{\frac{\sum_{m=0}^{K-1}\left[(k - SC(n))^2 \times |A(n,m)|^2\right]}{\sum_{m=0}^{K-1}|A(n,m)|^2}}, \tag{47.2}$$

where K is the order of the discrete Fourier transform (DFT), m is the frequency bin for the nth frame, $SC(n)$ is spectral centroid at the nth frame, and $A(n, m)$ is the DFT of the nth frame of the given signal. $SC(n)$ and $A(n, m)$ are computed as $SC(n) = \frac{\sum_{m=0}^{K-1} m \times |A(n,m)|^2}{\sum_{m=0}^{K-1}|A(n,m)|^2}$, $A(n,m) = \sum_{k=0}^{N_{samples}-1} x(k)e^{-j\cdot\left(\frac{2\pi}{W}\right)\cdot k\cdot m}$, where $N_{samples}$ is the total number of samples in the audio stream.

- Zero-Crossing Rate

Zero-crossing rate (ZCR) value is defined as the number of zero-crossings within a processing window, as shown in (47.3):

$$ZCR(n) = \frac{1}{N_{samples} - 1}\sum_{m=0}^{N_{samples}-1} |sgn[x(m)] - sgn[x(m-1)]|, \tag{47.3}$$

where $x(m)$ is the value of mth sample in the processing window, and $sgn(\cdot)$ is a sign function as mentioned in (47.1). Voiced and unvoiced speech sounds have

low and high zero-crossing rates, respectively. This results in high *ZCR* variation, whereas music typically has low *ZCR* variation.

47.2.1 Feature Vector Configuration

Audio streams are classified into the following five categories: *silence* (*SL*), *speech* (*SP*), *music* (*MU*), *speech with music* (*SWM*), and *speech with noise* (*SWN*). According to our experiments, statistical values of *RMS*, *SS*, and *ZCR* represent the characteristics of target audio signals for classification well. Consequently, this paper finally selects five audio features [L_R, σ_R^2, μ_{SS}, σ_{SS}^2, σ_{ZCR}^2] for more accurate audio classification, as shown in Fig. 47.1.

47.3 GA-Based FCM for Audio Classification

Let an unlabeled data set $X = \{x_1, x_2,\ldots,x_n\}$ represent n number of features. The FCM algorithm sorts the data set X into c clusters. The standard FCM objective function with the Euclidian distance metric is defined as follows:

$$J_m(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^m d^2(v_i, x_k), \tag{47.4}$$

where $d^2(v_i, x_k)$ represents the Euclidian distance between the centroid v_i of the ith cluster and the data point x_k, and u_{ik} is the degree of membership of the data x_k to the kth cluster, along with the constraint $\sum_{i=1}^{c} u_{ik} = 1$. The parameter m controls the fuzziness of the resulting partition, with $m \geq 1$, and c is the total number of clusters. Local minimization of the objective function $J_m(U, V)$ is achieved by repeatedly adjusting the values of u_{ik} and v_i according to the following equations:

$$u_{ik} = \left[\sum_{j=1}^{c}\left(\frac{d^2(v_i,x_k)}{d^2(v_j,x_k)}\right)^{\frac{1}{m-1}}\right]^{-1}, \; v_i = \frac{\sum_{k=1}^{n} u_{ik}^m x_k}{\sum_{k=1}^{n} u_{ik}^m} \quad where\ 1 \leq i \leq c. \tag{47.5}$$

As J_m is iteratively minimized, v_i becomes more stable. The iteration of the FCM algorithm is terminated when the terminal condition $\max_{1\leq i\leq c}\left\{abs\left(v_i^t - v_i^{t-1}\right)\right\} < \varepsilon$ is satisfied, where v^{t-1} are the centroids of the previous iteration, *abs*() denotes the absolute value, and ε is the predefined termination threshold. Finally, all data points are distributed into clusters according to the maximum membership u_{ik}. As noted in Sect. 47.1, FCM starts with randomly initialized centroids, which has strong effects on its performance. To deal with this drawback, this paper employs a GA for obtaining more accurate classification performance. To do this, centroids are initially selected by GA, and these centroids

are used for calculating membership values of FCM. According to [5], suitable coding (representation of chromosomes) for the problem must be devised before GA is performed. Likewise, the fitness function, which assigns a figure of merit to each coded solution, is required. During the process, parents must be selected for reproduction, and combined to generate offspring.

47.3.1 Initialization

The initial population, which consists of randomly produced initial individuals, is generated, whose number is M_{cla}. We set the population size as $M_{cla} = 100$. Data are normalized within the range [0:1], and then each individual is encoded in the population such as $Chrom^i = \left\{v_{11}^i, v_{12}^i, \ldots, v_{15}^i, v_{21}^i, v_{22}^i, \ldots, v_{25}^i, \ldots, v_{51}^i, v_{52}^i, \ldots, v_{55}^i\right\}$, where $Chrom^i$ is ith individual in population, $i \in [0:(M_{cla} - 1)]$.

47.3.2 Generating GA Operators

In this process, genetic operators such as selection, crossover, and mutation are set.

- *Selection*: The stochastic universal sampling method is utilized to select potentially useful chromosomes for recombination.
- *Crossover*: This paper selects an intermediate recombination technique which is suitable for use with real-valued code. In this method, different variable (or dimension) values of offspring are chosen somewhere near the values of parents. If there is a population of N_{var} dimension data in each individual, offspring are produced according to the following rule:

$$Var_i^0 = Var_i^{P_1} \times a + Var_i^{P_2} \times (1 - a), \tag{47.6}$$

where Var_i^0 indicates the value of the ith dimension of offspring, and $Var_i^{P_1}$ and $Var_i^{P_2}$ are values of the ith dimension of the first and second parents, respectively, and $i \in (1,2, \ldots, N_{var})$, $a \in [-d, 1 + d]$. Here, a is a scaling factor that is chosen by random over an interval $[-d, 1 + d]$ for each variable. The value of the parameter d defines the size of the area for possible offspring. In this paper, we set $d = 5$ because the number of features is five, and $N_{var} = 25$ because we classify audio data into five clusters with five features.

- *Mutation*: Individuals are randomly altered by mutation. These variations (mutation steps) are mostly small. The probability of mutating a variable is inversely proportional to the number of variables (dimensions). If one individual has more dimensions, the possibility of mutating a variable becomes smaller. In general, a mutation rate of $1/N_{var}$ has produced good results for a wide variety of objective functions. For audio classification, the mutation rate is set to 0.04 since

N_{var} is set to 25 for this paper. Furthermore, the objective function using (47.4) is computed in this process, and the fitness assignment is based on the objective function as follows:

$$f_i = \frac{1}{\sum_{l=1}^{c} \sum_{k=1}^{n} u_{ik}^{m} d^2\left(v_l^i, x_k\right)}. \tag{47.7}$$

47.3.3 Checking Termination Criterion

The optimization criterion is checked if $abs\left(f_{best}^{i} - f_{best}^{i-1}\right) \leq T_{clafinal}$ is satisfied or not, where f_{best}^{i} and f_{best}^{i-1} are the best fitness values for chromosomes in ith (current) and $(i-1)$th (previous) generations, respectively. Moreover, $T_{clafinal}$ is the predefined termination threshold. Therefore, if the terminal criterion is satisfied, we move on to the classification process. Otherwise, we turn back to *Generating GA operators* step.

47.3.3.1 Training and Classification

Based on the training process, we determine the training centroid of audio types $C = \{c_1, c_2, \ldots, c_5\}$ and define clusters based on minimum distance from the training centroid. We then utilize the highest membership of each data point in order to classify audio streams into proper clusters: *SP*, *MU*, *SWM*, *SWN*, and *SL*. For the training process, this paper includes 15 pieces of silence, 15 pieces of pure speech, 15 pieces of music which involve various musical instruments, 15 pieces of speech with music, and 15 pieces of outside interviews that are composed of different background noise levels.

47.4 Experimental Results

For audio classification simulation, we utilize two datasets composed of Korean news broadcasts obtained from Ulsan Broadcasting Corporation (www.ubc.co.kr). The first dataset was used for testing and the other dataset was used for training. We employ GA-based FCM to classify audio streams. In this experiment, the degree of the fuzziness and the termination condition for GA-based FCM were set to 2 and 0.001, respectively, because Bezdek et al. experimentally determined the optimal intervals for the degree of fuzziness and termination threshold, which range from 1.1 to 5 and 0.01 to 0.0001, respectively [6]. To evaluate classifications, we utilize the standard of correctness that has been widely accepted in recent studies, which is as follows:

Table 47.1 Classification result of the proposed method

		Classified					Correctness (%)
		SP	*MU*	*SWM*	*SWN*	*SL*	
Input audio stream	*SP*	54	0	0	0	0	100.0
	MU	0	6	0	0	0	100.0
	SWM	0	0	5	0	0	100.0
	SWN	2	0	0	21	0	91.3
	SL	6	0	0	0	51	89.5

$$Correctness = \frac{\#\ of\ correctly\ classified\ audio-segments}{\#\ of\ all\ audio-segments} \times 100\,(\%). \quad (47.8)$$

Table 47.1 presents classification results of the proposed approach and several misclassified results are obtained when attempting to distinguish between speech and silence, because some silence signals include speech components at the beginning and end. Likewise, misclassifications of speech with noise signals occurred mostly when the amplitudes of noise components were small and unclear.

47.5 Conclusion

This paper proposed a robust audio classification approach to address the rising demand for efficient multimedia content management. To classify audio streams into one of five categories (speech, music, speech with music, speech with noise, and silence), this paper explored 19 audio features. Among these audio features, this paper selected the five most suitable features such as $[L_R, \sigma_R^2, \mu_{SS}, \sigma_{SS}^2, \sigma_{ZCR}^2]$, and utilized these extracted features as inputs of GA-based FCM for audio classification. Our experimental results showed that the proposed classification method achieved very accurate classification performance.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. NRF-2013R1A2A2A05004566 and NRF-2012R1A1A2043644).

References

1. Foote J (1999) An overview of audio information retrieval. Multimedia Syst 7(1):2–10
2. Park D-C, Tran CN, Min B-J, Park S (2006) Modeling and classification of audio signals using gradient-based fuzzy C-means algorithm with a Mercer Kernel. Lect Notes Comput Sci 4009:1104–1108
3. Park D-C (2009) Classification of audio signals using fuzzy C-Means with divergence-based Kernel. Patt Recogn Lett 30(9):794–798

4. Cheng WS, Ji CM, Liu D (2009) Genetic algorithm-based fuzzy cluster analysis for flood hydrographs. In: International workshop on intelligent systems and applications, Wuhan, pp 1–4
5. Beasley D, Bull DR, Martin RR (1993) An overview of genetic algorithms: Part 1. Fundam Univ Comput 15:58–69
6. Bezdek JC, Keller J, Krisnapuram R, Pal N (2005) Fuzzy models and algorithms for pattern recognition and image processing. Springer, Berlin

Chapter 48
Anonymous Editing Prevention Method for Depth-Image-Based Rendering Stereoscopic Video

Ji-Won Lee, Hae-Yeoun Lee and Sang-Hyun Joo

Abstract Although many stereoscopic video watermarking techniques are being currently produced, these cannot be directly applied to stereoscopic video formatted in depth-image-based rendering (DIBR) due to the innate characteristics of DIBR. To rectify this situation, we first investigate technical issues in the design of watermarking schemes. We then propose a stereoscopic video watermarking method for DIBR formatted content to resolve the technical issues. The proposed method is robust against both 2D-3D conversion and frequently occurring distortions. Further, the imperceptibility is high enough to minimize visual fatigue, which is one of the most sensitive issues when it comes to stereoscopic video.

Keywords DIBR · Stereoscopic watermarking · Video watermarking · 2D-3D conversion

48.1 Introduction

There are two formats are widely used to represent the left and right views of stereoscopic content: stereo image recoding (SIR) and DIBR [1]. The SIR format stores both left and right views directly using the Side-by-side or Top–bottom

J.-W. Lee (✉) · S.-H. Joo
Next-Generation Visual Computing Research Section, Creative Content Research Laboratory, Electronics and Telecommunications Research Institute (ETRI), 218 Gajeong-ro, Yuseong-gu, Daejeon 305-700, Republic of Korea
e-mail: ez1005@etri.re.kr

S.-H. Joo
e-mail: joos@etri.re.kr

H.-Y. Lee
Department of Computer Software Engineering, Kumoh National Institute of Technology, Sanho-ro 77, Gumi, Gyeongbuk, Republic of Korea
e-mail: Haeyeoun.lee@kumoh.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_48, © Springer Science+Business Media Dordrecht 2014

method. Unlike SIR, DIBR consists of a monoscopic center view and an associated per-pixel depth map. Using these two components and given depth information from the viewer, left and right views are rendered by DIBR. DIBR has many merits that SIR does not have. Since human eyes are insensitive to small changes in depth level, a depth map can be compressed more highly than a monoscopic view. Moreover, since DIBR formatted content renders to left and right views using the center view and the depth map, viewers can configure the depth environment of synthesized view.

In spite of the wide usage of DIBR video, there exists one big problem to be solved in order to copyright-protect DIBR. Not only the center view but also the rendered left and right views can be illegally distributed as monoscopic content. Thus, we have to protect all three individual views in order to prevent illegal distribution. In this situation, the watermark information in the center view should be preserved in the rendering process. In addition, the watermarked view should be robust enough against commonly occurring distortions when the illegally leaked content is distributed on an unsecure transmission channel. Therefore, we need to design a stereoscopic watermarking technique that is robust against not only 2D-3D conversion, but also commonly occurring distortions.

48.2 The Proposed DIBR Watermarking Method

48.2.1 Technical Issues for DIBR Video Watermarking

Since not only the center view, but also the synthesized left or right view, can be used as monoscopic content, we have to protect each of these views from illegal distribution. There are three considerations in designing the stereoscopic watermarking for DIBR 3D video:

- The embedded watermark should survive DIBR 2D-3D conversion.
- The embedded watermark should be robust against scaling attack which is most frequently occurring when a leaked monoscopic video is distributed.
- The embedded watermark should be imperceptible.

48.2.2 Watermark Embedding

In the watermark embedding process, we need to design two parts by using two different methods: a method of inserting the watermark signal spatially into each single frame of the given video and a method of embedding the watermark information temporally into the given video. Watermark information can be represented as a sequence of meaningful watermark signals.

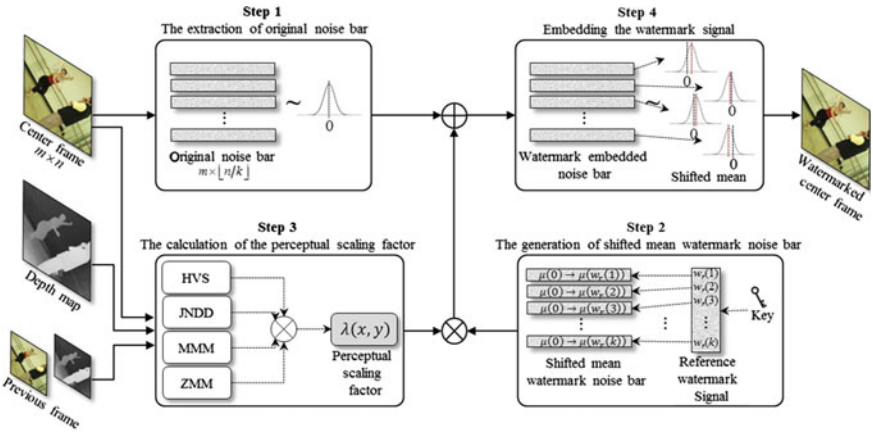


Fig. 48.1 Outline of the spatial embedding part

The spatial embedding process

Figure 48.1 illustrates the outline of the spatial embedding process. The details for each step are as follows:

Step 1 Extraction of original noise bar

When the original current center frame F_c whose resolution is $m \times n$ is given, we extract the noise image by subtracting the denoised center frame $\overline{F}_{curr}$ from F_c. We then horizontally slice the extracted noise image up the k noise bars with resolution $m \times \lfloor n/k \rfloor$, where k is the dimension of reference watermark signal w_r and $k \leq n$. The ith sliced original noise bar refers to nb_o^i.

$$F_c = \overline{F}_{curr} + \sum_{i=1}^{k} nb_o^i. \tag{48.1}$$

Step 2 Generation of shifted mean watermark noise bar

Next, we generate the shifted mean watermark noise bar nb_{sm}^i that will be added to nb_o^i. First, w_r, which follows Gaussian distribution with zero mean and constant variance c_r, is generated. We then generate k shifted mean watermark noise bars with resolution $m \times \lfloor n/k \rfloor$. Each nb_{sm}^i contains the ith information of w_r, denoted $w_r(i)$, by shifting the mean of the noise bar to be $w_r(i)$:

$$nb_{sm}^i(x,y) = \left(N^{(0,c_{sm})}(x,y) + w_r(i)\right) \times \lambda(x,y) \tag{48.2}$$

Here, $N^{(0,c_{sm})}(x,y)$ is a random variable that follows Gaussian distribution with zero mean and variance c_{sm}. $\lambda(x,y)$ is the perceptual scaling factor.

Step 3 Calculation of the perceptual scaling factor

In this step, we calculate $\lambda(x,y)$. We then find the product of it and each nb^i_{sm} to achieve better imperceptibility and to alleviate visual fatigue. In the proposed method, we exploit four perceptual masks to calculate $\lambda(x,y)$.

$$\lambda(x,y) = b_l + \frac{\left(\varphi(x,y) - \min_{i,j}(\varphi(x,y))\right) \times (b_u - b_l)}{\max_{i,j}(\varphi(x,y)) - \min_{i,j}(\varphi(x,y))} \tag{48.3}$$

where

$$\varphi(x,y) = \eta(x,y) \times \kappa(x,y) \times \phi_m(x,y) \times \phi_z(x,y) \tag{48.4}$$

Here, b_u and b_l respectively indicate the upper bound and the lower bound of $\lambda(x,y)$. $\eta(x,y)$ is a local weighting factor of the pixel from the noise visibility function (NVF) [2] and $\kappa(x,y)$ is a local weighting factor using the just noticeable difference in depth (JNDD) [3]. $\phi_m(x,y)$ and $\phi_z(x,y)$ represent the monoscopic motion mask (MMM) and z-axis motion mask (ZMM) [4], respectively.

The formulas of NVF, JNDD, and ZMM can be found in [2–4]. The monoscopic motion mask $\phi_m(x,y)$ can be defined by using the difference between the previous frame and the current frame as follows:

$$\phi_m(x,y) = \begin{cases} b_l, & \text{if } \left|\mu_l\left(F_p(x,y)\right) - \mu_l(F_c(x,y))\right| \le 3 \\ b_u, & \text{if } \left|\mu_l\left(F_p(x,y)\right) - \mu_l(F_c(x,y))\right| > 30 \\ 1, & \text{otherwise} \end{cases} \tag{48.5}$$

Here, the size $l \times l$ is 3×3 and F_p is denoted as a previous center frame.

The calculated $\lambda(x,y)$ are then multiplied with each nb^i_{sm} to make the watermark signal more imperceptible and to minimize the visual fatigue.

Step 4 Embedding the watermark signal

The final step is the embedding of each nb^i_{sm} into each nb^i_o. By adding nb^i_{sm} to nb^i_o, we can generate the watermark embedded noise bar nb^i_{we}. that has a shifted mean. Finally, nb^i_{we} is inserted into $\overline{F}_{curr}$ to make the watermarked frame F'_{curr}.

$$F'_{curr} = \overline{F}_{curr} + \sum\nolimits_{i=1}^{k} nb^i_{we} = \overline{F}_{curr} + \sum\nolimits_{i=1}^{k} \left(nb^i_o + nb^i_{sm}\right). \tag{48.6}$$

The temporal embedding process

In this process, we design the method to insert the watermark information along with the temporal axis of the given video content. Generally, every frame in a video content undergoes the same distortions. In this situation, we exploit the self-synchronizing signal [5] in this process. Detailed technique is in [5].

48.2.3 Watermark Extraction

Step 1 Extraction of embedded signal

When a target video $\tilde{V}$ is given, we first estimate the embedded signal from each frame. We extract the noise image $\tilde{F}_n$ from each frame $\tilde{F}$. After extracting $\tilde{F}_n$, we slice it into k horizontal noise bars $\widetilde{nb}^i_{we}$. We then calculate the mean of each $\widetilde{nb}^i_{we}$ to extract the estimated watermark signal w^j_e of the jth frame.

Step 2 Estimation of reference information and watermark information

After extraction of w^j_e from each frame, we accumulate them during the predefined time interval to estimate the reference information $\tilde{R}$ and watermark information $\tilde{W}$. During the pre-defined interval for $\tilde{R}$ and $\tilde{W}$, we can calculate the reference watermark signal $\tilde{w}_r$ and each watermark element $\tilde{w}_e$.

Step 3 Determination of embedded watermark information

To determine the embedded $\tilde{W}$, the normalized correlation Z_{nc} between $\tilde{w}_r$ and $\tilde{w}_e$ is calculated. In general, $\tilde{w}_r$ and $\tilde{w}_e$ have the same trace of certain distortions. Thus, the calculated Z_{nc} has high absolute value even if the target video $\tilde{V}$ is subjected to frequently occurring distortions such as resizing. If the absolute value of $Z_{nc}(\tilde{w}_r, \tilde{w}_e)$ exceeds the threshold, we determine each element of the watermark information by using the sign of Z_{nc}.

48.3 Experimental Results

In this section, we evaluate the fidelity and performance of the proposed method. In the experiments, the two DIBR formatted videos were used [6].

48.3.1 Fidelity Test

To measure the visual quality of the proposed method, we carried out objective and subjective fidelity tests. For the objective analysis, we calculated the PSNR, SSIM, and VQM between the original center, left, and right views and their watermarked views. For the subjective analysis, we adopted the DSCQS method and eight expert observers participated in the test.

The experimental results for the fidelity test are shown in Table 48.1. As shown in the table, both fidelity analysis results were good enough. These results confirm that the proposed method has good fidelity both for each monoscopic view and rendered stereoscopic view.

Table 48.1 Visual quality results of objective and subjective analysis

	Video clip 1 (ballet)				Video clip 2 (breakdancer)			
	C	L	R	S	C	L	R	S
PSNR	47.875	48.083	47.974	–	48.411	48.503	48.489	–
SSIM	0.9964	0.9963	0.9965	–	0.9968	0.9968	0.9948	–
VQM	0.4081	0.4078	0.4063	–	0.4334	0.4292	0.4240	–
MOS	4.9	5.0	4.9	4.9	4.7	4.8	4.8	4.8

C Center, *L* Left, *R* Right, *S* Stereoscopic view

Table 48.2 Comparison results of the proposed method and previously proposed schemes

Type of attack		2D-3D conversion	Format and frame rate conversion	Compression	Gaussian noise addition	Resizing
Proposed method	Center	–	○	○	○	○
	Left	○	○	○	○	○
	Right	○	○	○	○	○
Zhu's scheme [7]	Center	–	○	△	△	×
	Left	○	○	△	△	×
	Right	○	○	△	△	×
Lee's scheme [4]	Center	–	○	△	△	×
	Left	×	×	×	×	×
	Right	×	×	×	×	×

- Do not consider, ○ Possible, × Impossible, △ Partially possible

48.3.2 Robustness Test

To verify the robustness of the proposed method, we tested the proposed method and previous DIBR video watermarking schemes under various distortions. Table 48.2 shows the comparison results. The proposed method extracts the embedded watermark information stably under various distortions, whereas other schemes did not show enough tolerance against all distortions.

As shown in the results, the performance of the proposed method is better than that of previous schemes.

48.4 Conclusion

In this paper, we presented a robust stereoscopic video watermarking scheme for DIBR formatted content. We noted that not only the center view from the DIBR but also the left and right views rendered using 2D-3D conversion are illegally distributed, and that there is no technique to protect each monoscopic video in a blind way.

To rectify this situation, we first identified three technical issues that the DIBR video watermarking scheme should consider. Then, we designed the proposed watermarking scheme by considering the technical issues. To resolve all the issues, we employed three techniques: horizontal noise mean shifting, a self-synchronizing signal, and four different perceptual masks.

The proposed method proved highly robust and had high imperceptibility, as can be seen from the experimental results. Moreover, the proposed method outperforms previous DIBR watermarking schemes. However, there are a few drawbacks that need to be resolved. The time complexity of the proposed method is somewhat high when the perceptual masks are calculated.

Acknowledgments This research was supported by the ICT Standardization program of MISP (The Ministry of Science, ICT & Future Planning).

References

1. Jang GJ, Kim KH, Park IK (2004) Depth image-based rendering of 3D object on mobile device. In: Proceedings of intelligent signal processing and communication systems, Seoul, Korea, pp 630–633, Nov 2004
2. Voloshynovskiy S, Herrigel A, Baumgaertner N, Pun T (1999) A stochastic approach to content adaptive digital image watermarking. In: Proceedings of information hiding, Dresden, Germany, pp 211–236, Sept 1999
3. Silba VD, Fernando A, Worrall S, Arachchi HK, Kondoz A (2011) Sensitivity analysis of the human visual system for depth cues in stereoscopic 3D displays. IEEE Trans Multimedia 13(3):498–506
4. Lee MJ, Lee JW, Lee HK (2011) Perceptual watermarking for 3D stereoscopic video using depth information. In: Proceedings of intelligent information hiding and multimedia signal processing, Dalian, China, pp 81–84, Oct 2011
5. Oh TW, Lee MJ, Kim KS, Lee HY, Lee HK (2010) Robust high-definition video watermarking based on self-synchronizing signals against composite distortions. Opt Eng 49(9):1–14
6. Zitnick CL, Kang SB, Uyttendaele M, Szeliski R (2004) High-quality video view interpolation using a layered representation. ACM Trans Graphics 23(3):600–608
7. Zhu N, Ding G, Wang J (2008) A novel digital watermarking method for new viewpoint video based on depth map. In: Proceedings of intelligent systems design and applications, vol 2. Kaohsiung, Taiwan, pp 3–7, Nov 2008

Chapter 49
Different Knowledge Modeling Approach for Clinical Decision Supporting Service in u-Health Context

Jeong Ah Kim

Abstract The demand for a clinical decision supporting (CDS) system, which enables safe treatment and effective service, in the field of medicine is increasing with an increase in the demand for u-health care. In this paper, we suggested a knowledge authoring method required for CDS service application and a service development reference model based on the proposed method.

Keywords u-health · Knowledge authoring · Reference architecture model

49.1 Introduction

The demand for a clinical decision supporting (CDS) system, which enables safe treatment and effective service, in the field of medicine is increasing with an increase in the demand for u-health care. The CDS system helps a doctor during a treatment to decide on the diagnosis and treatment protocol and helps improve medical services by applying verified knowledge based on the existing accumulated patient information at the time of treatment.

For developing the clinical guidelines, several different approaches for modeling the clinical guideline have been proposed. Arden syntax [1], EON [2], PRODIGY [3], GUIDE [4], GLIF [5], and SAGE [6] (Standard-based Sharable Active Guideline Environment) are the most popular modeling formalism. With these guideline modeling formalisms, knowledge engineers specify the decision model and data for each clinical practice of point of care services or other administrative services. In u-health context, workflow and time-related knowledge representation is important but existing knowledge representations don't mention it.

J. A. Kim (✉)
Kwandong University, 522 NaeKok Dong, Kangnung, Gangwondo, South Korea
e-mail: clara@kd.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_49, © Springer Science+Business Media Dordrecht 2014

There are studies that propose the architecture required for implementing a u-health information system, such as [7] and [8]; however, these studies do not consider integration with a CDS service for customized treatment and a medical knowledge authoring method required for u-health service. Therefore, in this paper, we propose a knowledge authoring method required for CDS service application and a service development reference model based on the proposed method.

49.2 Requirements of Knowledge in Context of u-Health Portal

49.2.1 u-Health Portal

The u-health portal is a health management and medical service that monitors its users remotely anytime and anywhere by using wired and wireless networking technologies related to e-health. The e-health technology is defined as technology that uses modern information communication skills for fulfilling the necessities of citizens, patients, medical experts, medical service providers, and policy makers.

A care type in an e-health service can be defined as the application aiming at health management and disease management by using the Internet. The care type enables the development of self-treatment and expert systems through a model development with ubiquitous devices, consequently enabling health management and prevention or management of various diseases, including chronic diseases. The functions added to u-health for these purposes are self-treatment and treatment adjustment using measured information from the devices. In order to realize these services, the connection with a CDS system, which enables an exchange of patient information managed by a hospital and allows the medical staff and expert systems to intervene, is indispensable. The CDS system corresponds to a safety device that can protect patient safety and medical staff and support the intervention required in a clinical trial on the basis of completely verified knowledge. Such a system has become increasingly applicable to a real clinical system as medical staff has participated in the development of the CDS system from the initial stage and improved the accuracy of the medical service by providing its opinions to the CDS system during operations.

49.2.2 Modeling Guideline for Time-Related Knowledge

We identified three use cases for multi-entry knowledge representation [9]. (1) Reentrant input without time-related, (2) reentrant input with time-related but just one more time, (3) reentrant input with time related and many times.

1. **Reentrant input without time-interval**. In this case, guideline execution should suspend and guideline requests additional input values.
2. **Reentrant input with time-interval but just one more time**. In this case, guideline execution should be delayed until additional input values are provided after specific time interval. Additional inputs are normally new measurement in specific interval time. Guideline should notice the client how much interval is required for re-measurement.
3. **Reentrant input with time-interval and several times repeatedly**. In this case, guideline execution should be delayed until additional input values meet the specific goal value in the specific time interval. To encode the several times iteration of re-measurement. To encode this, we can apply guideline of reentrant input without time-interval and several times repeatedly.

49.3 Design of u-Health Portal and Knowledge Encoding Method

49.3.1 u-Health Portal Service Model Connected to CDS

For reinforcing a care-type service, self-treatment services, which a patient can carry out on the basis of verified information, should be added. Further, a self-treatment service can be provided by using a customized CDS based on the completely verified medical knowledge. u-health is a medical service provided to general patients; therefore, the CDS for u-health is designed for not only specialists but also general patients. Consequently, it is important to consider the level of the service user when designing the service or a proposal or recommendation customized with respect to patients.

49.3.2 Knowledge Authoring Method for u-Health Portal and CDS Connection

Let's consider weekly report of nourishment for patient in hypertension. Requirements define 3 kinds of input values and 3 output message fields. With these requirements, knowledge structure can be same but different value for input criterion or different values in same message expression. For example, same comparison criterion should be repeated for each month, first criteria is based on the how much calories patient took (<70, <100, >100 %) and second criteria is time interval for recoding (>6D, <6D). Also, in output message, almost message structure is same but specific value obtained from other rule execution can be different so that different message expression should be suggested to patients. This requirement is very easy for modeling but encoding result is so huge with

redundant activity graphs. For this, we suggested new knowledge encoding scheme. If controls variables, the kinds of comparison and the suggestion message structure are same but condition or value in messages are the only difference, parameterized function is better way to encode. Parameterized function makes different conditional expressions and different messages according to the criterion for input condition and values for message. With this parameterized functions the number of activity node can be decreased.

49.3.3 System Architecture for u-Health Portal and CDS Connection

A CDS primarily consists of a portal system, a CDS system, and a hospital information system. The portal system loads the modules that are responsible for services like content management from the portal, portal user management, and portal system application log management. The customized patient service, medical staff service, content management, data management, general user service, and user management service are general domains that can be changed depending on the service range and the contents of u-health.

The system application log management service is a service that must be developed to save general patient feedback and the medical staff feedback in order to analyze the effects gained from the application of CDS irrespective of the contents of the u-health service and to understand the need for service improvement and knowledge improvement. The CDS system domain is equipped with an engine for processing information provided to the portals. The hospital information system domain is a domain that loads the existing hospital information system, including the database that has patient information. The domains considered in this research for the development of the u-health portal are CDS system domain, hospital information system connection, and u-health device data collection module.

49.4 u-Health Portal Application Case and Evaluation

The u-health portal system for metabolic syndrome management was implemented to confirm the site application possibility of the architecture and the CDS service based the u-health system development process proposed in this paper. The knowledge was created according to the knowledge authoring method for the integration with the u-health portal, and the hospital connection module was developed and applied to the actual clinic.

The portal was implemented such that a user could verify the result when the CDS knowledge was executed on the basis of the patient providing information in three main areas—blood sugar, exercise, and nutrition management—for the implemented portal. We did not consider with the web application part in the

calculation of the overall performance of the portal because it was not significantly different from the existing web applications. However, we mainly evaluated the performance of the CDS system. The processing speed of the CDS system was analyzed on the basis of 8,887 response times, and the average response time was 285.638 ms even when the CDS system used for the metabolic syndrome u-health portal system was running in a PC environment. This shows that the system performance was sufficiently stable to provide services to a remote patient, even though we implemented the u-health architecture with added CDS based on knowledge obtained by using the proposed authoring method.

Also, we applied our approach to hypertension, metabolic syndrome, diabetes mellitus, and complication management guidelines for ubiquitous health care services of B electronics company. We have provided u-health services with CDS service for 6 months, average usage frequency of system in a month was 100. Based on these usages, service and knowledge have been improved. During these periods, the numbers of change requests were 45, 40 for service upgrades or updates, just 5 changes for knowledge upgrades. Since clinical knowledge is separated from the application codes, it was easy to analyze the change effects so that the average time for changing the application modules was 2 h, longest one was 4 h and shortest one was under 30 min.

References

1. Jenders RA, Corman R, Dasgupta B (2003) Making the standard more standard: a data and query model for knowledge representation in the Arden syntax. In: AMIA annual symposium proceedings, pp 323–330
2. Tu SW, Musen MA (2001) Modeling data and knowledge in the EON guideline architecture. In: Proceedings of MedInfo 2001, London, UK, pp 280–284
3. Johnson P, Tu S, Jones N (2001) Achieving reuse of computable guideline systems. Medinfo 10(Pt 1):99–103
4. National Institute for Health and Clinical Excellence (2009) The guidelines manual 2009. http://www.nice.org.uk
5. Boxwala AA, Peleg M, Tu S et al (2004) GLIF3: a representation format for sharable computer-interpretable clinical practice guidelines. J Biomed Inform 37(3):147–161
6. Tu S, Galsgow J (2006) SAGE guideline model: technical specification, SAGE Consortium
7. Hong JI (2005) An architecture for privacy-sensitive ubiquitous computing. Ph.D. thesis, University of Berkeley
8. Otto C et al (2006) System architecture of a wireless body area sensor network for ubiquitous health monitoring. J Mobile Multimedia 1(4):307–326
9. Kim JA et al (2012) Knowledge modeling guideline for reentrant features in SAGE, In: International conference on information technology and computer science (ITCS 2012), Porto, pp 76–81

Chapter 50
Formal Verification for Inter-Partitions Communication of RTOS Supporting IMA

Sachoun Park and Gihwon Kwon

Abstract The IMA (Integrated Modular Avionics) architecture is widely used to support multi avionics applications and execute those applications independently. It is important to ensure the fault containment and ease of verification and certification in IMA. However during the inter-partitions communication, because it is performed by copying a message between kernel memory areas, it is possible to break the wall to prevent any partitioned function from causing a failure in another partitioned function. In this paper, we show the possibility of the error propagation in the IMA and verify the properties in the inter-partition communication module of Qplus-653 kernel.

50.1 Introduction

In the recent aircrafts such as A380, IMA architecture is used and multiple software functions of different criticality level were integrated on single avionic computing devices in order to keep the volume, weight, power consumption and cost of avionic within reasonable limits. Before being applied IMA in this field, the classical concept "one computer has one function" was adopted, in which any faults could be put in a single computer. However, multi-function integration on a

This work was supported in part by the GRRC (Gyeonggi Regional Research Center) program of Gyeonggi Province (2013-0548, Developing and Industrializing Core-Technologies for Social Services Convergence Platforms).

S. Park
Electronics and Telecommunications Research Institute, Daejeon, South Korea
e-mail: sachoun@etri.re.kr

G. Kwon (✉)
Department of Computer Science, Kyonggi University, Suwon, South Korea
e-mail: khkwon@kyonggi.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_50, © Springer Science+Business Media Dordrecht 2014

single processor leads to in-transparent fault propagation, which significantly turns the reliability of the controllers down and maintenance cost up [1].

To deal with this problem, ARINC 653 Standard [2] was made by the AEEC (Airlines Electronic Engineering Committee), which is a software specification for space and time partitioning in Safety–critical avionics Real-time operating systems. It allows host multiple applications of different software levels on the same computer in the context of IMA architecture. In this specification, each function, application software, is called a partition which has its own memory space and time slot. In the concept of temporal partitioning, each partition is statically carried out in accordance with the predetermined order, however within a partition, multitasking is allowed. The spatial partitioning of each application is performed by partitioning the main memory area not to invade each other. The concept of these partitions facilitates the certification and plays a role in blocking error propagation from one partition to another [3].

Although each application is partitioned, there are some needs for the communication between them. For example, the information of flight control partition may send to the cockpit display to be visible, or values from each sensor device should be passed to mission and flight control partitions. Sometimes, the transmission of the data may be performed by copying from the memory area of a partition to the kernel memory area. Unfortunately, in the process of the transmission, it is possible to pass errors that lead to corrupt the concept of spatial partition. So it is important to formally verify the inter-partition communication at kernel level.

We developed Qplus-653 which is real-time operating system complied with ARINC 653, which has been certified with the DO-178B [4] level A to show the safety of it. As well known, the verification process of the DO-178B certification is based on the exhaustive testing and the review process. However, the review tends to be performed manually and it is hard to detect corner-stone errors even experts though. In this sense, the formal verification of the inter-partitions communication at the kernel level is proper to guarantee the safety. Contributions of this paper are follows:

- We found the error-prone part in the inter-partition communication when developing the operating system based on ARINC 653 (Sect. 50.2).
- We performed formal verification of our Qplus-653 kernel code with some safety requirements. In the verification SPIN model checker [5] was used and C code embedded in the PROMELA model (Sect. 50.3).

50.2 Inter-Partitions Communication of ARINC 653

50.2.1 The ARINC 653 APEX (Application/Executive)

Qplus-653 is a real-time operating system for safety–critical system and avionics embedded systems. The Qplus-653 provides a general-purpose APEX (Application/Executive) interface between Operating System (OS) of avionics computer

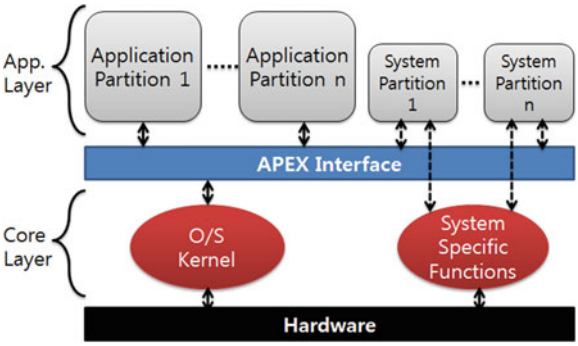


Fig. 50.1 Core module component relationships

resources and application software based on ARINC 653. Figure 50.1 depicts the core module component relationships. The software that resides on the hardware platform consists of:

- Application Partitions: the partitions of software specific to avionics applications supported by the core module.
- OS kernel: the API and behaviors defined within ARINC 653 specification, and supports a standard and common environment in which application software executes.
- System partitions: partitions outside the scope of APEX services.
- System Specific Functions: hardware interfaces such as device drivers, down loading, debug and built-in-test functions.

At the core module level, the OS manages partitions and their inter-partitions communication, the latter being conducted either within or across module boundaries.

A major part of ARINC 653 is the definition of communication between partitions. Its primary definition is communication between two of more partitions executing either on the core module or on different core modules. It may also mean communication between APEX partitions on a core module and non-ARINC 653. But the communication within a core module is the only consideration in this paper.

All inter-partition communication is conducted via messages, which are sent from a single source to one or more destinations. The basic mechanism for linking partitions by messages is the channel. A channel defines a logical link between one source and one or more destinations, where the source and the destinations may be one or more partitions. At the system configuration, channels, ports, and message type should be defined.

50.2.2 Inter-Partition Communication of the Qplus-653

Qplus-653 provides two types of port for user program, queuing port and sampling port. Queuing ports have a buffer to store data transferred and the OS kernel keeps

the data and sends to other partitions. Sampling port only keeps the up-to-date data and the kernel transfers the data to other partitions. Qplus-653 provides the port operations that are compatible to ARINC 653. Before build-time, ports, channels, and message type should be configured. On this phase model driven approach can be applied [6], but many commercial tools such as VxWorks-653 [7] just support an XML configuration editor. We developed the Esto-AIR as an IDE for the Qplus-653 providing the configuration editor and code translator from an XML to C files.

Those configuration files consist of configuration codes for the kernel and application partitions. In the kernel part of configuration codes, the queue for the inter-partition communication must be defined with a fixed size. A message used in the transmission is defined as a continuous block of data of finite length. The expression "continuous block" means a sequential data arrangement in the source and destinations memory areas. Sending a message is achieved by copying the message from memory to memory in Qplus-653. Conceptually, the port for the destination of a message belongs to a partition, and not a process within a partition. But in the implementation level, the message transmission, memory copy is performed within the memory area arraigned in the kernel.

Figure 50.2 depicts the flow of function call from application partitions to the kernel and the area of main memory for the inter-partitions communication of Qplus-653. In the situation that the "Partition_1" sends a message to the "Partition_2", the source "portA" transmits a data to the destination "portB". In the kernel boot-time, both ports are created by the service CREATE_QUEUING_PORT and allocated of each port in the main memory. Application partitions only can obtain the identifier of the port which is created by the creating service with the name of the port.

Both sending and receiving functions should use the memory copy function "memcpy" for transmitting the given messages. At the application level, the transmission between "portA" and "portB" is carried out with "A" and "B" in Fig. 50.3, which may be local or global variables in their own partitions. But the real transmission is conducted by "portA" and "portB" of kernel area in the memory. The port structure has the pointer to point the message queue and the size of this queue is predefined. We implement the circular queue for this communication. If there are more than two channels, then the queue is split into the number of channels. And each split queue becomes a small circular queue.

So far we explained the method of implementation for the inter-partition communication of Qplus-653. Then we can drive some safety requirements from the followings:

- Any split queue should not be involved each other.
- Each port should assess the split queue corresponding to the channel of the port itself.
- The index of the split queue pointed by a source port should be equal or bigger than one of the destination port.
- Any kernel object except ports should not assess the split queues.

In this paper, we are only interested in the kernel level requirements, not ones of application level nor APEX one.

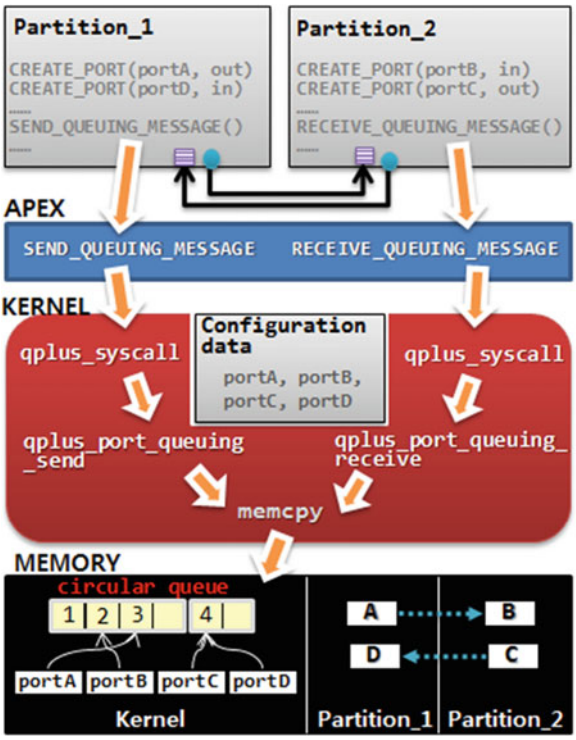


Fig. 50.2 Overview of inter-partition communications in Qplus-653

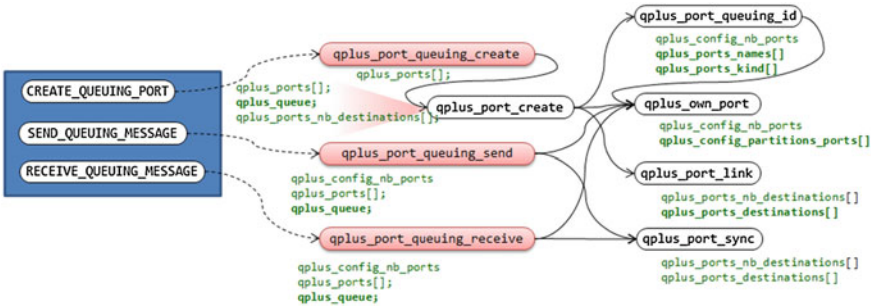


Fig. 50.3 Call graph for message transmission in queuing mode

50.3 Verification of Non-error-propagation in Qplus-653

In this section we explain the verification method and the results. We had to use SPIN model checker to verify the inter-partition communication. To avoid the manually construction of a verification model, SPIN versions 4.0 and later support

the inclusion of embedded C code within the verification model [5] and variables in the embedded C code can be abstracted. Because of the enormous state space of real code, the data abstraction is the inevitable step of formal verification. The important data structures are for queue and ports: "qplus_queue" and "qplus_ports". In order to avoid the state explosion, attributes are abstracted away except ones to access the "qplus_queue" and to be used in calculation of the index for accessing the queue, because they are enough to verify given properties.

Figure 50.3 depicts the call graph of "qplus_port_queuing_create", "qplus_port_ queuing_send" and "qplus_port_queuing_ receive", where the green texts are names of data structures configured in the initialization step. We decided the function flattening for facilitating the verification on a reasonable time. Only functions represented by the red oval were modeled and others have flattened in the major functions. Before explaining properties to be verified, it is need to explain details about the split queue and index attributes of the port to access the queue. Figure 50.4 shows the concept of mechanism of accessing the queue. The top of the figure is conceptual queue assigned for a port and attributes of the port, and the bottom one is implementation features of the queue of ports in Qplus-653. In the figure, both "portA.index" and "portX.index" are source ports and the values are determined at the configuration step. The "portA.index" attribute is the start point of the split queue for the portA.

Message send (receive) under the queuing mode would be summarized as follow:

1. Check parameters and wait for queue to be available
2. Calculate the size of the split queue for the given port and indexes of the circular queue
3. Memory copy with data and length
4. Increment the **port.end**(**port.begin**) index with the length of the message
5. Increment(Decrement) the **port.msg_count**
6. Synchronize the indexes of ports in the communication.

For sending a message, "end" and "msg_count" attributes may be increased. For receiving a message, the "begin" attribute may be increased and the "msg_count" attribute may be decreased. Of course "end" and "begin" attributes should be increased in the circular way. Now, the properties presented in the end of Sect. 50.2 are formalized with above attributes as below:

- portA.index + portA.end and portA.index + portA.begin are less than the portX.index.
- portA.end and portA.begin are bigger than equal to zero and less than equal to portA.max_size portA.max_count.

The last two properties are very hard to express with the above way and verify in the system. Fortunately the above properties are satisfied in the Qplus-653.

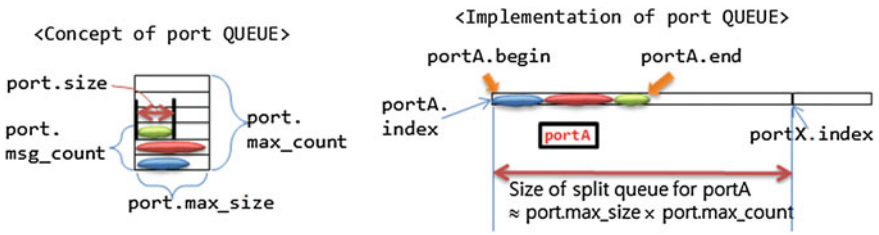


Fig. 50.4 Attributes of the port structure accessing the queue

50.4 Conclusions

In IMA architecture based on the partitioning concept, it is important to isolate the errors within a single partition, not being propagated outside the partition. However, there is the possibility of errors to be propagated into other partitions since inter-partitions communication is carried out by the memory copy operation at kernel level.

In this paper, we identified some safety requirements and conducted the formal verification for the communication part of Qplus-653 kernel developed to be used for an unmanned aerial vehicle.

References

1. Alena RL, Ossenfort JP, Laws KI, Goforth A, Figueroa F (2007) Communications for integrated modular avionics. In: Proceeding of IEEE aerospace conference 2007, pp 1–18
2. Airlines Electronic Engineering Committee (2006) Avionics application software standard interface—ARINC specification 653 Part 1—required services, Aeronautical Radio Inc.
3. Rushby J (2000) Partitioning in avionics architectures: requirements, mechanisms and assurance. DOT/FAA/AR-99/58
4. RTCA/EUROCAE, RTCA DO-178B/ED-12B Software Considerations in Airborne Systems and Equipment Certification, Dec 1992
5. Holzmann G, Joshi R (2004) Model-driven software verification. In: Proceeding of SPIN 2004, vol 2989, LNCS, pp 76–91
6. Lasnier G, Wrage L, Pautet L, Hugues J (2011) An implementation of the behavior annex in the AADL-toolset Osate2. In: UML&AADL—6th IEEE international workshop UML and AADL, Las Vegas, USA, April 2011
7. Wind River Systems (2010) VxWorks®653 configuration and build guide 2.3. Wind River Systems, Inc.

Chapter 51
3D Face Representation Using Inverse Compositional Image Alignment for Multimodal Face Recognition

Kanghun Jeong, Hyeonjoon Moon and Sanghoon Kim

Abstract A 3D model based approach for a face representation and recognition algorithm has been investigated as a robust solution for pose and illumination variation compared to 2D face recognition system. However, a 3D model based face recognition system is generally inefficient in computation time and complexity. In this paper, we propose a 3D face representation algorithm to optimize to have the same vertex number. Then, create an average model using processed 3D data. Finally, we evaluate fitting and face recognition performance based on 3D average model.

Keywords 3D face representation · Multimodal face recognition · 3D face database · Pixel-to-vertex map · 3D average face model

51.1 Introduction

3D face data provides absolute geometrical shape and size information, less sensitive to illumination [1]. Blanz and Vetter [2] found appropriate shape and texture coefficient vectors of the 3D morphable model by fitting process. However, this approach has complexity and inefficiency problems caused by very large vertices (about 80,000–110,000) despite excellent performance.

In this paper, we propose a novel 3D face representation algorithm based on a texture pixel to 3D shape information. It is possible to reduce the vertices which

K. Jeong · H. Moon
Department of Computer Science and Engineering, Sejong University, Seoul, South Korea

S. Kim (✉)
Department of Electrical, Electronic and Control Engineering, Hankyong National University, Ansung, South Korea
e-mail: kimsh@hknu.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_51, © Springer Science+Business Media Dordrecht 2014

can be aligned with correspondence information of a reference 3D shape data simultaneously. Based on proposed algorithm, a face can be represented less than 5,000 vertices. Our fitting processes provide increased recognition accuracy with less calculation time.

51.2 Face Representation

The 3D information is represented by close connections of many vertices and texture intensity in a 2D frame. Since the vertex number of the 3D face scans is different from the others, it is necessary to manipulate them to have the same number of vertices and remove texture illumination for consistent mathematical expression and drawing in scene [3].

We propose a novel 3D face representation algorithm based on a pixel to vertex map (PVM). It minimizes computational complexity while maintaining reasonable recognition performance.

51.2.1 Texture Correspondence

Texture data has different pose and lighting conditions from data acquisition process. In [4], they assume single point light source with variable illumination and fixed pose.

In our experiment, optical and SIFT flow [5] have been examined [6]. We choose optical flow to achieve texture correspondence which shows more uniform distribution of data since it has fixed lighting condition. Experimental results of both algorithms and face images are shown in Fig. 51.1 and Table 51.1. Our experimental result shows that the optical flow performs slightly better performance than SIFT flow in computation time and classification of face region.

51.2.2 Vertices Correspondence and Model Generation

A Pixel-to-Vertex Map (PVM) is an algorithm of binary image alignment, which classifies pixels in the masked face region into ones mapped to a vertex and the opposite (corresponding vertices). We call the former as active pixel (AP) and the latter as inactive pixel (IP) [7].

AP and IP are expressed as dark and bright pixels in an elliptical mask (Fig. 51.1c).

The procedures for the vertex correspondence using a PVM are as follows:

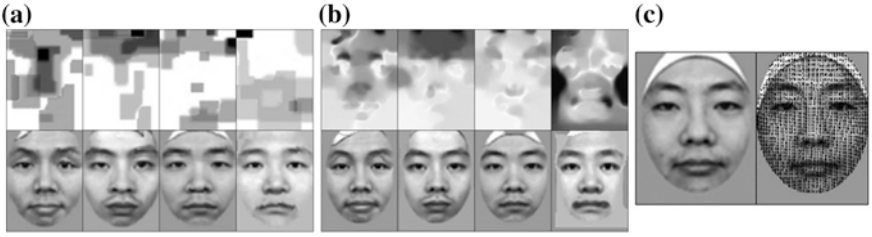


Fig. 51.1 Flow algorithm comparison and vertex correspondence result by PVM. **a** SIFT flow results. **b** Optical flow results. **c** PVM results

Table 51.1 Optical flow and SIFT flow result

Face DB file ID number	SIFT flow		Optical flow	
	Time (s)	Max flow	Time (s)	Max flow
1	0.526346	3.0000	0.410662	3.6233
2	0.457788	2.2361	0.417441	3.0577
3	0.470041	2.0000	0.417858	3.0577
4	0.469789	2.8284	0.415745	3.1779

1. Construct each PVM matrix of M + 1 3D face scans and build the vertex position matrix by stacking the position vector of the vertex mapping to each AP in a PVM. If the resolution of the texture frame is C by R, the PVM matrix of the *i*th scan, denoted by M_i and the vertex position matrix of the *i*th scan, denoted by P_i are obtained as

$$\mathrm{M}_i = \begin{bmatrix} m_{11}^i & m_{12}^i & \dots & m_{1c}^i \\ m_{11}^i & & & \\ \vdots & & \ddots & \vdots \\ m_{r1}^i & & \dots & m_{rc}^i \end{bmatrix}, \quad m_{rc} = \begin{cases} 0, & if\ p_{rc}\ is\ \mathrm{IP} \\ 1, & if\ p_{rc}\ is\ \mathrm{AP} \end{cases} \tag{51.1}$$

$$\mathrm{P}_i = \begin{bmatrix} \mathrm{V}_1^i & \mathrm{V}_2^i & \dots & \mathrm{V}_{s(\mathrm{M}_j)}^i \end{bmatrix} \tag{51.2}$$

where, p_{rc} is the pixel positioned at (r, c) in the texture frame and $\mathrm{s}(\mathrm{M}_j)$ is PVM size, AP numbers in the PVM. Also, $\mathrm{V}_j^i = [x_j\ y_j\ z_j]^T$ is the 3D position vector of the vertex mapping to the *j*th AP in the *i*th scan.

2. Select a Reference PVM, denoted by M^R, by maximizing this criterion.

$$\mathrm{M}^R = \arg\max_{\mathrm{M}_j} s(\mathrm{M}_j) \tag{51.3}$$

The size of the RPVM $\mathrm{s}(\mathrm{M}^R)$ means the vertex number of a reduced subset. Then, all scans will be in correspondence with the vertex number. Likewise, the vertex position matrixes of RPVM are denoted by P^R.

3. Compute each modified vertex position matrix of all scans except one selected for the RPVM.

$$\widehat{\mathrm{V}_k^i} = \begin{cases} \mathrm{V}_{p(k)}^i & if\ m_{p(k)}^i\ is\ \mathrm{AP} \\ \mathrm{V}^N & if\ m_{p(k)}^i\ is\ \mathrm{IP} \end{cases} \tag{51.4}$$

$$\mathrm{V}^N = \sum_{q=1}^{8} \mathrm{W}_q \mathrm{V}_q \tag{51.5}$$

where, $\widehat{\mathrm{V}_k^i}$ is a modified vertex position vector, which is the same to the position of the original vertex if mapped to AP, otherwise should be acquired by an interpolation method. And, the subscripted $\mathrm{P}(k)$ means the position of the pixel mapped to the vertex related to kth column in the P^R. We have to seek an appropriate 3D position V^N for a vertex mapped to IP using linear combinations of the positions of vertices mapped to 8 nearest neighbor APs in the PVM of the target scan as defined in Eq. (51.5).

All 3D vertices information is corresponded by PVM. Then we constructed separate models from shapes and textures of 100 Korean people by applying PCA and LDA [8] independently. The 3D morphable model was constructed with the 3D face samples acquired from a stereo based camera. Texture information of 3D face samples is well defined in the reference frame where one pixel corresponds to one 3D vertex perfectly. Also, we utilize an interactive tool to remove holes and spikes inevitably generated in 3D scanning of a human face. Models are generated by linear combination of the shapes and textures. Through the PVM, all 3D face samples have same number of vertices and polygons [7]. It means that all 3D face samples are existed in a consistent vector space, which is decided by the corresponded vertex number. Suppose that V is the vector space dimension for 3D face representation. Shape information is expressed in a matrix form as

$$A_i = \begin{pmatrix} x_1^{(i)} & x_2^{(i)} & \dots & x_V^{(i)} \\ y_1^{(i)} & y_2^{(i)} & \dots & y_V^{(i)} \\ z_1^{(i)} & z_2^{(i)} & \dots & z_V^{(i)} \end{pmatrix}, \quad (i = 1, \dots, N) \tag{51.6}$$

Then, we apply PCA to make an eigenspace for 3D shapes. PCA to create an eigenspace from sample shapes is achieved by the following sequential steps:

(a) Represent each shape of 3D face samples in column vector form.

$$Q_i = vec(A_i) \tag{51.7}$$

where, $vec(A)$ means that A is vectorized by stacking its column.

(b) Compute the mean shape from all samples. The shapes are centered, they are combined into a data matrix of size $3V \times N$, denoted by B.

$$\bar{S} = \sum_{j=1}^{N} Q_j, B = [b_1 b_2 \cdots b_N] = [(Q_1 - \bar{S})(Q_2 - \bar{S}) \cdots (Q_N - \bar{S})] \tag{51.8}$$

where, V and N are the number of vertices and 3D face samples.

(c) The data matrix is to create a covariance matrix as

$$\Omega = \frac{1}{N} BB^T = \frac{1}{N} \sum_{j=1}^{N} b_j b_j^T, \quad \Omega V_s = \Lambda V_s \tag{51.9}$$

Since Ω is a symmetric matrix, its eigenvectors are orthogonal. Where, V is the set of eigenvectors associated with the eigenvalues Λ.

(d) Arrange the eigenvectors $S_i \in V_s$ according to their corresponding eigenvalues λ_i, which are elements of the diagonal matrix Λ. Since the N vectors b_j are linearly dependent. We call the eigenspace.

$$V_s = [S_1 S_2 \cdots S_{N-1}] \tag{51.10}$$

(e) Reduce the dimension of the eigenspace. The higher order eigenvectors are operating as noise to reliable model generation. We eliminate the eigenvectors by selecting an appropriate eigenvalues N_s, less than $N - 1$.

$$V_s' = [S_1 S_2 \cdots S_{N_s}](N_s \leq N - 1) \tag{51.11}$$

(f) Implement a form of linear object class for shape based on the previous eigenspace generation steps. A formal shape and texture model is

$$S = \bar{S} + \sum_{j=1}^{N_s} \alpha_j S_j = \sum_{j=0}^{N_s} \alpha_j S_j, \; T = \bar{T} + \sum_{j=1}^{N_v} \beta_j T_j = \sum_{j=0}^{N_v} \beta_j S_j \tag{51.12}$$

(g) Where, $\alpha_0 = 1, S_0 = \bar{S}$ and S is a formal shape model. And, N_v is the number of basis spanning the eigenspace to textures. β_i will be used as a parameter (coefficient) for similarity measurement in face recognition system, including weight information of the projection onto the ith basis.

51.2.3 Fitting the Face Model Using Inverse Compositional Image Alignment

Shape and texture coefficients of the 3D model are estimated by fitting for the input face. Our choice of fitting algorithm is Inverse Compositional Image Alignment (ICIA) [2]. The ICIA algorithm is a fitting algorithm which guarantees

Table 51.2 Comparison of experimental configuration and performance to ours with the Max Planck Institute (MPI) based on ICIA fitting algorithm

Categories	MPI's	Ours
Used samples (for model generation)	100	100
Dimension of shape and texture model	50 %	50 %
Model vertices	18,404	4,014
Gallery/Probe number	68/544 (68 × 8)	21/51
Average recognition rate (frontal/sided)	99.5/94.5 %	89.9/83.2 %
Average fitting time (ms)	300	56

computational efficiency. Fitting steps are iterated until they converge to a given threshold value, minimizing texture difference between projected model and input image. During fitting process, texture coefficients are updated automatically.

51.3 Experimental Result

The performance evaluation of the proposed system was conducted with identification scenario [9, 10]. Face database used in our experiment are in 120 × 160 resolution. Average fitting time per an image of texture only and texture + shape (both) case is 0.9 and 5.6 s on Pentium-M 1.73 GHz CPU configuration. We made comparison of our experimental performance between MPI's as listed in Table 51.2 [2]. Based on the recognition rate, ours are on average 9.6 % in frontal case and 11.4 % in sided case less than each the MPI's. However, in observation for average fitting time, ours results shows far faster than the MPI's. There are two factors to allow such result. First factor is the number of vertex points constituting the generative 3D model. The complexities of fitting process depend on the number of vertex point and the number of selected shape and texture basis of the generative model. Vertex points of the MPI's are about 4.5 times as many as ours as presented in the second category for Table 51.2. The other factor is whether a manual registration exists or not. MPI's approach requires manual registration for initialization of fitting parameter of an input.

51.4 Conclusion

In this paper, we proposed an efficient 3D vertices correspondence algorithm for data minimization and face representation. This algorithm is called pixel-to-vertex map (PVM) which can be applied on preprocessed shape information. On the basis of the PVM, 3D face data including 30,000–40,000 vertices can be efficiently represented with 4,822 vertices. We have collected the 3D face database using

Geometrix FaceVision 200 which is a stereo-camera based imaging device. We have generated a 3D morphable model by each 3D face image synthesized with 4,822 vertices based on PVM. Shape and texture coefficients were estimated by fitting with an input face using the Inverse Compositional Image Alignment (ICIA) algorithm. Based on the experimental results, the proposed algorithm presents a reasonable recognition rate while maintaining computation complexity.

Acknowledgments This work was supported by a research grant from Gyunggi-do (GRRC) in 2013–2014 [(GRRC Hankyong 2012-B02)].

References

1. Papatheodorou T, Rueckert D (2005) Evaluation of 3D face recognition using registration and PCA. Lect Notes Comput Sci 3546:997–1009
2. Blanz V, Vetter T (2003) Face recognition based on fitting a 3D morphable model. IEEE Trans Pattern Anal Mach Intell 25:1063–1074
3. Xie X, Lam KM (2005) Face recognition under varying illumination based on a 2D face shape model. Pattern Recogn 38:221–230
4. Georghiades A, Belhumeur P, Kriegman D (2001) From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Trans Pattern Anal Mach Intell 23:643–660
5. Liu C, Yuen J, Torralba A (2011) SIFT flow: dense correspondence across scenes and its applications. IEEE Trans Pattern Anal Mach Intell 33:978–994
6. Jain R, Kasturi R, Schunck B (1995) Machine vision. McGraw-Hill, New York
7. Jeong K, Moon H (2011) Efficient 3D model based face representation and recognition algorithm using pixel-to-vertex map (PVM). Trans Internet Inf Syst 5:228–246
8. Chen X, Zhang J (2012) Optimization discriminant locality preserving projection of gabor feature for biometric recognition. IJSIA 6(2):321–328
9. Phillips PJ, Moon H, Rizvi S, Rauss P (2000) The FERET evaluation methodology for face-recognition algorithms. IEEE Trans Pattern Anal Mach Intell 22:1090–1104
10. Phillips PJ, Grother P, Micheals RJ, Blackburn DM, Tabassi E, Bone JM (2003) FRVT 2002: evaluation report. NIST

Chapter 52
Simulating Synthesized Automata from Decentralized Specifications in LTL

Ryoungkwo Kwon and Gihwon Kwon

Abstract In this paper, we present a technique for simulating the synthesized automata from decentralized specifications in linear temporal logic (LTL). Each automaton is synthesized from individual specifications written by generalized reactivity called GR(1) formula which is a restricted fragment of LTL. GR(1) specifications have a possibility of including mistakes so that it has to be validated. This paper describes the validation technique for simulating synthesized automata. As the result, we found out an inconsistent behavior of specifications.

52.1 Introduction

In this paper, we present a technique for simulating the synthesized automata from decentralized specifications in linear temporal logic (LTL). The specification is written by GR(1) formula in special class of LTL. For synthesizing automaton, we used the synthesis algorithm proposed in [1] which receives GR(1) specifications and checks whether there exists automaton realizing the specification. If such automaton exists, then it produces the automaton which can be considered as whole system.

This work was supported in part by the GRRC (Gyeonggi Regional Research Center) program of Gyeonggi Province. [2013-0548, Developing and Industrializing Core-Technologies for Social Services Convergence Platforms].

R. Kwon (✉) · G. Kwon
Department of Computer Science, Kyonggi University, Suwon, South Korea
e-mail: rkay8496@kyonggi.ac.kr

G. Kwon
e-mail: khkwon@kyonggi.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_52, © Springer Science+Business Media Dordrecht 2014

If we consider each automaton as one module belongs to the system, then we can't convince of behaviors of the system, because one module depends other one's behaviors. Even if each module is correct against a specification, interactions between these modules may cause inconsistent behaviors in the system.

To deal with this problem, we have developed a tool for simulating multiple automata to validate behaviors in the system. This tool automatically synthesizes multiple automata from decentralized specification written by GR(1) formula plus interactions manually added and then explores one infinite execution trace. From this trace, we can validate whether behaviors of the system is correct or not. From the experimental results, we have showed that a cause stems from an interaction between automata and modified specific specification into correct one.

This paper is organized as follows. In Sect. 52.2, we describe backgrounds about GR(1) synthesis. In Sect. 52.3, we describe the technique how decentralized specification is written. In Sect. 52.4, we show a tool we have developed for simulating and validating multiple automata. In Sect. 52.5, we present a case study and we will give a conclusion in Sect. 52.6.

52.2 Backgrounds

LTL [2] is based on the linear time which contains the propositional logic operators and some temporal ones. Therefore this logic expresses behaviors of the system on the sequences of the set of atomic propositions. The temporal means the order relation of behaviors, but it is used as an abstraction of concrete behavior from the system.

The synthesis is promising one among the formal methods in which a specification is written by logical formulas and then checked whether a system is realizing the specification. If such specification can be realizable, a model like automata is generated that it completely includes all desired behaviors in the specification. Otherwise, it shows that the specification is unrealizable.

In [1], Amir Pnueli proposed an effective algorithm of the synthesis. Logical formulas is written by a special class of the linear temporal logic, so called Generalized Reactivity(1), which is the fragment of full LTL. With restricting its syntax, GR(1) is used in synthesizing various applications. In general, the synthesis is regarded as automatic construction from specification to a system.

52.3 Decentralized Specifications

In this section we present a method how decentralized specifications are written. In prior to describe the method, we need to explain our idea we consider. Figure 52.1 depicts a big picture in this paper.

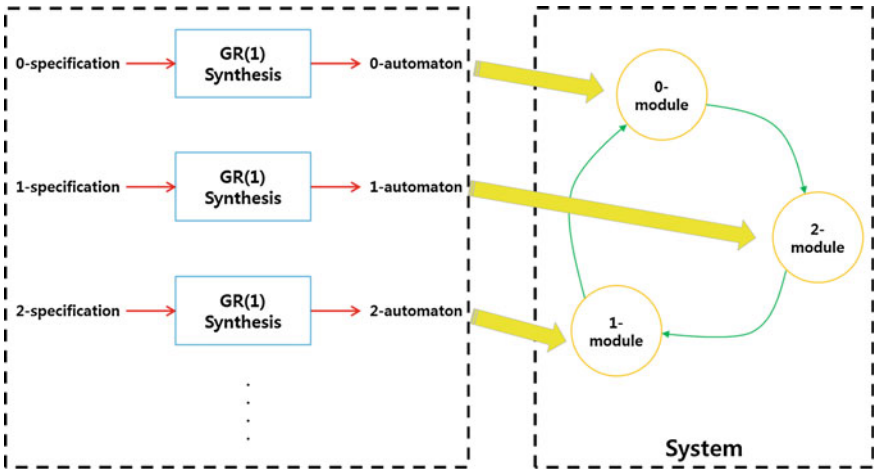


Fig. 52.1 Decentralized setting

We assume a decentralized setting in which a system is consist of multiple modules. In this setting, we can think that a module is considered as automaton and then from a specification, this automaton is generated by using GR(1) synthesis. So, multiple automata can behave single system. A module may interest with behaviors of other module. Some inputs of one module are based on some outputs of other module and this module behaves according to these inputs. These interactions between the modules are considered as relations of both inputs and outputs. These relations are only possible when specifications have already exist. A specification is written by GR(1) formula φ.

$$\varphi = \varphi_{\mathrm{i}} \Rightarrow \varphi_{\mathrm{o}}$$

Above formula has two parts, an input specification φ_{i} and a output specification φ_{o} respectively. A output specification φ_{o} represents desired behavior of a module itself, while an input specification φ_{i} represents behaviors from other one. Most important difference with GR(1) formula which is proposed in [2], is that an behaviors of other module are only as true. In our setting, above formula is acceptable, because this specification is only based on behaviors of other module's output. Thus, in this paper, we write a specification as below.

$$\varphi = \mathrm{TRUE} \Rightarrow \varphi_{\mathrm{o}}$$

Another important difference with GR(1) formula is restriction of the syntax. Traditionally GR(1) formula is also possible to combine variables of inputs with a temporal operator "next" as well as output variables. But we restricted this syntax to combine only output variables with it. Such restriction stems from some reasons. It is to reflect characterization of behaviors of execution in the system. Each module interacts with other module so that one module receives most current output from other model as an input.

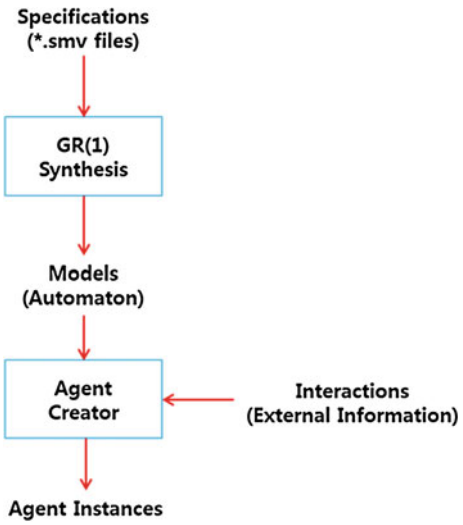


Fig. 52.2 Overall processes in the tool

52.4 Simulating Synthesized Automata

The overall process of a tool we have implemented is Fig. 52.2.

We have implemented a tool for simulating automata to validate behaviors in the system. This tool automatically synthesizes multiple automata from decentralized specification written by GR(1) formula plus interactions manually added and then explores one infinite execution trace. For developing this tool, we are using JADE (Java Agent Development Framework) [3] that is a platform that can implement agents programmatically and execute them. In JADE, synchronous or asynchronous communication between agents is provided as a high-level API.

Each specification is written by an input format of the NuSMV model checker. For an algorithm of the GR(1) synthesis, we have used the JTLV [4] library implemented the JAVA. Automaton can be generated from each specification, if a specification is realizable. To execute automata in JADE, each automaton is converted into an instance of a software agent. Also interactions between automata should be given additionally and this tool automatically added them for agents. Figure 52.3 shows a screenshot of our tool.

52.5 Case Study

We have validated a scenario through our technique. Figure 52.4 shows the setting of this scenario. The target scenario is that an arbiter and two robots exist. Each robot should traverse some regions and but r_4 should be passed by only one robot at a time. Whenever a robot sends a request to an arbiter for passing r_4 and this

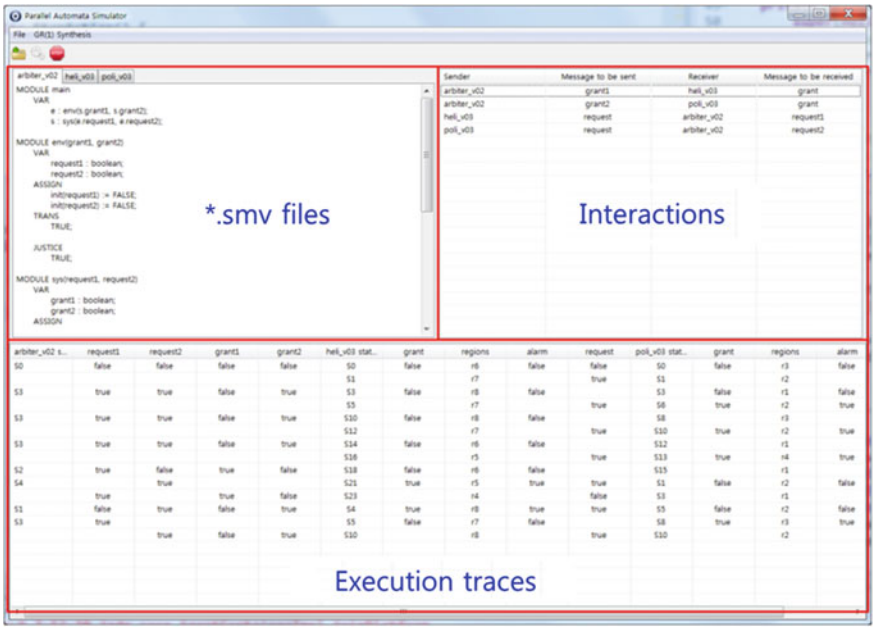


Fig. 52.3 Screenshot of the tool

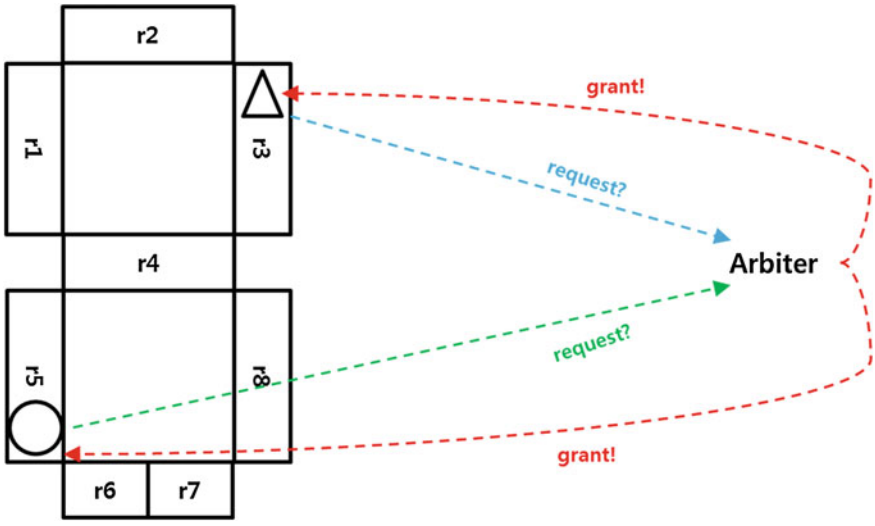


Fig. 52.4 Setting of the scenario

robot doesn't pass r_4 until granted. If a robot is granted, immediately go to r_4. An arbiter can grant to other robot when a robot which has arrived lives. Behaviors of an arbiter are considered as a critical section problem.

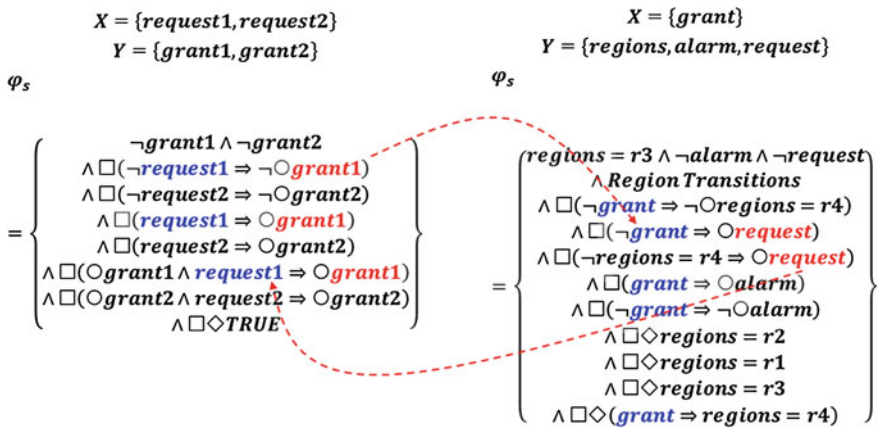


Fig. 52.5 Specifications

Figure 52.5 shows two specifications that are initially used in this scenario. Left one is an aribter's specification, and right one is a robot's specification. All variables are boolean. X and Y are input and output variables respectively. Variable request means that each robot requested for passing r_4. Variable grant means that an arbiter granted to a robot which has requested. So, each robot should be granted before traversing a shared region. In this case, interactions between each robot and an arbiter, can be identified. Therefore, the arbiter's input is a request from the robot and the robot's input is grant from the arbiter. So we can identify such relations of input and output on the specification.

As a result of simulation, we have obtained an execution trace of the arbiter below.

$$\{\neg grant1, \neg grant2\}, \ldots, \{grant1, grant2\}, \ldots$$

From this execution trace, we have found an inconsistent behavior of the arbiter, because the arbiter's behavior includes that both robots are granted at same time. So, we have modified two logical formulas, fourth and fifth from the top, in the arbiter's specification for solving this problem. Modified version of these formulas is shown below.

$$\Box(\text{request1} \wedge \neg \bigcirc \text{ grant2} \Rightarrow \bigcirc \text{grant1})$$

$$\Box(\text{request2} \wedge \neg \bigcirc \text{ grant1} \Rightarrow \bigcirc \text{grant2})$$

52.6 Conclusions

In this paper, we present a technique for simulating synthesized automata from decentralized specifications in linear temporal logic. We also show how to specify decentralized specifications in LTL and how to validate the specifications by simulating the synthesized automata from the specifications using the simulator developed in this paper. As the result of applying our technique to the scenario of an arbiter and robots, we found out the inconsistent behavior of the arbiter.

References

1. Piterman N, Pnueli A, Sa'ar Y (2006) Synthesis of reactive(1) designs. In: Proceedings 7th international conference on verification, model checking, and abstract interpretation (VMCAI'06), LNCS vol 3855, pp 364–380
2. Clarke EM, Grumberg O, Peled DA (1999) Model checking. MIT Press, Cambridge
3. JADE (Java agent development framework), http://jade.tilab.com
4. Pnueli A, Sa'ar Y, Zuck, LD (2010) Jtlv: a framework for developing verification algorithm. CAV, pp 171–174

Chapter 53
Peers-Based Location of Mobile Devices

Mordechai Guri, Anatoly Krasner and Yuval Elovici

Abstract As the smart mobile device popularity is rapidly growing, numerous location-based services try to aid us in daily tasks, offering new patterns of consumerism and personal productivity. The wide range of location oriented services also increase the risk of services being abused by receiving fabricated location of the mobile device. Particularly, modern threats such as mobile viruses, botnets and malicious applications can spoof subscriber location for fun and profit. In this paper we suggest a technique which allows a service to verify a client's location based on confirmation from devices located nearby. Using this scheme makes it significantly harder to trick the service into accepting a spoofed location. Practically, this technique may introduce a new set of services which can rely on the reported location in high confidence. We show that with the continuous trends of prevalent Wi-Fi capable devices in modern environment, this scheme becomes practical, and discuss design considerations and implementation issues.

Keywords Security · Location · Peer-to-peer · Wireless

53.1 Introduction

Smartphone, tablets and mini-tablets are becoming more and more popular [1] even overtaking the desktop PC sales recently [2, 3]. These changes have motivated the creation of many location-based services: navigators, social games, coupons, location-based advertising, child and car tracking and many more. Generally,

M. Guri · A. Krasner (✉) · Y. Elovici
Department of Information Systems Engineering,
Ben-Gurion University of the Negev, Beersheba, Israel
e-mail: anatolyk@post.bgu.ac.il

Y. Elovici
Telekom Innovation Laboratories, Ben-Gurion University, Beersheba, Israel

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_53, © Springer Science+Business Media Dordrecht 2014

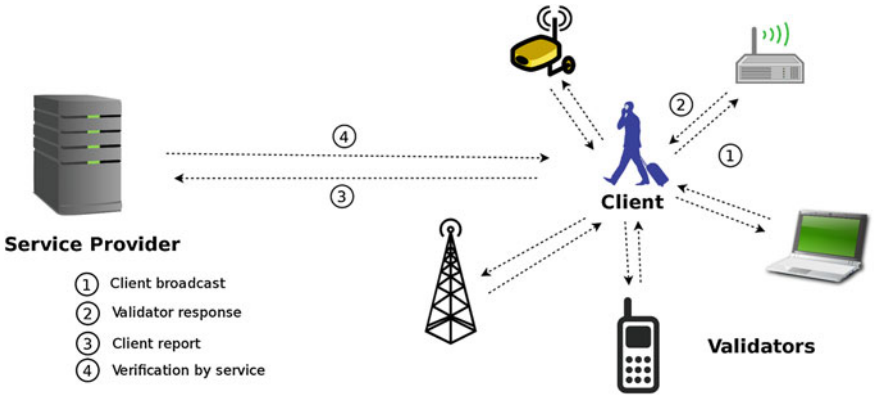


Fig. 53.1 Client and five surrounding devices

a service determines the subscriber location by requesting it from the device through an installed application or while visiting a web page.

However, the location data can be spoofed at several levels of the operating system or even at the GPS signaling channel [4], providing a service with wrong information. Such manipulations can be used to gain a personal profit as in tricking a stolen phone to report wrong location to phone tracking services. In another real-world example it has been demonstrated that Waze clients can spoof their location and report fake events [5]. In this paper we suggest a new method to authenticate the device location. The main idea is to use surrounding mobile phones and other Wi-Fi capable devices as a network of sensors which provides greater reliability to device reported location (see Fig. 53.1).

In order to spoof the device location, the adversary will have to subvert the surrounding devices, which significantly complicates its attack.

The presented concept is especially relevant in a modern environment which is continuously introduced with new Wi-Fi capable devices such as mobile phones, tablet, laptops, Wi-Fi routers, wireless cameras and hotspots. These trends are expected to continue in the future.

53.2 Motivation

The peer-verified location technique can be used to solve problems of existing location techniques, and also to introduce new applications.

(a) *Location authentication for area-based coupons:* A service may offer coupons and bonuses to the user based on the device location, for example offer coupons for businesses located near the user, or "rewarding" the user for staying in the mall. If a device can send spoofed location data, it becomes easy

to trick the service and cause it to send many coupons regardless of the real location of the device. A peer-based approach will make such services more reliable.

(b) *Avoid feeding fabricated data to collective services:* People count on navigation services when planning which roads to take. Services can also decide which routes are faster based on the amount of traffic. Some social games allow the user to detect and interact with users located nearby. All these services rely on user location reports; a peer-based technique will make these services more reliable, less vulnerable to attacks.

(c) *Reliable detection of phone/car/child location:* Such services operate reliably only as long as the map of users is correct. In crowded places, a peer-based approach may even be more accurate and faster than GPS, which is helpful when looking for a lost child.

(d) *Devices without location capabilities:* A device without GPS capabilities will be able to locate itself using the assistance of nearby peers whose locations are known.

(e) *Places with no location signals (GPS, cellular):* In closed places such as buildings, or where GPS signals are blocked, device can utilize the peer-based technique to receive location information from nearby devices which are exposed to signals. For example, devices near windows which do have GPS signal reception.

53.3 Model and Requirements

The model is based on the following assumptions.

1. Client devices have broadcasting capabilities. In this paper, we choose the Wi-Fi interface to broadcast messages to the surrounding environment. While Wi-Fi is chosen due to its highest availability, other interfaces such as Bluetooth and cellular are also possible in other contexts. Each validator device keeps its own estimated location updates. With mobile devices it is usually done through GPS, while in fixed routers and access points it can be done automatically by external services (Google Geolocation) or manually, by the administrator.
2. Validator devices have to share an encryption key and unique ID with the service provider. Main service providers such *Google* or *Microsoft* can function as central trusted services. Sharing of the key and ID has to occur only once, e.g. through the Internet or manually. This requirement is applied only to the secured version of the scheme, as will be explained in the following section.

53.4 Scheme

We designed two versions of the scheme, basic and secure. The basic version is suitable for cases where the availability of location mechanism is important, and has less concern about security issues. In these cases, the main benefit is the ability to locate devices quickly and reliably, even though the device itself doesn't have location hardware. The secure version is more complex and provides better security, stability and immunity at the cost of complexity.

First, in order to demonstrate the main idea of peer assisted location we present an outline of the basic algorithm, and then we go to the detailed secure algorithm.

53.4.1 Basic Scheme

The basic scheme consists of the following steps. (1) The service provider sends a location request to client wishing to be served. (2) The client then broadcasts location request paired with temporary code T_i using a wireless interface. T_i will tag the current request to distinguish it from other requests from same or other devices. (3) Location request is received by surrounding devices. (4) Each device estimates its own location, and sends the pair $<T_i, location>$ to the requesting device. (5) Client calculates its location based on the received pairs and sends it back to the service.

As mentioned, in this version we assume that client is trusted, and will not modify the received triples. The client provider may receive more than one response from nearby devices; it can then take the most accurate pairs given the sender type or RSSI (Received Signal Strength Indication), or calculate the mean value, using intersection triangulation methods.

53.4.2 Secure Scheme

The secure scheme assumes that the requesting device is untrusted. Several security enhancements such as messages encryption and signing have been added to scheme.

We denote S to be a centralized service which requests client location, C to be a mobile device wishing to report its location to the service and V to be the list of devices located near the client. $V_i.ID$ is the unique ID of the device, $V_i.K$ is the encryption key. $V_i.ID$ and $V_i.K$ are known to client and service provider. The following pseudo code outlines the functionality of each of the three components participating in this scheme.

Registration Phase: In registration phase, the service provider and validator device agree on parameters such as device ID ($V_i.ID$) and encryption key ($V_i.K$). In

a straight-forward implementation, the service provider generates the identifier and the encryption key. Typically, the device registers with the service provider once.

```
Algorithm 4.1: RegisterValidator(S, V_i)

  V_i.ID ← AllocateNewDeviceID(S)
  V_i.K ← GenerateEncryptionKey(S)
  AddToDatabase(S, V_i.ID, V_i.K)
```

Client Side: The following code demonstrates client functionality.

```
Algorithm 4.2: RequestService(C, S}

  T_i ← GenerateTemporaryCode(C)
  BroadcastRequest(C, T_i)
  Wait()
  P ← CreateReportPackage()
  V ← GetValidators()
  For each V_i in V
    AddReportToPackage(P, V_i.ID, GetReport(V_i))
  SendPackage(C, S, P)
```

The client broadcasts a temporary request code. It then collects validator replies for a short period of time. After that the client packs the validator-signed replies into a package (e.g. an encrypted file) and sends it to the service provider. The replies themselves are encrypted using the validators' private keys $V_i.K$; the client only passes them to the service provider, without decoding them.

Validator Side: When a validator device receives a broadcasted location request it act as follows.

```
Algorith 4.3: Validate(V_i, C, S, T_i)

  l ← EstimateSelfLocation(V_i)
  h ← GenerateTimeHash()
  r ← EncodeResponse(C, V_i.ID, l, h)
  Encrypt(r, V_i.K)
  SendReponse(C, r)
```

Upon client request, the validator get its own location. In addition to the location, the validator adds a time hash as an indication of the relevance of the validation. It encrypts the reponse, which serves as a signature for the service provider to verify the content.

53.5 Implementation and Supported Technology

A peer assisted location scheme needs software support at both client and devices. Such an authenticating scheme may be supported in the future by adding built-in service applications, which are running in the background persistently. Such

pre-installed applications already exist in both Android and iPhone, examples are *Google Maps* for Android devices and *iTunes* for iOS. We describe two hardware technologies which already exist and support the communication scheme.

53.5.1 Wi-Fi Direct and Wi-Fi Tethering

The communication between devices, i.e. between the client and the validators, can be established using Wi-Fi Direct technology [6]. With this technology, devices at Wi-Fi range are able to communicate directly and receive data at Wi-Fi bandwidth without negotiation of a central access point, and without an Internet connection. Using this technology, a mobile device may broadcast messages to surrounding Wi-Fi capable devices. Smartphones running the Android operating system offer an API for Wi-Fi Direct since version Android 4.0 [7]. This functionality is also referred as Wi-Fi Peer-to-Peer (P2P). In the same manner, mobile devices may broadcast through 'Wi-Fi Tethering' technology. In Wi-Fi Tethering, a device can share an internet connection with surrounding devices through Wi-Fi or Bluetooth messages. Devices can use the same broadcasting capabilities to send location requests to nearby devices. An implementation of the basic scheme is based on the following template: All phones have Wi-Fi enabled, as they periodically scan available location requests broadcast through the Wi-Fi interface. When a client needs to verify its location, it runs a Wi-Fi discoverable service for a short period of time. Phones around it discover the request, receive the broadcasted data and send a response using the described scheme.

53.5.2 Wi-Fi Routers and Access Points

Wi-Fi Routers, Access Points Routers and access points are inherently coming with an internal Wi-Fi card, which obviously ns. Such support can be applied by a firmware update to router and acable to monitor channels for messages in both Wi-Fi direct and Wi-Fi Tethering beacocess point, e.g. by adding a Linux background 'daemon' to monitor location requests.

53.6 Conclusion

In this paper we described a new approach to locate a mobile device, based on verification from surrounding devices. We show that the technique is useful when trusted and reliable location is needed, and can be utilized to locate devices lacking self-location capabilities. Finally, we show how such a scheme can be implemented using current technologies such as Wi-Fi direct and Wi-Fi Tethering.

References

1. Smartphone sales grew 46.5 percent in second quarter of 2013. http://www.gartner.com/newsroom/id/2573415
2. World's top PC vendor now sells more smartphones and tablets than PCs, BGR. http://bgr.com/2013/08/15/lenovo-smartphone-sales-q1-2014/
3. How to fool a GPS, TED. http://www.ted.com/talks/todd_humphreys_how_to_fool_a_gps.html
4. Tippenhauer NO, Pöpper C, Rasmussen KB, Capkun S (2011) On the requirements for successful GPS spoofing attacks. http://www.syssec.ethz.ch/research/ccs139-tippenhauer.pdf
5. Jeske T (2013) Floating car data from smartphones: what google and waze know about you and how hackers can control traffic, Blackhat. https://media.blackhat.com/eu-13/briefings/Jeske/bh-eu-13-floating-car-data-jeske-wp.pdf
6. Wi-Fi direct on Wikipedia. http://en.wikipedia.org/wiki/WiFi_Direct
7. Android WiFi P2P API. http://developer.android.com/guide/topics/connectivity/wifip2p.html

Chapter 54
Platform Design for Data Sharing

Yun-Hee Son, Min-A Kang, Woo-Seung Jo, Kyu-Jin Choi and Kyu-Chul Lee

Abstract Data sharing became important all over the world so that the need for data sharing systems has been increasing as well. Currently many people all around world are using data sharing systems developed in the United States and United Kingdom. In accordance with this recent needs, more data sharing model for valuable service research is needed. As a result this paper suggests a module for data sharing and platform design.

Keywords Data sharing · Platform design · CKAN · LOD2

54.1 Introduction

Recently the importance of data sharing has been acknowledged. Because of this efficient platform that can share data has been required. The purpose of this paper is making of open platform architecture for sharing data and creating standards for data sharing.

Y.-H. Son · M.-A. Kang · W.-S. Jo · K.-J. Choi · K.-C. Lee (✉)
Department of Computer Engineering, Chungnam National University,
220 Gung-Dong, Yuseong-gu, Daejeon, Korea
e-mail: kclee@cnu.ac.kr

Y.-H. Son
e-mail: mellow211@cnu.ac.kr

M.-A. Kang
e-mail: kma1101@cnu.ac.kr

W.-S. Jo
e-mail: champion29@cnu.ac.kr

K.-J. Choi
e-mail: rbwls31@cnu.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_54, © Springer Science+Business Media Dordrecht 2014

This paper deals with the importance in the design for data sharing platform architecture. It will be an essential element for the data sharing model study. To do this, existing studies and necessary functions are analyzed. So this study will be useful in various industries.

This paper is organized as follows: the next section introduces the related works. In Sect. 54.3, we propose a description of Platform Design. At last in Sect. 54.4, the conclusions and future works related to this paper.

54.2 Related Works

54.2.1 CKAN (Comprehensive Knowledge Archive Network)

CKAN [1] is an open data platform developed at the OKF (Open Knowledge Foundation), the goal is to making all data available to anyone. The functional structure of CKAN is shown in Fig. 54.1. Data publisher add data through web Interface or API and data user can find the shared data. In addition to the data catalog, analytics, visualizations function are provided.

Currently The United States [2] and the United Kingdom [3] have been using CKAN as a platform and more than 20 countries make up the instance.

However, CKAN provides limited data and information. Because of that connection between data is needed in order to make the data more valuable.

54.2.2 LOD (Linked Open Data)

LOD describes a method of publishing structured data so that it can be interlinked and become more useful. URI, XML, RDF, RDFS, and SPARQL of the semantic web layer cake is basic requirement of LOD as shown in Fig. 54.2. In addition to the basic information via LOD, all related data can be found. LOD is necessary technologies to increase the value of the data.

54.2.3 LOD2

LOD2 [4] is a large-scale project, 11 European countries are participating as a partner. LOD2's goal is to define a life cycle for generating LOD, also adopted a step by step tool to its life cycle. LOD2 life cycle and tools is shown in Fig. 54.3.

Tools were integrated using Debian Packaging System converting the raw data to LOD. And also these are connected to each other. So it is possible to make the information more valuable. Further, it can give synergistic effect to be fused connection to share.

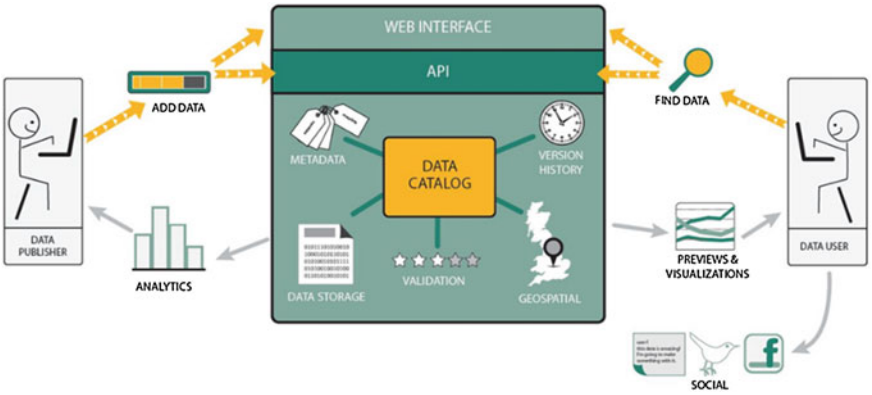


Fig. 54.1 Functional structure of CKAN

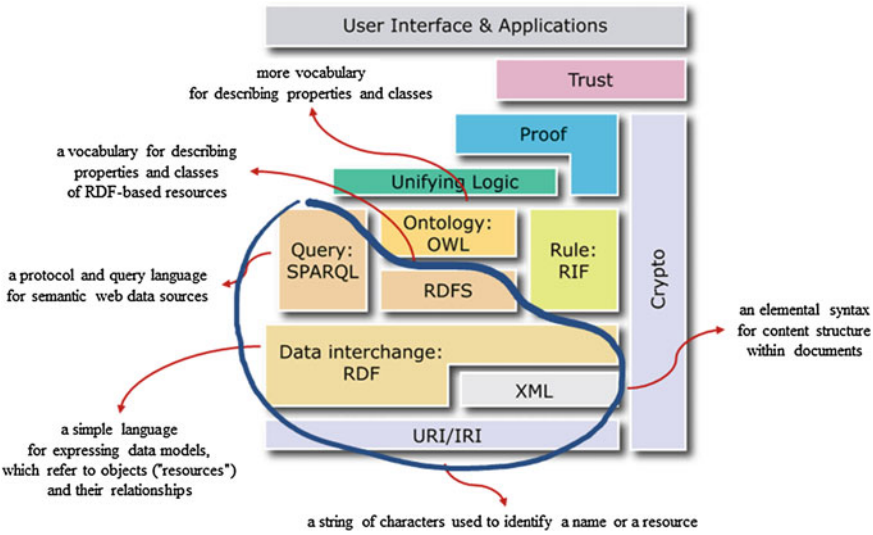


Fig. 54.2 The semantic web layer cake

54.3 Platform Design

This section describes platform design for data sharing and connections. The platform is designed to enable data sharing and LOD connect between data. Through users who can share their data, as well as through additional connections to make their data is valuable.

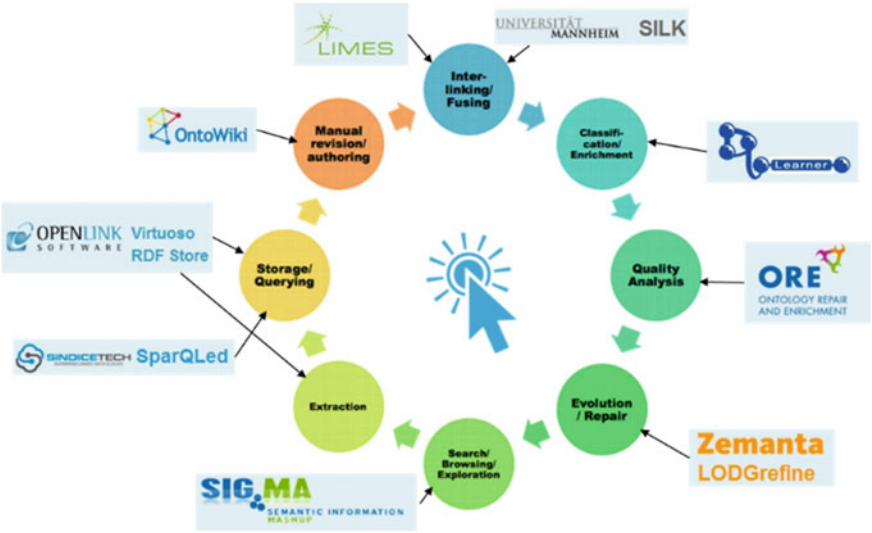


Fig. 54.3 LOD2 life cycle and tools

54.3.1 Module Definition

Through Analysis of existing research have defined Modules for data sharing and LOD. The six modules are shown in Fig. 54.4. Each module described in details in the next step.

54.3.2 Platform Architecture

The platform is largely composed of Data Sharing and LOD. Platform Architecture is shown in Fig. 54.5.

Data Sharing module can take advantage of various features to add and retrieve a user's data like CKAN. In addition it can be used linking with LOD function through Extraction module.

LOD is composed of five modules which are Extraction, RDF Storage/Querying, Ontology Authoring/Enrichment, Interlinking, and Inconsistency Detecting/Repair.

Extraction is automatically converted in the form of a traditional RDB data to LOD. At this time, Ontology is required in order to create RDF.

Ontology Authoring support you can easily create ontology. And it helps making of Ontology through Ontology Enrichment which is find top-level ontology and then the converted data in RDF Store. The Store data can be queried through SPARQL Endpoint.

Finally, using Interlinking Internal LOD can be connected with external LOD.

Fig. 54.4 Data sharing and LOD module

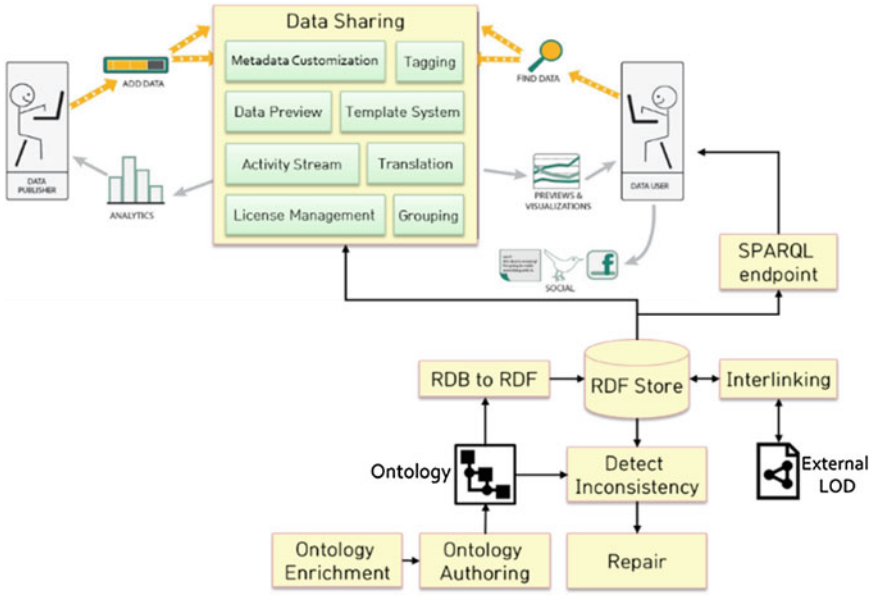


Fig. 54.5 Platform architecture

54.4 Conclusions

So far, the data has been isolated and exist independently. However, many governments and organizations were actively trying to share data and so they can make some deliverables. Good examples are data.gov, data.gov.uk. There are reason why sharing data are solving the old problems and getting more data by user. In this Paper, according to this need for data sharing platform was designed. This can be helpful to increase the value of data. Also, anyone can easily access to data and use that. We need to develop appropriate sharing service.

Acknowledgments This Research has been performed as a collaborative research project of project Building the System for Sharing and Convergence of Scientific Big Data and supported by the Korea Institute of Science and Technology information (KISTI).

References

1. CKAN. http://ckan.org
2. The United State. http://www.Data.gov
3. The United Kingdom. http://www.Data.gov.uk
4. LOD2. http://lod2.eu
5. Saltlux (2010) Linked open data. White paper
6. Auer S, Buhmann L, Dirschl C, Erling O, Hausenblas M, Isele R, Leh mann J, Martin M, Mendes PN, Nuelen BV, Stadler C, Tramp S, Wil liams H (2012) Managing the life-cycle of linked data with the LOD2 stack. In: international semantic web conference volume 7650 of lecture notes in computer science, LOD2 collaborative project. 2010. Deliverable 1.3 architecture and system De Sign, Springer

Chapter 55
Reasonable Resolution of Fingerprint Wi-Fi Radio Map for Dense Map Interpolation

Wonsun Bong and Yong Cheol Kim

Abstract In Wi-Fi based location fingerprinting, the cost of building a radio map is high. Though interpolating a coarse radio map into a dense one may reduce the cost, low accuracy at space discontinuity such as walls deteriorates the reliability of interpolated radio map. We present an interpolation of radio map, which preserves the discontinuity of received signal strength in accordance with the layout of a building. Furthermore, a closed form solution of reasonable resolution of radio map is derived. Experimental results show that, with sampling density >35 %, the localization accuracy is comparable to that of a full density map.

Keywords Wi-Fi · Indoor localization · Interpolation · Radio Map

55.1 Introduction

For location based service for a mobile device (MD), the localization of a MD is an essential requirement. When the MD is inside a building, GPS signal is not strong enough. As an alternative, Wi-Fi based indoor localization techniques have been actively studied [1], as most smart phones have Wi-Fi functions.

Received signal strength (RSS) gets weaker as a MD moves away from an access point (AP). RSS is used to estimate the distance from the signal source. In indoor propagation, RSS is modeled to have a linear decrease in [dBm] with reference to the log distance [2] as shown in (55.1). $P(r)$ and $P(r_0)$ are the signal power at distance r and at reference distance r_0, with a curve slope of n. W_A and w are the wall attenuation factor and the number of walls in the propagation path.

W. Bong · Y. C. Kim (✉)
Department of Electrical and Computer Engineering, University of Seoul, Seoul, South Korea
e-mail: yckim@uos.ac.kr

W. Bong
e-mail: gaam@uos.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_55, © Springer Science+Business Media Dordrecht 2014

$$P(r) = P(r_0) - 10 \cdot n \cdot \log\left(\frac{r}{r_0}\right) - w \cdot W_A \tag{55.1}$$

In reality, radio signal propagation in a building is perturbed due to the reflections and scattering by the obstacles in the propagation path, which results in a complicated attenuation pattern. Location fingerprinting [3] has been introduced to overcome this instability. Radio map is a database of RSS at many reference points (RP) from nearby access points (AP). The position of a MD is identified by matching its RSS pattern (fingerprint) $[v_1, v_2, \ldots, v_n]$ with the RSS vector $[s_1, s_2, \ldots, s_n]$ in the radio map and finding the position with the most similar pattern. Usually, Euclidean distance between patterns is used as the similarity measure.

While fingerprinting provides reliable location functionality, the cost of radio map is high due to the calibration of measured RSS data at a large number of RPs and updating the database as APs are added or removed. Interpolation of a coarse radio map into a dense one may reduce the cost of a radio map. An interpolated radio map, however, has low accuracy near a wall. Figure 55.1 shows the RSS curve superimposed on the interpolated values of RSS out of six RSS measurements, based on the path loss model. RSS exhibits a considerable drop on the wall boundary. The path loss model and the actual data exhibit a large discrepancy, especially at the wall boundaries.

We present a method of constructing a high-density radio map by discontinuity preserving smoothing (DPS). Image processing techniques such as adaptive smoothing [4], bilateral filter [5] and guided filter can serve DPS functionality. Among them, we employ adaptive smoothing since it is simple and fast in localized smoothing in accordance with the wall layout of indoor building.

The contribution of this paper is as follows. First, we propose to apply the adaptive smoothing to radio map interpolation. We examine what is the lower bound of sampling density for a reasonable accuracy in localization, when compared with the performance with a full density radio map. For performance evaluation, a fraction of full density radio map are randomly selected with a sampling density varying from 10 to 90 %. Second, a closed-form solution of the reasonable resolution of radio map is presented.

The organization of this paper is as follows: In Sect. 55.2, we present the basic idea of DPS. In Sect. 55.3, a closed-form solution for a reasonable resolution for a specified amount of RSS fluctuation is presented. In Sect. 55.4, we present experimental results and performance comparison between DPS, inverse distance weighting (IDW) and Voronoi tessellation. Finally, we draw a conclusion.

55.2 Proposed Interpolation

An intrinsic problem of interpolation is that it does not provide a deterministic framework of discriminating RPs in opposite sides of a wall. Such a functionality can be implemented by a localized smoothing and by limiting the region of smoothing such that no walls lie inside the region.

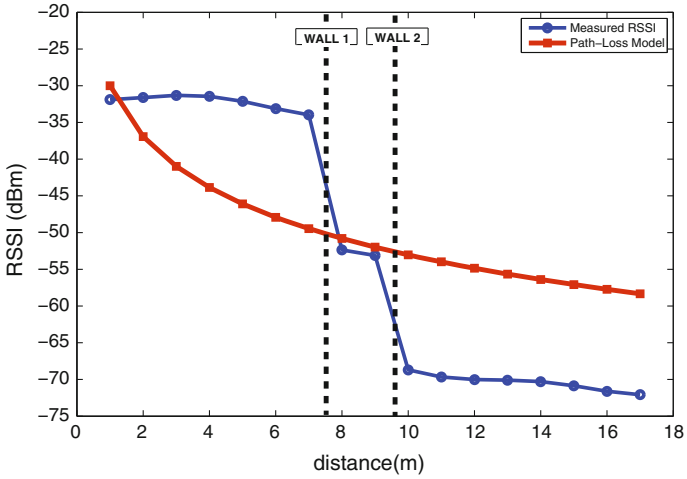


Fig. 55.1 Comparison of DPS and path loss model [2]

Adaptive smoothing can be used for that functionality. The smoothing strength of pixel intensity is controlled in accordance with the magnitude of signal gradients. As objects boundary is preserved by allowing large discontinuity, large RSS discontinuity can be preserved in adaptive smoothing which discerns the presence of walls. Figure 55.2 shows the DPS effect of adaptive smoothing. The random noise in the signal is smoothed out while the skeleton of the signal is maintained.

The process of adaptive smoothing is performed in iterative mode, as shown in (55.2)–(55.5). The superscript t represents the iteration number. $s^{(t)}(x,y)$ is the signal in the t-th phase and $w^{(t)}(x,y)$ is the smoothing kernel. The value of p in (55.4)–(55.5) is 10 if the cell is blocked from the center of the kernel by a wall, otherwise it is 1.

$\left|s^{(t)}(x+1,y)-s^{(t)}(x,y)\right|$ and $\left|s^{(t)}(x,y+1)-s^{(t)}(x,y)\right|$ are used as the magnitude of gradient in the x-direction and in the y-direction, respectively.

$$s^{(t+1)}(x,y)=\frac{\sum_{i=-1}^{1}\sum_{j=-1}^{1}s^{(t)}\cdot w^{(t)}(x+i,y+j)}{N^{(t)}} \tag{55.2}$$

$$N^{(t)}=\sum_{i=-1}^{1}\sum_{j=-1}^{1}w^{(t)}(x+i,y+j) \tag{55.3}$$

$$w_x^{(t)}(x,y)=\exp\left(-\frac{\left|s^{(t)}(x+1,y)-s^{(t)}(x,y)\right|^2}{2k^2}\cdot p\right) \tag{55.4}$$

$$w_y^{(t)}(x,y)=\exp\left(-\frac{\left|s^{(t)}(x,y+1)-s^{(t)}(x,y)\right|^2}{2k^2}\cdot p\right) \tag{55.5}$$

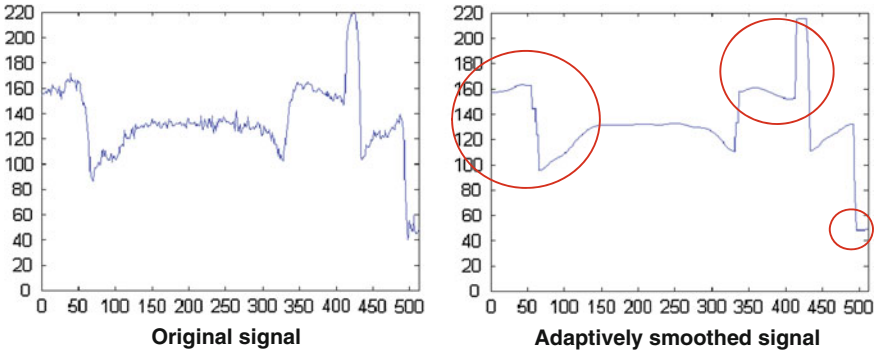


Fig. 55.2 Original noisy signal (before and after smoothing)

55.3 Reasonable Map Resolution with RSS Fluctuation

In addition to the positional randomness, RSS exhibits fluctuation over time on every measurement. RSS fluctuation affects the accuracy of localization. As derived from (55.1), a RSS difference of Δ_{RSS} corresponds to a radial difference of d_R.

$$d_R = r_0\left(10^{\frac{\Delta_{RSS}}{10n}} - 1\right) \tag{55.6}$$

In the construction of radio map, the measurement error of RSS can be decreased by taking the average of N measurements. Assuming each measurement is independent of one another, the magnitude of error decreases as $\frac{1}{\sqrt{N}}$.

On the MD side, however, the RSS is measured only once or twice in real-time localization. If the measurement error is 1.5 dBm, it results in a positional error of 1.9 m when $r_0 = 10$ m. In this case, a radio map with a resolution finer than 1.9 m is actually a waste of efforts. A reasonable value of resolution of a radio map, R_{map}, would be of the order of positional error resulting from RSS fluctuation as shown in (55.7), where δ_{RSS} is the standard deviation of RSS over time. Figure 55.3 shows that a reasonable value of resolution increases as the standard deviation of RSS gets larger.

$$R_{map} = r_0\left(10^{\frac{\delta_{RSS}}{10n}} - 1\right) \tag{55.7}$$

In experiments with four APs, the standard deviation of RSS (for 80 measurements) are 1.21 dB (AP 1), 1.68 dB (AP 2), 1.51 dB (AP 3) and 1.47 dB (AP 4), respectively. These correspond to R_{map} of about 2 m.

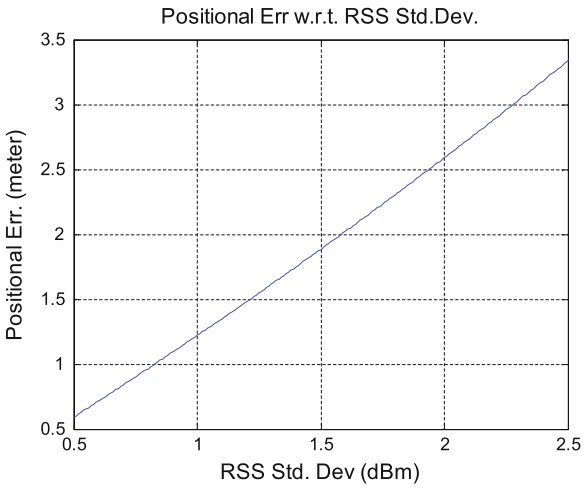


Fig. 55.3 Reasonable resolution of radio map w.r.t. std dev of RSS

55.4 Experimental Results

The performance of DPS-reconstructed radio map was tested and compared with two other methods: IDW interpolation [6] and Voronoi interpolation [7]. The location is identified by matching its RSS pattern with the reconstructed radio map and finding the one with the most similar RSS pattern in terms of Euclidean distance. The test was performed in three rooms and along an elongated hallway in the 6th floor of IT building [8]. Each RP is a grid point of 1.2 m by 1.2 m. The number of RPs is 145.

In the initial gathering, the RSS at each RP is measured 100 times during 200 s and the average is taken to reduce the effect of random fluctuation. RSS vectors of all 145 RPs are randomly selected with a sampling density varying from 10 to 95 % in 5 % step. This way, we constructed a series of low density 10–95 % radio map to find the lower bound of sampling density for a reasonable performance of localization. Parameter k in (55.2), the thickness of boundary lines, is set to be 3.

Figure 55.4a shows the average error in localization for IDW, Voronoi and DPS method. At the point designated as 100 %, no interpolation is performed. We can find that the radio map reconstructed by DPS outperforms IDW-interpolated map and Voronoi-interpolated map. Figure 55.4b shows the a detailed result of DPS with the sampling density varying from 80 to 100 % in 1 % step. The average error continues to decreases as the sampling density increases.

We find a few observations: First, when the sampling density is higher than 35 %, the accuracy in localization is approaching that of the genuine full density radio map. Second, with sampling density higher than 60 %, the performance of DPS-reconstructed radio map is even better than that of the original full density map. The RSS of a full density radio map without interpolation has a large fluctuation and this is a major source of localization error in fingerprinting. In adaptive

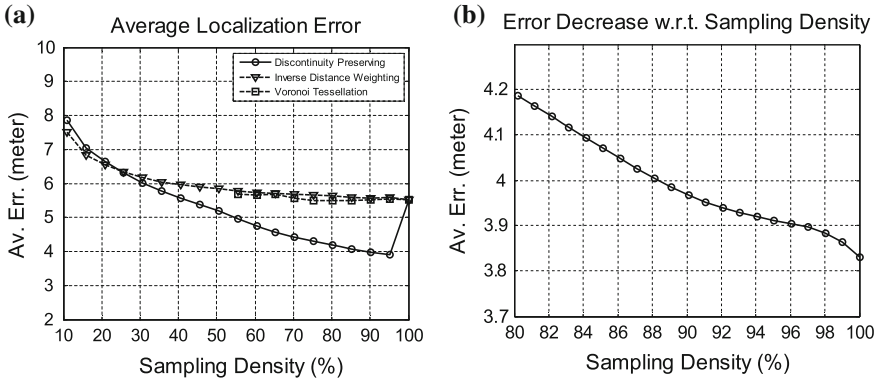


Fig. 55.4 **a** Average localization error. **b** Improvement as sampling density increases

smoothing, the RSS value over all RPs, including the original low-density seed RPs, are iteratively smoothed. In this continuous spatial smoothing, the magnitude of random fluctuation gets further reduced and this brings higher accuracy in localization.

55.5 Conclusion

We presented a method of constructing a high density fingerprinting radio map by interpolating a low density map. We used an adaptive smoothing, which can preserve the discontinuity.

We reconstructed a full density radio map out of low density 10–90 % radio maps, using different interpolation methods: IDW, Voronoi and DPS. DPS radio map provides higher accuracy than IDW map or Voronoi map.

We found that the cost of RSS measurement can be reduced to 35 %. The accuracy of localization is comparable to that of a genuine full density map when the sampling density is higher than 35 %.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (grant number NRF-2013R1A1A2012745).

References

1. Skyhook Wireless Inc. (2007) Location-based services that choose location algorithms based on number of detected access points within range of user device. US Patent 7,305,245 B2, Dec 2007
2. Seidel S, Rapport T (1992) 914 MHz path loss prediction model for indoor wireless communications in multi-floored buildings. IEEE Trans Antennas Propag 40(2):207–217

3. Bahl P, Padmanabhan V (2000) An in-building RF-based user location and tracking system. IEEE Infocom 2:775–784
4. Saint-Marc P, Chen J (1991) Adaptive smoothing: a general tool for early vision. IEEE Trans Pattern Anal Mach Intell 13(6):514–529
5. Tomasi C, Manduchi R (1998) Bilateral filtering for gray and color images. In: Proceedings of the 1998 IEEE ICCV, Bombay, India, 1998
6. Shepard D (1968) A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of the 1968 ACM national conference, pp 517–524
7. Lee M, Han D (2012) Voronoi tessellation based interpolation method for Wi-Fi radio map construction. IEEE Comm Lett 16(3):404–407
8. Bong W (2012) Generation of fingerprint radio map by adaptive smoothing. MS thesis, University of Seoul, Feb 2012

Chapter 56
Sensor Based Ubiquitous Application Modeling Using Process Definition Language

Sang-Young Lee

Abstract The Unified Modeling Language (UML) is a language for specifying, visualizing, constructing, and documenting the artifacts of software systems. It provides rich notations for representing, analyzing architecture and behaviors of systems. Business Process Definition Languages such as XPDL, BPML, and BPEL4WS have been so far specified with different basis and goals. Among them, XPDL proposed by WfMC is considered as an XML-based language for process definition exchange. RFID application domain is an especial environment that requires precise measurement and precision calculation of real-world factor. Therefore in the paper new extended iconic stereotypes for better modeling RFID application in the UML Diagram are proposed, and the Implementation of a program called Stereotype Creator, which is able to create iconic stereotypes used in one of the most popular visual modeling tools for software development, Rational Rose, will be also proposed.

Keywords Unified modeling language · Extension · RFID

56.1 Introduction

The Unified Modeling Language (UML) is a language for specifying, visualizing, constructing, and documenting the artifacts of software systems. It provides rich notations for representing, analyzing architecture and behaviors of systems [1]. Because among these notations, the UML activity diagram is well-known for describing systems' dynamic behaviors, it is useful to model business process and workflow [2]. RFID application domain is an especial environment that requires

S.-Y. Lee (✉)
Department of Health Administration Namseoul University, 21 Maeju-ri, Seongwan-eup, Cheonan, South Korea
e-mail: Sylee@nsu.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_56, © Springer Science+Business Media Dordrecht 2014

precise measurement and precision calculation of real-world factor. The conventional modeling element of class in UML is not powerful enough to present the spatial feature and temporal feature that RFID objects embody. Therefore, new modeling elements of class for RFID application with UML are needed [3]. This paper will propose several new iconic stereotypes formulating modeling elements for RFID application. Also in this paper the implementation of a program called Stereotype Creator, which is able to create iconic stereotypes used in Rational Rose, will be proposed.

56.2 Iconic Stereotypes for Class Meta-Model Element

In Fig. 56.1, the visual notations are used to represent georeferenced classes which are distinguished from conventional classes. Main elements of georeferenced classes are:

- a graphical representation with a symbolistic icon,
- an iconic notation for geographic types (points, lines and polygons),
- the class name,
- attributes,
- operations.

For example, the Fig. 56.2 depicts the visual representation of a class named "Building" with a symbolistic icon and an iconic notation for geographic types (corresponding to a polygon) on the left side of the class name, Building. The polygon symbol means that each object "Building" is associated to a polygon. Attributes are "address" and "inhabiting_area". The only operation associated to the class is "build".

The georeferenced classes allow the representation of complex objects and composed objects. Each complex object is related to several points, lines or polygons. For example, a waterway may be considered as a lake (a polygon) connected to a river (a line). Graphically speaking, overlapped iconic figures express complex objects. An example of a complex object class is given in Fig. 56.2b. Composed objects have several representations in function of scale's point of view. Figure 56.2c presents the graphical notation of a composed class "City". We do not consider composed-complex objects in this paper.

In many cases the value of an attribute of an object varies during all the life cycle of the object [4, 5]. It is possible to associate a temporality to an attribute x of an object a. In that case, during all the life cycle of a, the object is able to "know" all previous values of x. Graphically, an icon representing a clock is placed on the right side of the attribute in the class (Fig. 56.3a). In the same way, a temporality may be associated to the geometry of an object. In that case, an icon representing a clock is placed on the right side of the iconic notation for geographic types (Fig. 56.3b). The Fig. 56.4 depicts the implementation of Class Mobile Station presented in Fig. 56.3a.

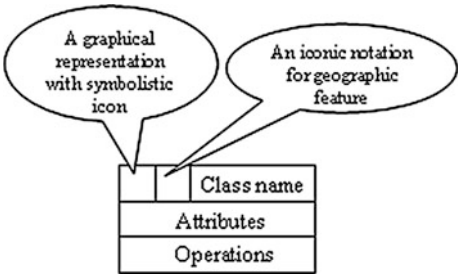


Fig. 56.1 Graphical representation of class meta-model element

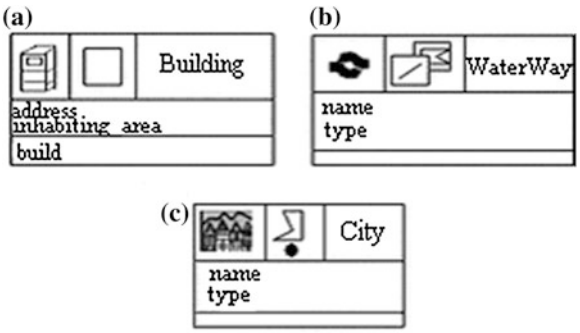


Fig. 56.2 Examples of georeferenced class

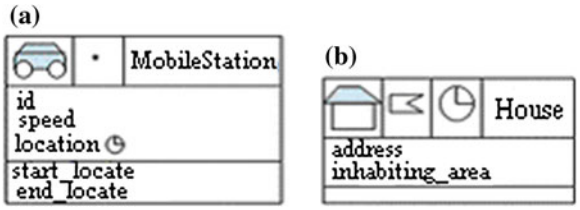


Fig. 56.3 Examples of georeferenced class with temporality

56.3 Implementation of Stereotype Creator

According to the specification of extended iconic stereotype of class meta-modeling element for a georeferenced class in UML class diagram, many kinds of the iconic stereotypes of class can be designed. Stereotype Creator is a software tool used to automatically create all kinds of user-defined iconic stereotypes that can be used in Rational Rose. It can be considered as an extended tool for Rational Rose [6]. The tool can create stereotypes for different modeling elements in Rational Rose that include class, attribute, association, dependency, use case etc. [7].

The whole process is that users input some simple information required to create stereotype, such as diagram icon, small/big toolbar icon, stereotype name

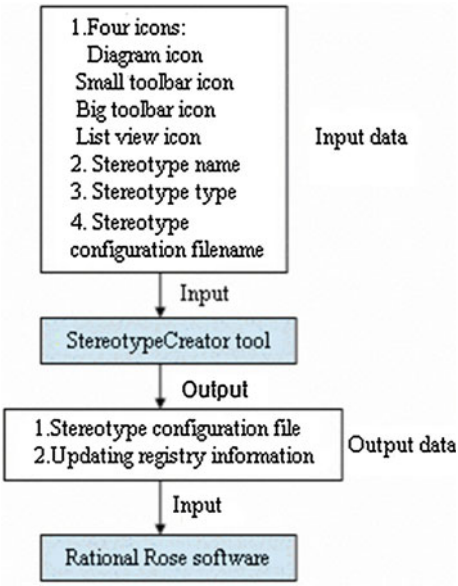


Fig. 56.4 Whole process of stereotype creator

and so on, into Stereotype Creator program, and that Stereotype Creator program processes and converts these data information into output as stereotype configuration file and updating registry information which can be recognized and loaded when Rational Rose software initiates.

56.3.1 Stereotype Configuration File

The format of the stereotype configuration file must be known before the creation of stereotype configuration file. Stereotype configuration file is a text file with extension name of INI. A stereotype configuration file may include one or more stereotypes information.

The Fig. 56.5 presents the general format of stereotype configuration file. In the format of stereotype configuration file, the general section contains the information of specific setting (Fig 56.6).

56.3.2 Update Registry Information

After creating the stereotype configuration file, another task is to update the registry information for stereotype configuration file in order to make stereotype configuration file loaded correctly when Rational Rose software initiates.

Fig. 56.5 The format of stereotype configuration file

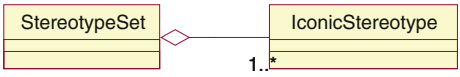


Fig. 56.6 Relationship between stereotype set class and iconic stereotype class

56.4 Conclusion

This paper proposes extended iconic stereotypes of class meta-model element for GNSS application in the UML Diagram and provides its implementation as a tool of Stereotype Creator for Rational Rose. Current research is so little that a great deal of work remains and should be done in the future. Future research can be positioned in customizing iconic stereotypes for other modeling elements such as

association, generalization, attribute and so on, in GNSS application. Now the tool of Stereotype Creator just supports the creation of iconic stereotype for the modeling element of class in class diagram, but in the future Stereotype Creator for Rational Rose will support more kinds of iconic stereotype for different modeling element.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2007598).

References

1. Leick A (1990) GPS satellite surveying. Wiley, New Jersey
2. Bajaj R, Ranaweera SL, Agrawal DP (2006) GPS: location-tracking tehnology. Computer 35(3):92–94
3. Pinet F, Lbath A (2010) Semantics of stereotypes for type specification in UML: theory and practice. Electronic Edition, pp 339–353
4. Borges KAV, Davis CA, Laender AHF (2011) OMT-G: an object-oriented data model for geographic applications. GeoInformatics 5(3):221–260
5. Parent C, Spaccapietra S, Zimanyi E (1999) Spatio-temporal conceptual models: data structure + space + time. In: ACM GIS'99
6. Rational (2011) Using rose, electronic document, VERSION: 2011A.04.00. http://www.rational.com
7. Pinet F, Lbath A (2012) An algorithm for stereotype deduction in UML-based formalism and its application in geographic information systems. In: proceedings of the IEEE symposium on visual/multimedia approaches to programming and software engineering (Human Centric Computing'2012), pp 296–303
8. Booch G, Jacobson Y, Rumbagh J (1999) The unified modeling language user guide. Addison-Wesley, Boston

Chapter 57
Smart Device Locking System Provides Handy and Improved Security

Hyun-Woo Kim and Young-Sik Jeong

Abstract The touch screen, the fruit of recent IT development, is incorporated into many applications through a variety of touch screen smart devices. For smart phones, it provides convenient use of time and space for many people by replacing many desktop PC functions. Although this convenience has gained popularity among the public, the security is usually neglected. Inherent features are generalized, but the features also increase risk of exposing personal information. As a result, smart devices provide a variety of locking features to protect personal information. Although password and pattern lock features exhibit some degree of security, they are vulnerable to shoulder surfing or smudging. In this paper, vulnerable security points of smart devices are complemented and the locking system for security improvement (LSSI), in which intuitive user interface provides convenience, is proposed.

Keywords Smart phone · Smart device · Touch screen · Locking system · Security · Secure smart device

57.1 Introduction

From general user perspective, smart devices should not only carry out basic functions, such as schedule management, memo, phone book, phone, and alarm, but also various additional content, including Internet navigation, Augmented

H.-W. Kim · Y.-S. Jeong (✉)
Department of Multimedia Engineering, Dongguk University,
30 Pildongro 1 Gil, Jung-gu, Seoul, Republic of Korea
e-mail: ysjeong@dongguk.edu

H.-W. Kim
e-mail: z4538@nate.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_57, © Springer Science+Business Media Dordrecht 2014

Reality (AR), Social Networking Service (SNS), e-mail, and network games, to increase convenience in everyday life. From a business perspective, a mobile office is found to be effective in increasing work efficiency and reducing costs [1–7].

Though they may be physically small, they can store a lot of information, which is precisely why various locking systems are built in for data security. However, in spite the various available locking systems, the security function is minimal or the input method is inconvenient. Even if the locking system has a security function, it may not be safe from shoulder surfing or smudge attacks, calling for a new approach to the locking system [8–10].

In this paper, a locking system for security improvement (LSSI) with improved security is proposed to protect the information stored on smart devices.

57.2 Related Works

57.2.1 Previously Developed Locking Systems

Common locking systems in smart devices include patterns, PINs, and passwords. Table 57.1 shows the names of the locking systems built into smart devices and explains their functions.

57.3 Locking System of LSSI

57.3.1 DSD

The starting point of Draw Sequential Direction (DSD) is anywhere within the touch screen. Figure 57.1 shows the process of cognition in the east and southwest.

(1) from Fig. 57.1 shows the first touch. (2) shows dragging to the east from the starting point, and cognition is being made outside the circle. As shown in (3), the new input can be made by re-entering the circular area while still pressing the screen. (4) shows cognition outside the circle for input to the southwest. This makes it possible to input additional directions. These cognitions enhance security by providing variety in the secret patterns and in the ways to input the same pattern.

57.4 Design of LSSI

The LSSI design proposed in this paper is shown in Fig. 57.2.

The user interface consists of a direction, setting, and mode.

Table 57.1 Diverse locking systems built into smart devices and their functions

Embedded locking systems	Lock function
Pattern	Pattern is a common locking system, similar to drag. The number of recognized patterns is limited, and the security strength depends on the locking pattern set by the user. Screen locking through a pattern provides an easy interface and security
Personal identification number (PIN)	The PIN is composed of a combination of numbers, and the weaker the relationship between the numbers and the user, the higher the security. However, since the PIN it is made up of only numbers, its security strength is weaker than that of a password
Password	It can be composed of numbers, letters, special characters, or a combination of these. It offers very strong security, but inputting a password to unlock the locked screen also poses a risk of leakage through screen exposure

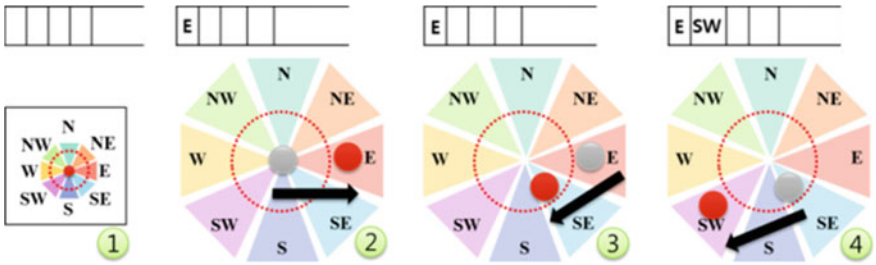


Fig. 57.1 Direction cognition of DSD in LSSI

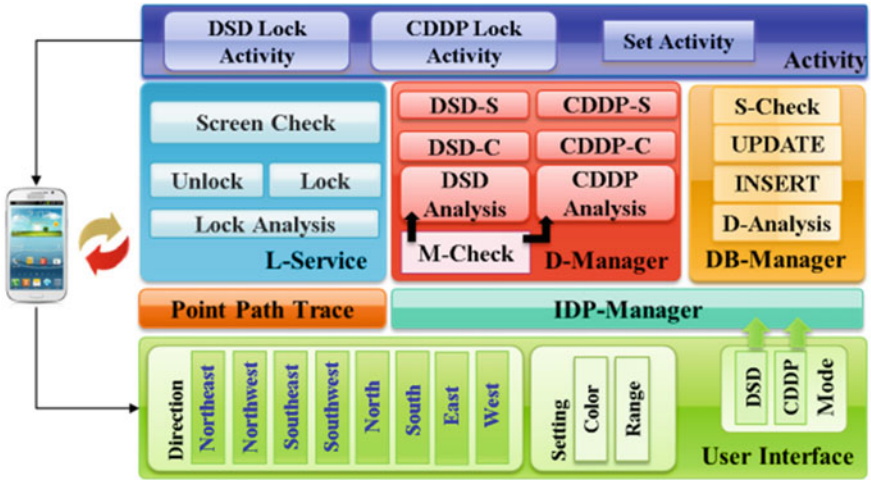


Fig. 57.2 Architecture of LSSI

Fig. 57.3 LSSI initial execution screen

The point path trace plays the role of calculating the distance from the point the user pressed to the point where the user moved in the user interface.

The D-manager is composed of D3D analysis, D3D-C, D3D-S.

The IDP-manager operates according to whether the user set. The random direction pattern is induced to input directionality with low frequency in the user's input to unlock the locked screen.

The DB-manager is composed of function for system database.

L-service is comprised of lock analysis that analyzes whether screen locking has been executed, unlock and lock that are executed according to the lock analysis, and screen check that is intended to perform a test of the movement direction when screen locking is performed.

Activity consists of D3D lock activity, which appears when the user sets up the lock and the smart device or the screen is turned off.

57.5 Implementation of LSSI

Figure 57.3 shows a screen where LSSI has been executed for the first time. The screen consists of a touch area for inputting the pattern, a cancel button for canceling the pattern, a revise button to revise when a directional pattern has been wrongly input, and a continue button to perform the next progression after inputting the directionality.

Figure 57.4 shows a screen when a total of ten directions—north, south, north, south, east, west, east, west, north, and south—are input through the perception of

Fig. 57.4 Screen upon receiving 10 directions from user

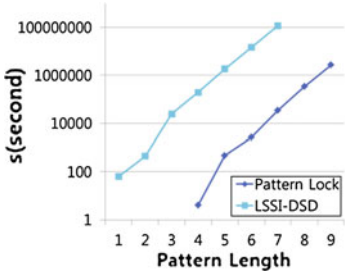


Fig. 57.5 Input time for all possible patterns depending on pattern length

directionality. By showing the input directionality to the user, he/she may verify it, and when the user wishes to revise it, he/she may do so using the revise button at the bottom of the screen.

57.6 Performance Evaluation

Among the generally used locking systems, pattern locks are more secure than drag locks. A pattern lock has limited number of lock pattern it is restricted to 3×3 access points. Figure 57.5 compares the time to input all possible patterns in pattern lock, DSD, assuming that one pattern takes 1 s to input. When a pattern is incorrectly inputted, for five wrong trials, 30 s of waiting time is applied in pattern lock.

57.7 Conclusion and Future Works

LSSI addresses the vulnerability of touch screens to shoulder surfing and smudge attacks. LSSI perceives directionality and, therefore, the touch screen may be reused, enabling the setting of a screen lock with many directional patterns. When the screen is locked, LSSI provides several methods for unlocking it that differ from the previous method of inputting the same pattern; therefore, it is difficult for a malicious person to infer the pattern.

In the future, the researcher aims to study a middleware that may transmit the location of the device to a contact number set by the user.

Acknowledgments This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2013-H0301-13-4007) supervised by the NIPA (National IT Industry Promotion Agency).

References

1. Mulliner C, Vigna G, Dagon D, Lee W (2006) Using labeling to prevent cross-service attacks against smart phones. In: DIMVA 2006, LNCS 4064, pp 91–108
2. Carey M (2005) Enterprise risk management: how to jumpstart your implementation efforts. International Risk Management Institute
3. Park M (2011) The evolution of the mobile phones with touchscreen and the prospect of future: focused on the SRI-Tech. Master thesis, Incheon University
4. Wang G, Zhou W, Yang LT (2013) Trust, security and privacy for pervasive applications. J Supercomput 64(3):661–663
5. Jaegal B (2010) Trend of smartphone market and mobile OS. Samsung Wireless Division
6. Gong Y (2010) Implications and agreement of smartphone. Korea Inf. Soc. Dev. Inst. 22(4):480
7. Kim C, Yoon S, Lee M (2010) Shoulder-surfing resistant password input method for mobile environment. J. Korea Inst Inf Secur Cryptol 20(3):93–104
8. ITU-T (2010) Security aspects of mobile phones. T09 SG17 100407 TD PLEN 1012
9. Peng K (2013) A secure network for mobile wireless service. J Inf Process Syst 9(2):247–259
10. Ahn J, Han R (2012) An indoor augmented-reality evacuation system for the smartphone using personalized Pedometry. Human-centric Comput Inf Sci 2(1):1–23

Chapter 58
A Study on the Efficient Management of Android Apps Using a Whitelist

Uijin Jang, Changjin Suh and Hyungmin Lim

Abstract Currently, many people use the smart phone throughout the world, and the government and enterprise also use it to improve business efficiency. However, there is a risk that the confidential information of the government or enterprise can be disclosed, if the malicious or forged/altered App is installed and executed in the smart phone used for work. Apps distributed by the domestic App Store are verified in advance, using the independent verification system of each App Store. However, those App Stores don't use the common verification system, and their verification system is lacking security features. Besides, the Android Apps can be installed and used, using the alternative method (Android black market and APK file), instead of the official App Store. As a result, the use of the non-verified App is rarely controlled. To prevent the alternative method, mobile device management (MDM) and other methods are employed but are not quite effective. To supplement these shortcomings, this paper proposes a method of providing a reliable smart phone environment to users, by verifying the safety of the App using the common verification method for Android phones, and managing the verified App with a whitelist technology, and testing the installed App periodically.

Keywords Proxy signature scheme · ID-based partially blind signature · Proxy partially blind signature · Gap Diffie-Hellman problem · Bilinear-pairing

U. Jang
Korea Copyright Commission, Department of Computer Engineering, Soongsil University, Seoul, Korea
e-mail: neon7624@gmail.com

C. Suh · H. Lim (✉)
School of Computer Science and Engineering, Soongsil University, Seoul, Korea
e-mail: atskyo@ssu.ac.kr

C. Suh
e-mail: cjsuh@ssu.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_58, © Springer Science+Business Media Dordrecht 2014

58.1 Introduction

With the appearance of the smart phone, the trend of mobile device and computing has changed, and a new industry was formed. These changes began to occur from 2001 in full scale, and the hardware specification of the smart phone has recently been improved to the level similar to the PC. In addition, the government and enterprise create and utilize many Apps to improve business efficiency using the smart phone, as the smart phone market expands rapidly.

The App for administrative works developed by the government to carry out various administrative works using mobile devices is secure, because security features are sufficiently reviewed in advance. However, there is a risk that the App for administrative works can be forged and altered due to the malicious App installed in the smart phone. There is another risk that the confidential information can be leaked due to the malicious App and forged and altered App. The App can be installed in the Android phone using the App market, or can be directly installed using the APK file.

App Store performs a pre- or post-verification procedure to see if the App is secure or not, when registering a program of the developer in the Mobile App Store. Google "Android Market" performs a post-verification, whereas domestic App Stores (App Store managed by mobile operators such as LG U+, KT, and SKT; and the market managed by Samsung and Naver) adopt a prior-verification system to filter out the malicious App. However, malicious App and forged/altered App are installed, when the App is directly installed using the black market and APK file, which can cause many problems in smart phone security. To solve these problems, this paper proposes a method of managing Apps securely using the whitelist technology, among many security technologies.

58.2 Related Research

58.2.1 Mobile Administrative Service

Mobile services can be grouped into the mobile administrative service targeting civil servants and the mobile public service targeting the public, according to the service receiver. The mobile administrative service can be divided into the mobile internal administrative service, which migrate the PC-based service to the mobile-based service, and the mobile field administrative service for field work support. The internal administrative service, field administrative service, and public service of the mobile administrative service can be divided into several phases (3, 2, and 3 phases respectively). The service that can be accessed with simple authentication or without authentication for the person who receives a simple retrieval service belongs to the first phase service. The second phase service refers to the service that requires identity confirmation, such as retrieving the vehicle with unpaid tax,

making a reservation, and providing civil service results. The third phase refers to the service that requires strong identity confirmation. For identity confirmation, a public certificate is used by default, and a security card, security token, or OTP can be used as an additional means. The third phase service includes accessing the electronic approval system, retrieving tax payment, and retrieving the sensitive information. For the public service, phases are mixed by type.

58.2.2 Mobile Malicious Code

Mobile malicious code means a malicious program that performs vicious behavior in the mobile device, such as the information disclosure, terminal break-down, and illegal billing. As an environment is created that can perform many jobs using the smart phone, risks also are increasing proportionately, such as the information disclosure, and unauthorized billing and payment. Recent smishing incidents can be a good example. Table 58.1 shows the type of mobile malicious code. Creation and distribution of malicious code became easier with the increase of open type terminals.

In addition, mobile malicious code is spreading via various infection ways as described below, due to the diversification of the external access method.

1. Distributing malicious code on the online market by disguising as a normal App.
2. Mutual infection through direct connection between the PC and smart phone.
3. Infection by communication like Bluetooth.
4. Malicious code infection that exploits browser's vulnerabilities.
5. Sending manipulated packets by exploiting network vulnerabilities (e.g., WiFi).

58.2.3 Mobile Device Management

The MDM system can manage the mobile device remotely at any time, and anywhere, using the Over the Air (OTA) technology, if the mobile device power is turned on. The demand on MDM is on the rise, as the diffusion of smart phones has recently increased sharply and the mobile device becomes diversified (e.g., tablet PC). Also, the required function of MDM is also diversifying. There are many configuration methods for MDM, depending on the mobile phone OS. Generally, MDM is composed of a gateway server, MDM management server, MDM registration server, and MDM agent.

Table 58.1 Type of mobile malicious code

Type	Symptom	Example of mobile malicious code
Terminal failure causing type	Functions other than a call cannot be used	Skulls
	Keys cannot be used	Locknut
	A call cannot be placed	Gavno
Battery consuming type	Proliferated through Bluetooth. Continuous Bluetooth scanning	Cabir
Billing causing type	SMS sending	RedBrowser
	Displaying a payment request screen	Kiazha
Information leak type	Terminal security setting change	Infojack
	Sending the serial number, and OS and application information to the outside	
	Spyware type malicious code that sends call logs and SMS to a particular server	Flexspy, PBStealer
Cross platform type	Spreading to the PC through a memory card and other media	Cardtrap.A

58.2.4 Blacklist and Whitelist

A blacklist function gives a warning and blocks the execution of the malicious App, by comparing the App identified as malicious code and the App installed in the user's smart phone. It is important to add a newly detected malicious App to the blacklist App database, besides the existing malicious App. On the contrary, a whitelist function allows the execution of the App that was proven to be safe only. That is, all Apps except the allowed App will be blocked, if the whitelist function is used.

58.3 The Proposing System

The proposing system uses whitelist security technology to provide a safe App to the user.

As shown in Fig. 58.1, the proposing system is composed of the App management program installed in the terminal for the whitelist, whitelist interface server, management server, App verification system, and MDM server and agent for Android terminal control. Table 58.2 describes the components of the proposing system, and the function of each component. To prevent the installation and execution of the malicious App, the proposing system puts the App into the verification system and saves/manages the result in the whitelist, even though the App has already been verified in App Store. Also, the proposing system allows the user to install the App in the whitelist only. If the App has been forged/altered, or downloaded from the abnormal route, the App will be checked whether it is included in the whitelist. If the App is not on the whitelist, its execution is

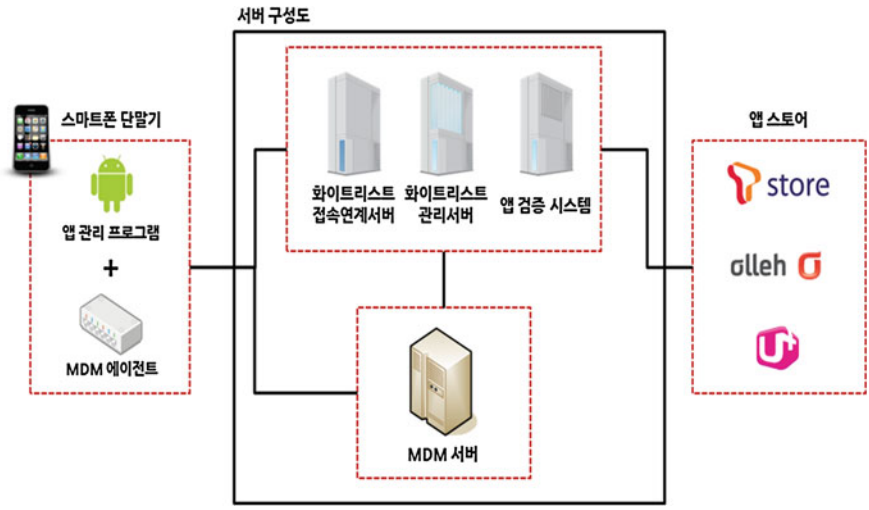


Fig. 58.1 Configuration diagram of the proposing system

Table 58.2 Components of the proposing system

Type	Component	Description
Whitelist	App management program	Installed in a terminal to control the function related to a whitelist
	Whitelist interface server	The interface server between the App management program and whitelist management server
	Whitelist management server	The server used to manage the whitelist system directly
	App verification system	The system that verifies whether the Android App is malicious or not
MDM	MDM agent	Android mobile control
	MDM server	

prohibited. In addition, MDM checks whether the registered App matches the App installed and used in the user's Android system, by comparing the hash of the two Apps. If the hash doesn't match, App deletion is recommended to the user.

58.4 Functional Evaluation

The proposing system is designed to prevent the installation and execution of the App containing malicious code, and forged/altered App, by managing the App using a whitelist. Currently, vaccine software, forgery and alteration prevention

Table 58.3 Items to compare with the proposing system

No.	Type	Comparative item
1	Malicious App	Controlling malicious App installation
2		Controlling malicious App execution
3	Forged and altered App	Controlling the execution of the forged and altered App
4	Convenience	User's use convenience in using the mobile App

system at App Store, and MDM are used to prevent malicious code and forged and altered App. Comparing with the proposing system, those solutions have shortcomings.

Table 58.3 shows the main functions of the proposing system to compare with other systems. If the items in the "Items to compare" are blocked, the malicious App installed in the Android smart phone, as well as security threats can be fundamentally prevented. The comparison target for the functional evaluation includes the App verification system of App Store that is currently used, illegal copy prevention system, vaccine for Android, and MDM.

58.4.1 App Verification System of App Store

Domestic Android App Stores (Olleh Market, T Store, and U+ Store) adopt a prior verification system to verify the App.

Table 58.4 shows the verification items by App Store. Security check on malicious code is performed, when inspecting the App before registering it. Olleh Market and T Store check virus and malicious code during the security verification and operation verification phase respectively.

The verification system of the App Store can prevent the installation of the malicious App, as the registration of the App infected by malicious code can be screened in advance. However, the system cannot affect the App, once it is successfully installed. Therefore, it is weak at preventing the execution of forget/altered App, if such App has already been installed.

58.4.2 Illegal Copy Prevention System of App Store

Domestic App Stores prevent illegal copy prevention methods for each store, to prevent illegal copy. Table 58.5 shows the name and method of the illegal copy prevention technology by App Store.

The illegal copy prevention technology of the App Store doesn't provide the function of controlling App installation, and preventing the execution of the malicious App but can control execution of the forged/altered App. However, there is a shortcoming that the illegal copy prevention technology provided by App

Table 58.4 The verification items by App store

Market name	Verification item
Olleh market	1. Registration information 2. Policy/harmful features 3. Function 4. Network 5. Signaling 6. Security 7. Billing
T store	1. System 2. Harmful features 3. Operation
U+ store	1. The developer verifies the source/test coverage using a code verification tool 2. Registering the report result file 3. Review using the report file

Table 58.5 The name and method of the illegal copy prevention technology by App store

	Olleh market	T store	U+ store
Name	KT Android illegal copy prevention	Application DRM (ARM)	Application rights management (ARM)
Illegal copy prevention method	1. Check illegal copy by accessing the license server 2. Issue and save the license of the App, if the App is normally purchased 3. Afterwards, check illegal copy using the saved license when running the App		

Store calls the API within the App, and the application method is opened. Therefore, forged and altered App can be distributed using various detour techniques, because the Android App can be easily decompiled, due to the characteristics of the Android App.

In conclusion, the illegal copy prevention technology of the App Store can be easily bypassed but the execution of the forged and altered App can be controlled. However, control over App installation doesn't exist.

58.4.3 Android Vaccine

The Android vaccine has more diverse function than vaccine S/W for the regular PC. Table 58.6 shows the test items used to certify the Android vaccine in AV-Test.

The items checked by the AV-Test tell us that the malicious code detection function is checked but the forged/altered App is not detected, if the App contains malicious code. The malicious App can be detected during installation and

Table 58.6 The test items used to certify the Android vaccine in AV-Test

Type	Check item
Protection	Detection of a representative set of malicious Apps discovered in the last 4 weeks (AV-Test reference set)
Usability	Performance: The App does not impact the battery life
	Performance: The App does not slow down the device during normal usage
	Performance: The App does not generate too much traffic
	False warnings during installation of legitimate software from Google Play and their usage
Features	Anti-theft (Remote-Lock/Remote-Wipe/Locate): locate, lock or wipe your device when it is lost or stolen
	Call blocker: Block calls from specific or unknown numbers
	Message filter: Filter messages and/or mails for unwanted content
	Safe browsing: Protection of malicious websites and/or against phishing
	Parental control: Features to control or observe the activity of children on the device
	Backup: Personal data can be saved to SD-card or cloud storage
	Encryption: Any kind of encryption is supported (e.g. device encryption, SD-card encryption or VPN)

execution, if the vaccine program is available. However, installation and execution cannot be controlled. The user should control App installation and execution manually, when malicious App is detected.

58.4.4 MDM

MDM was introduced in Chap. 2, "Related research". Table 58.7 shows the App management functions of MDM.

The malicious App and forgery/alteration can be controlled, using a blacklist and whitelist. In addition, as MDM provides a function of deleting the malicious App and forged/altered App; and the installation and execution of the malicious App and forged/altered App can be controlled.

However, there is a shortcoming that the malicious App can be executed until the App is deemed malicious and registered in the blacklist, if the App is registered with a blacklist. It is also same for the forged and altered App until the App is registered in a blacklist. If the App is managed in a whitelist, user's convenience drops significantly, because only the App registered in MDM can be used.

Table 58.8 compares the proposing system with the existing system, based on the description provided above.

The proposing system makes the use of MDM, as described in Chap. 3 (Configuration of the proposing system). Therefore, MDM's App control functions can be used without changes. In addition, user's convenience can be improved by checking and registering the malicious App that is wanted by the user, after requesting the management App to register the App wanted by the user in a whitelist server.

Table 58.7 The App management functions of MDM

Type	Function
App management	Blacklist control
	Whitelist control
	Remote installation and deletion
	File distribution management

Table 58.8 Results of comparing with the proposing system

	Controlling malicious App installation	Controlling malicious App execution	Controlling the execution of the forged and altered App	User's use convenience in using the App
Whitelist	Supported	Supported	Supported	Normal
App store App verification system	Supported	Not supported	Not supported	Convenient
Illegal copy prevention function of the App store	Not supported	Not supported	Partially supported	Convenient
Android vaccine	Detectable (manual control)	Detectable (manual control)	Not supported	Convenient
MDM	Supported	Supported	Supported	Inconvenience

58.5 Conclusion

Although there are many ways of preventing the installation and execution of malicious App in the Android App, and execution of the forged/altered App, existing methods have their shortcomings, as described in the functional evaluation section of Chap. 4.

The proposing method of managing the whitelist App designed a more perfect management method, by adding a new procedure using MDM, in order to supplement existing methods. This system can prevent the execution of the malicious App and forged/altered App, and the execution of the malicious App installed in the smart phone to the maximum extent, by managing Android Apps using the MDM whitelist, while maximizing convenience improvement.

This method is quite effective when the App with strong security requirements is used, such as the App using the e-Government service, and financial transaction App. If this management method is institutionalized and its use is forced, reliability on mobile services will be enhanced significantly. To achieve the goal, research on the systemization is required that fit into each field (finance and administration). In addition, more research is needed to make use of other OS (e.g., iOS, Windows Mobile), besides Android.

Acknowledgments This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the Convergence-ITRC (Convergence Information Technology Research Center) support program (NIPA-2013-H0401-13-1004) supervised by the NIPA.

References

1. Kang H et al (2012) Mobile broadband and mobile business model. Korea Information Society Development Institute
2. AhnLab (2011) AhnLab V3 Mobile+ for transaction, 5
3. Cheon W et al (2012) Mobile security diagnosis system to cope with smart phone malicious code. J. Korean Inst. Inf. Technol. 22:537–544 (3rd edition)
4. Tatli EI, Stegemann D, Lucks S (2005) Security challenges of location-aware mobile business
5. Lee K, Tolentino RS, Park GC, Kim YT (2010) A study on architecture of malicious code blocking scheme with white list in smartphone environment
6. Ongtang M, McLaughlin S, Enck W, McDaniel P (2009) Semantically rich application-centric security in Android
7. McAfee (2013) McAfee threat report: fourth quarter 2012
8. Reddy N, Jeon J, Vaughan J, Millstein T, Foster J (2011) Foster application-centric security policies in unmodified Android
9. Krishnan P, Hafner S, Zeiser A (2011) Applying assurance techniques to a mobile phone application

Chapter 59
A Workflow Scheduling Technique to Consider Task Processing Rate in Spot Instance-Based Cloud

Daeyong Jung, JongBeom Lim and Heonchang Yu

Abstract Cloud computing is a computing paradigm in which users can rent computing resources from service providers according to their requirements. A spot instance in cloud computing helps a user to obtain resources at a lower cost, but it may be unreliable. In this paper, we propose a workflow scheduling scheme to consider task processing rate. This scheme reduces the task waiting time and the rollback time when an out-of-bid situation occurs in an instance.

Keywords Cloud computing · Spot instances · Workflow · Price history · Fault tolerance

59.1 Introduction

In recent years, due to the increased interest in cloud computing, many cloud projects and commercial systems such as Amazon EC2, have been implemented. In most cloud services, the concept of an instance unit is used to provide users with resources in a cost-efficient manner. Generally, instances are classified into on-demand instances and spot instances. On-demand instances allow the user to

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2013056346).

D. Jung · J. Lim · H. Yu (✉)
Department of Computer Science Education, Korea University, Seoul, Korea
e-mail: yuhc@korea.ac.kr

D. Jung
e-mail: karat@korea.ac.kr

J. Lim
e-mail: jblim@korea.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_59, © Springer Science+Business Media Dordrecht 2014

pay for computing capacity by the hour, with no long-term commitments. This frees users from the costs and complexities of planning, purchasing, and maintaining hardware, and transforms what are usually large fixed costs into much smaller variable costs [1].

The spot instance is configured by spot market-based cloud environment. In the spot instance environment, variations of spot prices are depended on the supply and demand of spot instances. The environment affects the successful completion or failure of tasks depending on the variation of spot prices. Spot prices have a market structure and follow the law of demand and supply. Therefore, cloud services (Amazon EC2) can provide a spot instance when a user's bid is higher than the current spot price. Further, a running instance stops when a user's bid becomes less than or equal to the current spot price. After a running instance stops, it restarts when a user's bid becomes greater than the current spot price [2–4].

There is the problem that the preformed task may fail according to current spot price. To solve the problem, our proposed scheme is expanded from our previous work [5] and includes a workflow scheduling algorithm. In our proposed method, the system immediately stops the spot instance when a user bid becomes less than or equal to the current price. We refer to this situation as an out-of-bid event or a failure. The out-of-bid situation increases the execution time of the entire task. In existing proposed checkpoint techniques, checkpoints are taken considering the expected failure point of an instance. The expected failure point of each instance is calculated based on the average execution time of the task. Therefore, the existing scheme reduces the loss of task occurred due to the fluctuations of the spot price and the total task execution time due to decrement of the recovery time. However, the instance stops when the spot price is higher than the user's bid and the task waiting time is occurred until the instance restarts. To solve this problem, we propose a workflow scheduling scheme considering task processing rate.

59.2 Related Works

In cloud computing, recent researches have been focused on two different environments such as reliable environments (with on-demand instances) and unreliable environments (with spot instances). Our study is based on the latter. Typically, spot instances are used in unreliable environments, and studies on spot instances focus on performing tasks at low monetary costs. The spot instances in the Amazon EC2 are offered at lower prices at the expense of reliability [2]. Cloud exchange [6] provides the actual price history of EC2 spot instances.

A workflow is a model that represents complex problems with structures such as Directed Acyclic Graphs (DAG). Workflow scheduling is a kind of global task scheduling as it focuses on mapping and managing the execution of interdependent tasks on shared resources. The existing workflow scheduling methods have limited scalability and are based on centralized scheduling algorithms. Consequently, these methods are not suitable for spot instance-based cloud computing. In spot

instances, the available time and instance costs have to be considered for job execution. Fully decentralized workflow scheduling systems use a chemistry-inspired model to determine the instance in a community cloud platform [7]. The throughput maximization strategy is designed for transaction-intensive workflow scheduling that does not support multiple workflows [8]. Our proposed scheduling scheme guarantees an equal task distribution across available instances in spot instance-based cloud computing. Our workflow scheduling method distributes tasks considering the task execution time in the available instances. And, the scheduling method performs redistribution of the tasks based on task processing rate.

59.3 Workflow Scheduling Method

59.3.1 System Architecture

Our proposed scheme expands on our previous work [5] and includes a workflow scheduling algorithm. Figure 59.1 presents the relationship between workflows and instances, and illustrates the roles of the instance information manager, the workflow manager, and the resource scheduler. The instance information manager obtains the information required for job allocation and resource management. The workflow manager and resource scheduler extract the necessary execution-related information from the instance information manager. First, the workflow manager generates the workflow for the request job. The generated workflow determines the task size according to the VM performance, the execution time and costs, and the failure time in the selected instance. Second, the resource scheduler manages the resource and allocates the task to handle the job.

59.3.2 Workflow Scheduling Technique Considering Task Processing Rate

The scheduling scheme is depicted in Fig. 59.2. The instances I_i, I_t, and I_k have high, medium, and low performance, respectively. The instance I_k belongs to a positive group and the other two instances (I_i, I_j) belong to a negative group. The scheduler distributes a task size to allocate available instances and considers performance of instances. Task size recalculation points divide the fourth quarter based on the expected task execution time and recalculate each quarter to except the last quarter. The task size rate is determined based on the average of task execution time of each instance by the recalculated point. And, the modified task size in each instance allocate to consider the task size rate.

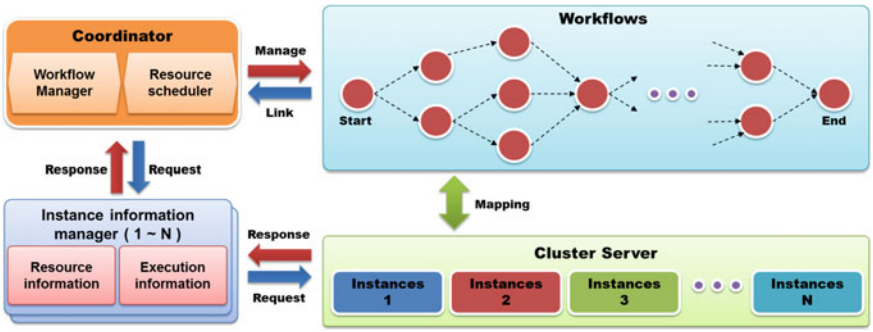


Fig. 59.1 The mapping relation of workflows and instances

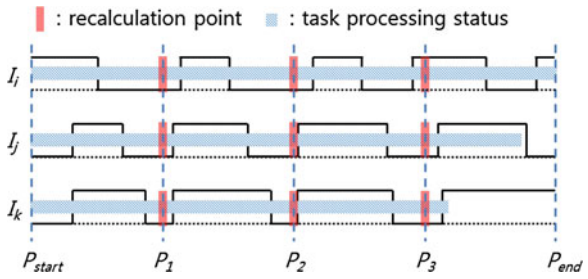


Fig. 59.2 The recalculation point of the task size

Figure 59.3 shows the recalculation point of the task size from the P_1 position in Fig. 59.2. In Fig. 59.3, we assume the processing rate of instances is proportional to the performance of instances. The left side of Fig. 59.3, 'before recalculation', represents the tasks assigned to each instance. The right side, 'after recalculation with relocation', shows the result of task migration based on the average task execution time in each instance. After a recalculation operation, we perform the rearrangement of tasks. The rearrange method sorts tasks in increasing order of their indices.

Our task distribution method determines the task size in order to allocate a task to a selected instance. Based on a compute-unit and an available state, the task size of an instance $I_i(|T_{I_i}|)$ is calculated as follows.

$$|T_{I_i}| = \left(\frac{U_{I_i} \times A_{I_i}}{\sum_{i=1}^{N} (U_{I_i} \times A_{I_i})} \right) \times \frac{1}{U_{I_i}} \times |T_{request}| \times U_{baseline}$$

where $|T_{request}|$ represents the total size of tasks required for executing a user request. In an instance I_i, U_{I_i} and A_{I_i} represent the compute-unit and the available state, respectively. The available state A_{I_i} can be either 0 (unavailable) or 1 (available).

We use the instance rate $T_{rate}^{I_i}$ for determining the criteria to divide groups. $T_{rate}^{I_i}$ represents the unit taken for the processing of a task size in the instance I_i.

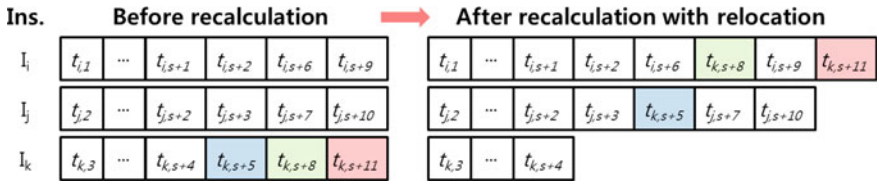


Fig. 59.3 The recalculation operation of the assigned task

$$T_{rate}^{I_i} = \left(T_{execution}^{I_i} + T_{failure}^{I_i}\right) / T_{execution}^{I_i}$$

where $T_{execution}^{I_i}$and $T_{failure}^{I_i}$ represent the task execution time and the task failure time, respectively.

And, we define the *avg* to classify groups. The *avg* is the average of available instances such as T_{rate}^{avg} and T_{avg} represent the average of the $T_{rate}^{I_i}$ and T_{I_i}, respectively. The set of instances is classified into two groups—positive and negative—based on T_{rate}^{avg}. The positive group G_P is the set of instances with $T_{rate}^{I_i}$ greater than T_{rate}^{avg}.

$$G_P = \left\{I_i | T_{rate}^{I_i} \geq T_{rate}^{avg}, \quad 1 \leq i \leq N\right\}$$

We calculate the task size to transfer from instance $I_i(Tr_{I_i})$ in G_P as follows:

$$Tr_{I_i} = \left[\left(|T_{I_I}| - T_{execution}^{I_i}\right) \times I_{rate}^{I_i} - \left(|T_{avg}| - T_{execution}^{avg}\right) \times I_{rate}^{avg}\right] / I_{rate}^{I_i}, \quad 1 \leq i \leq N$$

In the group G_P, the task size of each instance I_i is given as $|T'_{I_i \in G_P}|$. We are able to get $|T'_{I_i \in G_P}|$ by considering Tr_{I_i} after the transfer operation.

$$I_{rate}^{avg} \times \left|T_{avg}\right|,$$

The negative group G_N is the set of instances I_i with $T_{rate}^{I_i}$ less than T_{rate}^{avg}.

$$G_N = \left\{I_i | T_{rate}^{I_i} < T_{rate}^{base}, \quad 1 \leq i \leq N\right\}$$

The tasks are allocated according to the instance performance U_{I_i}. The task size to receive, R_{I_i}, is allocated according to the task size of each instance I_i. In the group G_N, the task size of each instance is given as $|T'_{I_i \in G_N}|$. After the receive operation, R_{I_i} is added to $|T_{I_i \in G_N}|$.

$$R_{I_i} = \frac{U_{I_i}}{\sum_{i \in G_N} U_{I_i}} \times \sum\nolimits_{i \in G_P} (Tr_{I_i} \times U_{I_i}) \times \frac{1}{U_{I_i}}, \quad 1 \leq i \leq N$$

$$|T'_{I_i \in G_N}| = |T_{I_i \in G_N}| + R_{I_i}, \quad 1 \leq i \leq N$$

```
1:  Boolean S_flag = false  // a flag representing occurrence of a task execution
2:  while (search user's job) do
3:   if  (require job execution by the user) then
4:     take the cost and total execution time by the user;
5:     S_flag = true;
6:   end if
7:   if (S_flag) then
8:    invoke  initial_workflow ( ); // thread function
9:   while (task execution does not finish) do
10:   if(meet the recalculation point by instance)then
11:    invoke  recalculation_workflow ( ); // thread function
12:   end if
13:  end while
14:  end if
15: end while
```

Fig. 59.4 Workflow scheduling algorithm

```
1:  Thread_Function initial_workflow ( ) begin
2:    forall instance I_i ∈ Ins do
3:     retrieve an instance information to meet the user's requirement in an instance I_i;
4:     analyze an available execution time and cost in an instance I_i;
5:     store the analyzed available instance to a queue_instance;
6:    end forall
7:    calculate on priority list for the priority job allocation;
8:    forall instance  I_i ∈ queue_instance do
9:     allocate tasks to the instance I_i ;
10:   end forall
11: end  Thread_Function
12: Thread_Function recalculation _workflow ( ) begin
13:   forall instance I_i ∈ Ins do
14:
15:    retrieve the information  T_rate^I_i  to an instance I_i;
16:    calculate the modified task size;
17:   end forall
18: end  Thread_Function
```

Fig. 59.5 The initial and recalculation workflow function

In addition, we compare the expected execution time of *base* instance, $I_{rate}^{avg} \times \left|T_{avg}\right|$, with that of an instance I_i, $I_{rate}^{I_i} \times |T_{I_l}|$. The exclusive group G_E is the set of instances I_i with $I_{rate}^{I_i} \times |T_{I_l}|$ greater than $I_{rate}^{avg} \times \left|T_{avg}\right|$.

$$G_E = \left\{ I_i | I_{rate}^{I_i} \times |T_{I_l}| > I_{rate}^{avg} \times |T_{avg}|, \quad 1 \le i \le N \right\}$$

Table 59.1 Information of resource types

Instance type name	Compute unit	Virtual cores	Spot price min	Spot price average	Spot price max
m1.small (standard)	1 EC2	1 core (1 EC2)	$0.038	$0.040	$0.053
m1.large (standard)	4 EC2	2 cores (2 EC2)	$0.152	$0.160	$0.168
m1,xlarge (standard)	8 EC2	4 cores (2 EC2)	$0.076	$0.080	$0.084
c1.medium (high-CPU)	5 EC2	2 cores (2.5 EC2)	$0.304	$0.323	$1.52
c1.xlarge (high-CPU)	20 EC2	8 cores (2.5 EC2)	$0.532	$0.561	$0.588
m2.xlarge (high-memory)	6.5 EC2	2 cores (3.25 EC2)	$0.532	$0.561	$0.588
m2.2xlarge (high-memory)	13 EC2	4 cores (3.25 EC2)	$0.532	$0.561	$0.588
m2.4xlarge (high-memory)	26 EC2	8 cores (3.25 EC2)	$1.064	$1.22	$1.176

Table 59.2 Parameters and values for simulation

Simulation parameter	Task time interval	Baseline	Distribution time	Merge time	Checkpoint time	Recovery time
Value	43,200 (s)	m1.xlarge	300 (s)	300 (s)	300 (s)	300 (s)

We calculate the task size to exclude from instance I_i (E_{I_i}) in G_E as follows:

$$E_{I_i} = I_{rate}^{I_i} \times |T_{I_I}| - I_{rate}^{avg} \times |T_{avg}|, \quad I_i \in G_E,\ 1 \le i \le N$$

In the group G_E, the task size of each instance I_i is given as $|T'_{I_i \in G_E}|$. We are able to get $|T'_{I_i \in G_E}|$ by considering E_{I_i} after the transfer operation.

$$|T'_{I_i \in G_E}| = |T_{I_i \in G_E}| - E_{I_i},\ 1 \le i \le N$$

We calculate the task size to transfer from instance I_i (ER_{I_i}) excluded instances in the group G_E among instances in G_N as follows:

$$ER_{I_i} = \frac{U_{I_i}}{\sum_{i \in G_N, i \notin G_E} U_{I_i}} \Bigg/ \times \sum\nolimits_{i \in G_E} (E_{I_i} \times U_{I_i}) \times \frac{1}{U_{I_i}}, \quad 1 \le i \le N$$

For instances excluded instances in the group G_E among instances in G_N, the task size of each instance I_i is given as $|T'_{I_i \in G_N,\ I_i \notin G_E}|$. We are able to get $|T'_{I_i \in G_N,\ I_i \notin G_E}|$ by considering ER_{I_i} after the exclusive operation.

$$|T'_{I_i \in G_N,\ I_i \notin G_E}| = |T_{I_i \in G_N,\ I_i \notin G_E}| + ER_{I_i}, \quad 1 \le i \le N$$

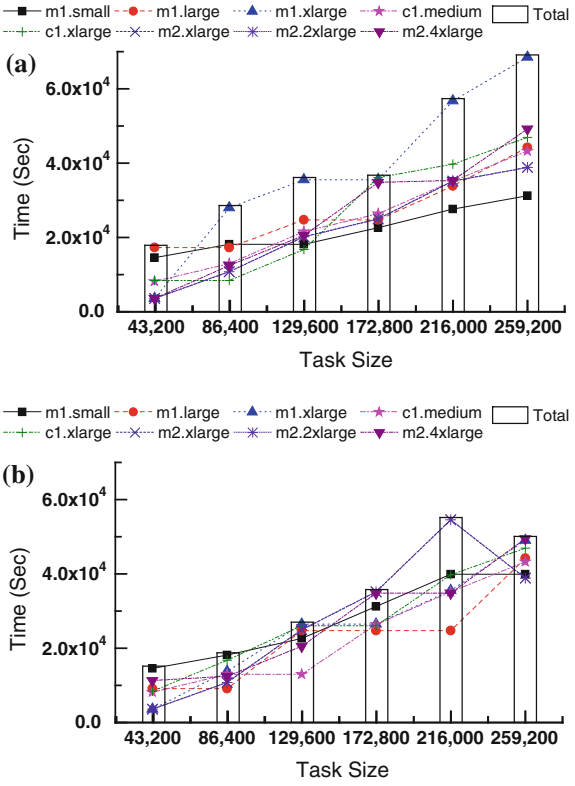


Fig. 59.6 Comparison of total execution time

We propose a workflow scheduling algorithm based on the above equations. Figures 59.4 and 59.5 show the workflow scheduling algorithm and the workflow recalculation function, respectively.

59.4 Performance Evaluation

The simulations were conducted using the history data obtained from Amazon EC2 spot instances [6]. The history data before 10-01-2010 was used to extract the expected execution time and failure occurrence probability for our checkpointing scheme. The applicability of our scheme was tested using the history data after 10-01-2010.

Table 59.1 shows the various resource types used in Amazon EC2 and in our simulation. In this table, resource types are comprised of a number of different instance types. First, standard instances offer a basic resource type. Second, high-CPU instances offer more compute units than other resources and can be used for

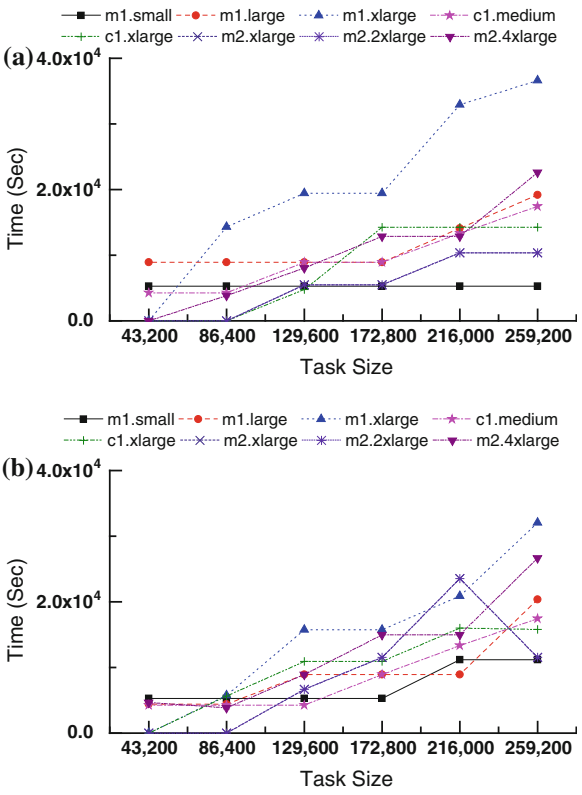


Fig. 59.7 Comparison of failure time

computation-intensive applications. Finally, high-memory instances offer more memory capacity than other resources and can be used for high-throughput applications such as databases and memory caching applications. In the simulation environments, we compared the performance of our proposed scheme with that of the existing schemes without considering task distribution based on the task time.

Table 59.1 shows the information related to resource types in each instance, and Table 59.2 shows the parameters and values for the simulation. The spot price information was extracted from the spot history data from 11-30-2009 to 01-23-2011. The user's bid was taken as the average of the spot prices from the spot history data. The task size was decided by the compute-unit rate based on the baseline. Initially, the baseline denoted an instance m1.xlarge. Using this baseline, the task size of an instance m1.small was calculated as follows:

$$|T_{m1.small}| = U_{m1.xlarge}/U_{m1.small} \times |T_{original\,task}|$$

Figures 59.6, 59.7, 59.8, and 59.9 show the execution results of the actual data based on the task processing rate, before and after applying our proposed scheme. In the figures, Total_T denotes the total time taken for the distribution and merging

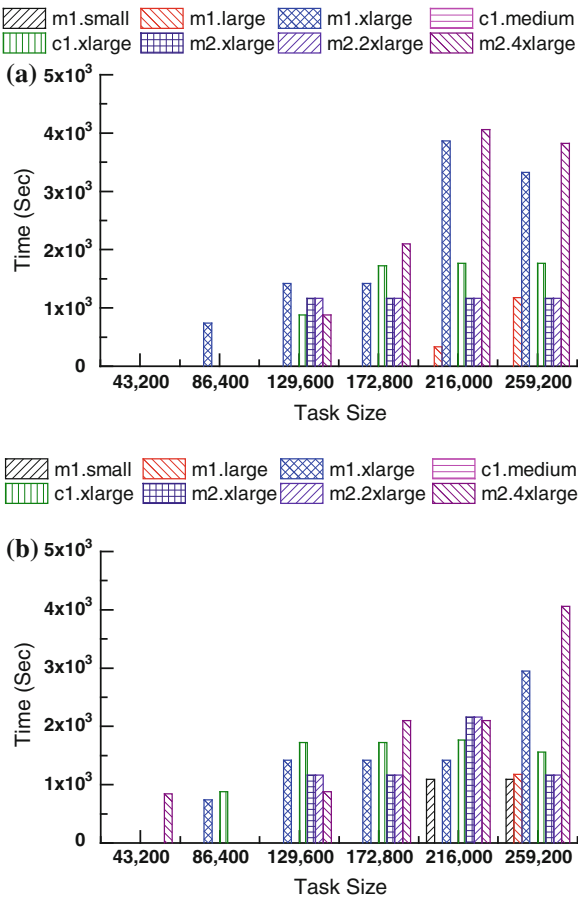


Fig. 59.8 Comparison of rollback time

of tasks. Total_C denotes the sum of costs of task execution in each instance. Figure 59.6a and b show that the total execution time is reduced by an average of 18.8 % after applying ours as compared to without our scheme. Figure 59.7a and b show that the failure time after applying our proposed scheme was increased on 6.68 % as compared to before applying ours. However, in Fig. 59.8a and b, the rollback time after applying our proposed scheme showed an average performance improvement of 4.3 % when compared to the rollback time without applying ours. The rollback time is calculated from a failure point to the last checkpoint time. Figure 59.9a, b show that the total costs after applying our scheme decreased by an average of $0.37 when compared to the cost before applying ours. There are two facts deduced from these results. One is increase of failure time. The other is the improvement of total execution time through an efficient task distribution. And, the task execution loss was reduced when the out-of-bid situation is occurred.

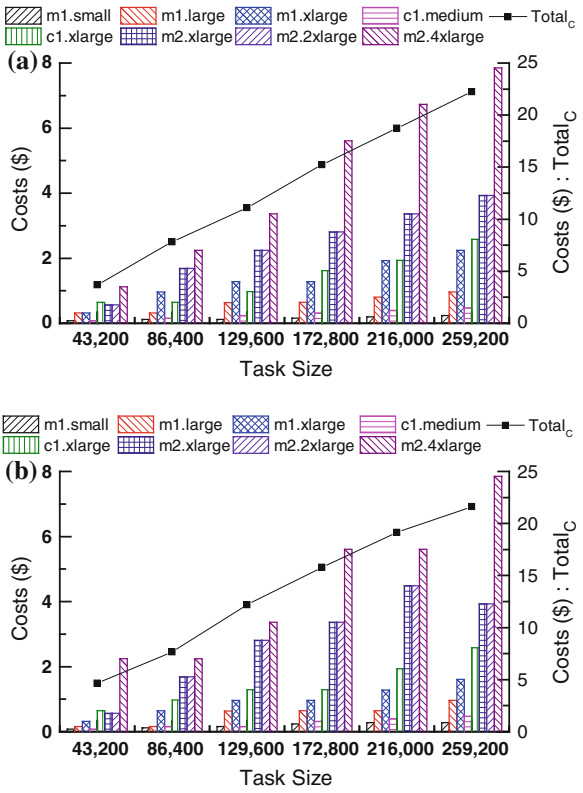


Fig. 59.9 Comparison of costs

59.5 Conclusion

In this paper, we proposed a workflow scheduling technique to consider task processing rate in unreliable cloud computing environments. The workflow scheduling scheme recalculates the task size based on task processing rate in the recalculation point. In additional, the checkpoint scheme that we had proposed earlier performed checkpointing based on two kinds of thresholds: price and time. The simulation results showed that the average execution time in our scheme was improved by 17.8 % after applying our proposed scheme as compared to before applying our proposed scheme. And, our proposed scheme was represented approximately the same cost as compared to before applying our scheme.

References

1. Foster I, Yong Z, Raicu I, Lu S (2008) Cloud computing and grid computing 360-degree compared. In: Proceedings of 2008 grid computing environments workshop, pp 1–10
2. Amazon EC2 spot Instances. http://aws.amazon.com/ec2/spot-instances

3. Yi S, Kondo D, Andrzejak A (2010) Reducing costs of spot instances via checkpointing in the amazon elastic compute cloud. In: Proceedings of the 2010 IEEE 3rd international conference on cloud computing, pp 236–243
4. Yi S, Heo J, Cho Y, Hong J (2007) Taking point decision mechanism for page-level incremental checkpointing based on cost analysis of process execution time. J Inf Sci Eng 23(5):1325–1337
5. Jung D, Chin S, Chung K, Yu H, Gil J (2011) An efficient checkpointing scheme using price history of spot instances in cloud computing environment. In: Proceedings of the 8th IFIP international conference on network and parallel computing, NPC'11, pp 185–200
6. Cloud exchange. http://cloudexchange.org
7. Fernandez H, Obrovac M, Tedeschi C (2012) Decentralised multiple workflow scheduling via a chemically-coordinated shared space. INRIA research report, RR-7925, pp 1–14
8. Liu K, Chen J, Yang Y, Jin H (2008) A throughput maximization strategy for scheduling transaction-intensive workflows on SwinDeW-G. Concurrency Comput: Pract Experience 20:1807–1820

Chapter 60
Enhanced Beacon Scheduling of IEEE802.15.4e DSME

Sung-wook Nam and Kwang-il Hwang

Abstract This paper investigates the validity of the beacon scheduling algorithm specified in the IEEE802.15.4e draft, more specifically Deterministic Synchronized Multichannel Extension (DSME) capability, and presents a concrete guideline for its practical implementation. Based on experimental results, we revise the DSME beacon scheduling by resolving several problems step by step, and finally propose an enhanced beacon scheduling method. In addition, we also prove the performance superiority of the enhanced beacon scheduling, through experiments under various environments.

Keywords IEEE802.15.4 · Beacon scheduling · Superframe · IEEE802.15.4e

60.1 Introduction

The IEEE802.15.4 [1] has become a representative standard for low rate wireless personal communications (LR-WPAN). One of the main advantages in using the IEEE802.15.4 is low power operation, which can be accomplished by beacon-enabled mode. However, the beacon-enabled mode requires that all the devices in the network be synchronized with pan coordinator (PC) and thus PC manages a superframe to maintain an active duration and inactive duration within a beacon interval. This requirement also makes it difficult for a WPAN to be extended to multi-hop networks. One of the critical problems is a collision among beacon frames transmitted by different devices.

Therefore, the beacon scheduling to address beacon collision problem has been considered to be one of the significant challenges in multi-hop networks consisting

S. Nam · K. Hwang (✉)
Department of Embedded Systems Engineering, Incheon National University,
Incheon 402-772, Korea
e-mail: hkwangil@incheon.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_60, © Springer Science+Business Media Dordrecht 2014

of the IEEE802.15.4 systems. In order to address the beacon collision problem, various beacon scheduling methods have been studied so far. The beacon scheduling can be largely categorized into tree-based approach and mesh-based approach. Tree-based approaches [2–8] introduced various beacon scheduling approaches based on a tree topology. The tree network has advantage of low routing overhead, but it is prone to make a network isolation partially in the case of link failure of an intermediate node. On the other hand, a mesh based network can provide more flexible topology management and thus some researches [9–13] emphasized the importance of beacon scheduling for mesh networks. In particular, a IEEE802.15.4 task group (TG4e) realized the need of the beacon scheduling for mesh networks and they involved an efficient beacon scheduling method utilizing a specific bitmap for the neighboring superframe duration slot management in the IEEE802.15.4e draft [12] more specifically Deterministic and Synchronous Multi-channel Extension (DSME) Capability. However, the draft provides only a conceptual idea, so that the concrete algorithm details for its implementation as well as validity of the algorithm through experiments are not yet provided.

In this paper, through experiments, we evaluate the validity of the DSME beacon scheduling algorithm specified in the IEEE802.15.4e draft and presents a concrete guideline for the algorithm implementation. The experiment is conducted using ns-3 [15], which is a popular network simulator based on open source and we find out some problems in the pure DSME beacon scheduling algorithm by analyzing the experimental result. Therefore, based on the results, we revise the pure DSME beacon scheduling algorithm step by step and finally we propose an enhanced DSME beacon scheduling including a limited permission notification coupled with a proper superframe duration slot selection method.

60.2 IEEE802.15.4e DSME Beacon Scheduling

The DSME involves a novel beacon scheduling for mesh networks. Even though the DSME supports multi channel utilization, the multi channel operation is limited only in contention free period (CFP) for the GTS usage. Actual superframe operation is achieved using a single channel, so that the DSME focuses on avoiding beacon collisions among different WPANs by introducing beacon scheduling method in which different pan coordinators in a complex mesh network are synchronized by conducting beacon scheduling based on multi superframe structure which allows a number of superframes to co-exist in a beacon interval(BI). In DSME beacon scheduling, each prospective device, first performs a scan procedure over the available channels. Each DSME device has superframe duration (SD) index table to manage SD information of neighboring nodes. In addition, the SD index information of a node is represented as a bitmap, included in a macSDBitmap field of the beacon frame, and transmitted periodically to notify current SD index allocation information of its neighbors. If a prospective node receives a beacon of an active node that SD index is already allocated, the node

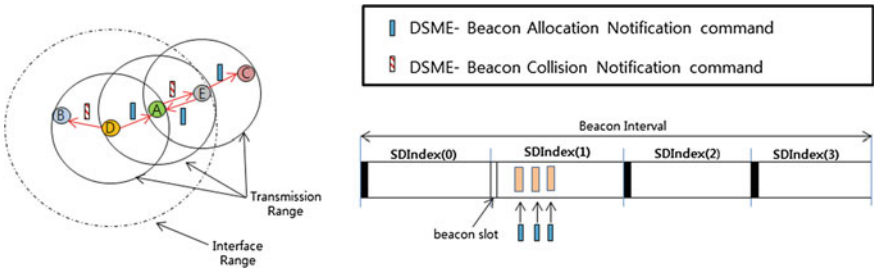


Fig. 60.1 DMSE beacon slot collision avoidance

selects a vacant slot which is represented as '0' in the received macSDBitpmap, sets the corresponding bit to '1', and broadcasts DSME-Beacon allocation notification command frame to its neighbors. The neighboring nodes who receive the notification command, first check if the bit is being used by other neighboring nodes, and then update its SD index table if the slot is available.

However, in the SD index allocation process as mentioned above, a collision situation might occur when more than two devices make an attempt to occupy the same slot. Figure 60.1 illustrates this beacon collision situation. When node D and E, which are neighbors of node A but cannot communicate with each other, receive a beacon of node A, both can select the same slot out of vacant slots in the SDbitmap (Hidden node problem). Therefore, the two nodes have the same SD index so that beacon transmission slot is overlapped. That is, since beacons of the two nodes come to collide, node A cannot hear any beacon transmitted from the two nodes. To address this problem, the DSME uses an additional frame, DSME-Beacon collision notification command. If the two nodes want to use the same SD index by sending allocation notification message, node A allows for the first arrival node to allocate the SD slot, and after that if another node requests to use the already occupied slot, the node A makes the new requester to select other slot by sending collision notification command. Eventually, this procedure can avoid overlapped allocation of SD indexes among neighboring nodes. This method provides a simple but powerful beacon scheduling which is not solved in IEEE802.15.4. In particular, it is possible that superframe duration of two hop neighboring nodes as well as directly neighboring nodes is scheduled in a distributed manner.

60.3 Performance Evaluations of IEEE802.15.4e DSME

The IEEE802.15.4e DSME beacon scheduling as described in the previous Section can provide an efficient scheduling among neighboring WPANs. However, the standard introduces just an abstract concept without any concrete outline and implementation details. Therefore, in this Section, we first present a guideline for

implementation of the DSME beacon scheduling, and then evaluates the performance and validity of the algorithm.

60.3.1 Experimental Environments

We implemented the beacon scheduling of IEEE802.15.4 DSME based on lr-wpan module of ns-3 [15]. For a channel model, single model spectrum channel is used and communication channel is constructed based on the log distance propagation loss error model. Specific primitives and frames for DSME are involved in lr-wpan mac layer and additional PIB attributes and constants are defined in MAC header. In addition, instead of a superframe structure used in original IEEE802.15.4, a multi superframe structure, which is a specific feature of the DSME, is used.

60.3.2 The 1st DSME Experiment

60.3.2.1 SD Index Selection Criteria

If a prospective node receives a beacon, superframe slot of the node can be allocated by selecting a vacant slot. The allocated slots are represented as '1' in the SDBitmap so that the node should choose a slot out of '0' bits. As a matter of fact, since the allocation method is distributive and SDBitmap only represents its neighboring SD index of the sender, slot allocation order in the received SDBitmap is varied according to network topologies. That is, when a node receives a SDBitmap, the SDBitmap might present sparse allocation distribution. Intuitively, we can consider three different selection methods: Least Available Bit (LAB), Most Available Bit(MAB), and random. However, we also presume that the beacon scheduling performance might be varied according to a selection method used.

60.3.2.2 Evaluation of the 1st Experiment

In this subsection through experiments we evaluate the validity and performance of DSME with respect to different SD index slot selection methods. For the experiments, we considered two representative topology models, sparse and dense, consisting of 4×4 nodes as shown in Fig. 60.2. The first experiment is conducted by pure DSME beacon scheduling specified in the IEEE802.15.4e DSME draft.

To evaluate the performance of DSME beacon scheduling, we first observed successful SD slot allocation ratio with respect to individual selection method, LAB, MAB, and random, in sparse and dense topology models. As shown in Fig. 60.3, the result shows that the MAB selection method is superior to LAB and random. In particular, the LAB selection method shows the worst performance.

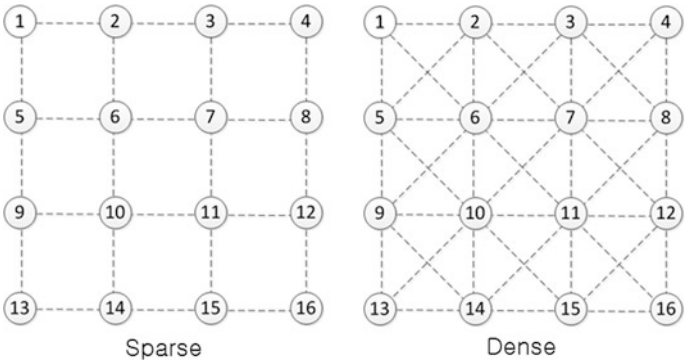


Fig. 60.2 IEEE802.14.4e DSME topology

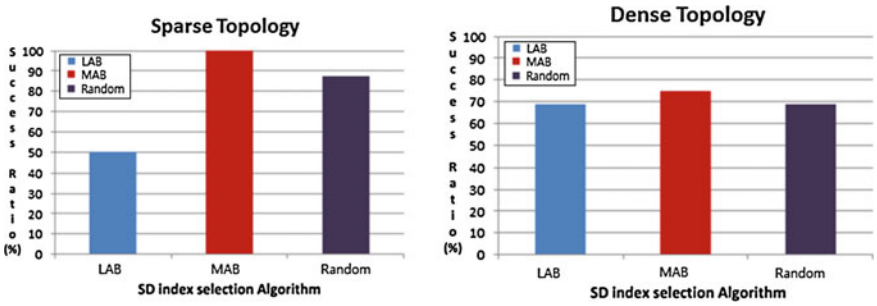


Fig. 60.3 Experimental result of the 1st experiment

That is because the LAB caused a number of collisions among the nodes which chose the same SD slot, as we expected. However, the result also reveals that even the MAB, which outperforms other SD index selection method, shows allocation failure ratio more than 20 % in the dense topology. This results from beacon collision problem that is not yet completely resolved.

60.3.3 Enhanced DMSE Design (1)

The experimental results revealed that beacon scheduling algorithm of the DSME has some critical problems in a realistic WPAN network construction. Therefore, it is necessary to redeem the beacon scheduling algorithm of the pure DSME by resolving collisions. One of the significant problems in pure DSME is collisions among command frames, such as allocation notification, and collision notification, and another one is a beacon collision which is caused by using the overlapped SD slot. This is because a node allocates its SD index slot for itself based on the received SDBitmap information.

So, we first employ a distributed permission notification, which enhances a collision notification proposed in the pure DSME. While the pure DSME sends a collision notification only when the newly allocated SD slot is overlapped with my other neighbor' one, the distributed permission method allows only the node, which receives permission notification message after sending allocation notification message, to complete SD index allocation. A prospective node, which sent allocation notification message, waits until a permission notification is received, and the neighboring active node, which receives the allocation notification of the prospective node, checks whether the requested SD index is available or not, and then if the slot is available it broadcasts permission notification. The neighbors of the active node, which broadcasts permission notification, also come to know the information of a SD index of a new node apart from two hops by hearing this permission notification. This might result in the reduction of collisions. Furthermore, it is expected to cope well with complex topology model.

This enhancement has been expected to significantly improve the performance of the pure DSME. However, on the contrary to our expectation, the experimental results were worse than the pure DSME. As shown in Fig. 60.4, the successful beacon scheduling ratio shows lower value in both of sparse and dense models.

As a result of analyze problem, we found out a problem about the 1st revision. The distributed permission method can avoid the collisions by learning SD index information of two hop distance nodes through a permission notification, but all the nodes, which receive an allocation notification message, have the right to send the permission notification and thus the nodes, which do not have to be allocated, also complete SD allocation. Therefore, they allocate the same SD slot, and eventually beacon collision occurs. The collisions also occur without regard to the SD index selection method used.

60.3.4 Enhanced DSME: The 2nd Revision

The first revision of the enhanced DSME is to limit the permission right only into the originator of the latest beacon. Restricting the node that can send the permission into the sender of the beacon can avoid the abuse of the SD allocation caused by unnecessary permission notifications performed by neighboring active nodes. The second approach of the enhanced DSME employees a dedicated period, a permission only period (POP), which is located at the end of superframe and is used to allow only permission notification message transmission of the originator of the beacon. Transmissions of all the command frames except for the permission notification message are restricted into its superframe duration to protect the permission notification message from colliding with other command messages during POP.

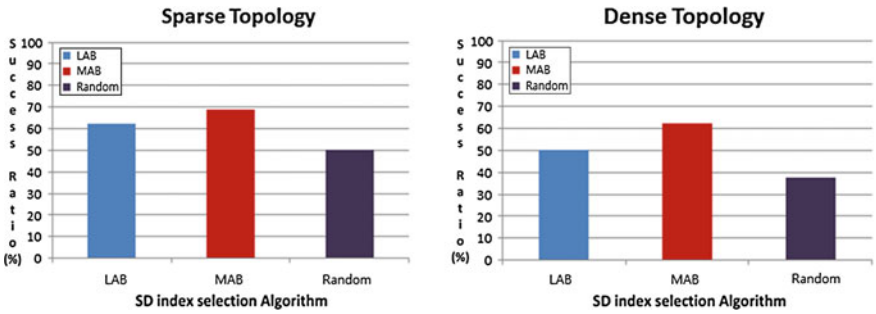


Fig. 60.4 Experimental result of the 1st revision

60.3.5 Validity of the Revised Algorithm

The enhanced DSME has been conducted on the same environment as previous experiments. As shown in Fig. 60.5, the enhanced DSME shows 100 % allocation success ratio when the MAB SD index selection method is applied, in both of sparse and dense model. Compared to the previous method, pure DSME and the distributed permission notification, this performance improvement of the enhanced DSME beacon scheduling might result from utilizing a limited permission notification and employing permission only period. Furthermore, the result demonstrates that the MAB SD index selection algorithm is the most suitable for the DSME beacon scheduling compared to other SD index selection methods, LAB and Random.

60.3.5.1 Beacon Scheduling Completion Time

We also observed the association completion time of the enhanced DSME beacon scheduling. The association completion time represents the total time required to complete SD index allocation of all the nodes in the network. We observed the association completion time with respect to varying the number of devices at different superframe order (SO), 5, 6 and 7. As shown in Fig. 60.6, the result shows that completion time is getting longer as the BO size becomes smaller. That is because smaller SO can make more number of SD slots within the same beacon duration. Since the SO size is also proportional to the duration of each superframe, short superframe duration might deal with only a small number of allocation requests, and eventually the remaining nodes which are not allocated at the round have to wait until the next neighbor's beacon is received.

The number of multi superframe slots is equal to the beacon bitmap length and the length is represented by SDBitmaplength $= 2$ (BO $-$ SO). So, the enhanced DSME beacon scheduling can apply the result after calculating the expected number of slots in advance of a network formation according to the network size.

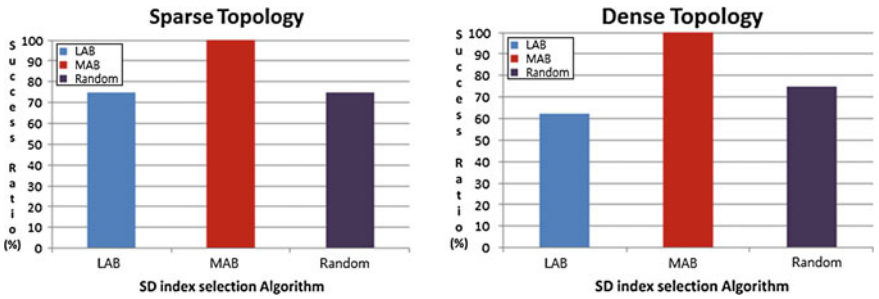


Fig. 60.5 Experimental result of the enhanced DSME

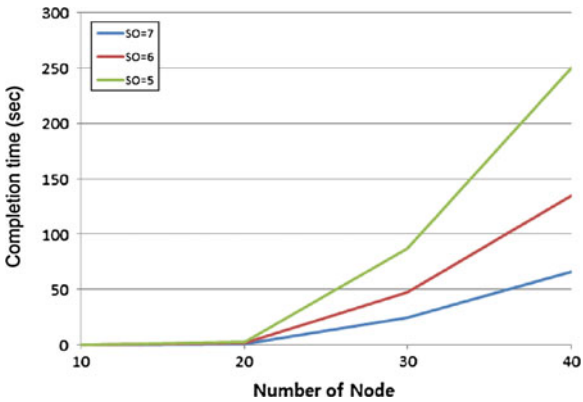


Fig. 60.6 Association Completion Time

Moreover, even though the network size is extended during run time, it is possible to dynamically change the number of expected slots prior to network formation, since each beacon contains BO and SO information.

60.4 Conclusion

In this paper, we conducted in-depth analysis of the beacon scheduling algorithm specified in the IEEE802.15.4e DSME capability, addressed several problems presented in the draft through experiments, and revised the beacon scheduling algorithm step by step. The experimental results show that the enhanced beacon scheduling method is capable of minimizing beacon collisions during beacon scheduling process and guarantees a reasonable beacon scheduling completion time. Therefore, it is expected that the enhanced DSME beacon scheduling method will contribute to various network design based on multi-superframes.

Acknowledgments This paper was supported by the Incheon National University research grant in 2013.

References

1. IEEE Std 802.15.4 (2003) Part 15.4: wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (WPANs). In: IEEE standard for information technology, 2006D. Lenton, "The small screen [TV to Mobile Devices]," IEE Rev., vol 49, no. 10, pp 38–41, Oct 2003
2. Lee M, Zheng J, Liu Y et al (2004) Combined beacon scheduling proposal to IEEE802.15.4b. IEEE802.15-04-0536-00-004b, Sept 2004
3. Koubaa A, Cunha A, Mario A et al (2007) A time division beacon scheduling mechanism for IEEE802.15.4/Zigbee cluster-tree wireless sensor networks. In: 19th Euromicro conference on real-time systems (ECRTS' 07)
4. Cho JJ, An SS (2009) An adaptive beacon scheduling mechanism using power control in cluster-tree WPANs. Wireless Personal Commun. 50(2):143–160
5. Ahn SY, Cho JJ, An SS (2008) Slotted beacon shceduling using ZigBee Cskip Mechanism. In: Proceedings of the second international conference on sensor technologies and applications, 25–31 Aug 2008
6. Yeh LW, Pan MS, Tseng YC (2008) Two-way beacon shceduling in ZigBee tree-based wireless sensor networks. In: Proceedings of 2008 IEEE international conference on sensor networks, ubiquitous, and trustworthy computing
7. Yen LH, Law YW, Palaniswami M (2008) Risk-aware beacon scheduling for tree-based ZigBee/IEEE802.15.4 wireless networks. In: Proceedings of the 4th annual international conference on wireless internet, WICON '08
8. Chen S, Almeida L, Wang Z (2011) Analysis and experiments for dual-rate beacon scheduling in ZigBee/IEEE802.15.4. INFOCOM Workshop, Apr 2011
9. Burda R, Wietfeld C (2007) A distributed and autonomous beacon scheduling algorithm for IEEE802.15.4/ZigBee Networks, MASS 2007. In: IEEE international conference on mobile ad hoc and sensor systems, 2007, 8–11, pp 1–6, Oct 2007
10. Lee M, Zhang L, Zheng J et al (2010) IEEE802.15.5 WPAN mesh standard-low rate part: meshing the wireless sensor networks. IEEE J Sel Areas Commun 28(7):973–983
11. Muthukumaran P, de Paz R, Spinar R et al (2009) MeshMAC: enabling mesh networking over IEEE802.15.4 through distributed beacon scheduling. In: The first international conference on ad hoc networks (AdHocNets09), Niagara Falls, Ontario, Canada. Sep 2009
12. Lee WY, Hwang KI, Jeon YA et al (2011) Distributed fast beacon scheduling for mesh networks. IEEE MASS 2011, Mesh Tech, Oct 2011
13. IEEE P802.15.4e Draft (2011) Amendment 5: amendment to the MAC sub-layer for low-rate wireless personal area networks (WPANs) Amendment 5. In: Proceedings of IEEE standard for information technology
14. ZigBee Alliance (2006) ZigBee Specification v1.0, 12
15. ns-3, http://www.nsnam.org

Chapter 61
Building Recognition Guided by Wi-Fi Based Localization

Wonsun Bong and Yong Cheol Kim

Abstract We propose a method of building recognition on imagery from a mobile device, when an approximate location of the user is available. The query image taken by a mobile device is matched with images of nearby buildings in database. We use SIFT descriptors for feature matching to get an invariance of scale and rotation between images taken at various shooting conditions. A rough localization of the mobile user is obtained by triangulation or fingerprinting of the received Wi-Fi signal strength from Wi-Fi access points. The proposed method is demonstrated on buildings with access points in vicinity. In the experiments, simulated street view images are generated to serve the role of on-line street view data. Results from our work on Wi-Fi based localization and radio map interpolation are also presented

61.1 Introduction

Given an image of one building, it is a difficult task to tell which building is in the image. There would be several buildings which look quite alike in a big city. Hence recognition by searching for the most similar-looking building is a difficult high-level computer vision problem. However, if the search space can be reduced to just tens of buildings, the recognition of the building becomes a tractable problem.

In this paper, we propose a method of building recognition on imagery from a mobile device when the approximate location of the building is available. The motivation is to retrieve specific information about that building and the surrounding area. This information will be useful for location based service for mobile users [1].

W. Bong · Y. C. Kim (✉)
Department of Electrical and Computer Engineering, University of Seoul, Seoul, Korea
e-mail: yckim@uos.ac.kr

W. Bong
e-mail: gaam@uos.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_61, © Springer Science+Business Media Dordrecht 2014

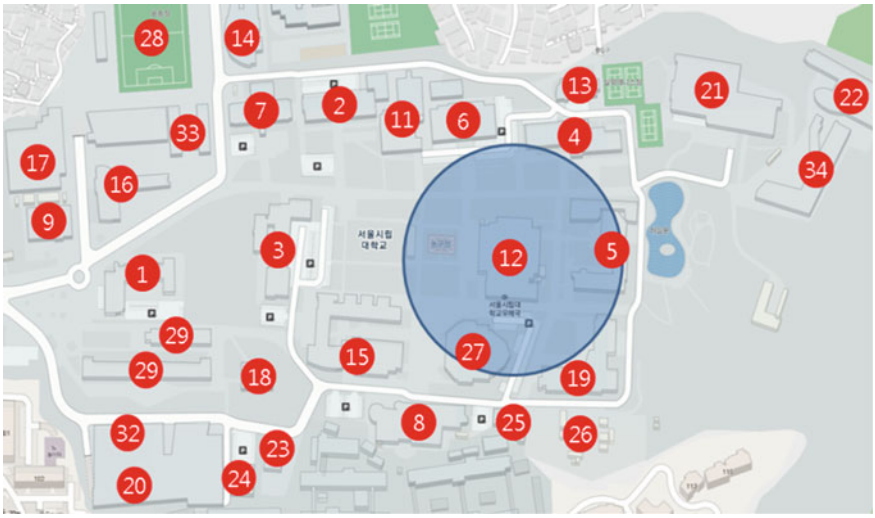


Fig. 61.1 *Circle* represents the range of search space

In Fig. 61.1, the building locations used in the SIFT matching are shown. The circle represents the range of search space obtained by Wi-Fi localization. A thumbnail for the building images is in Fig. 61.2.

The query image taken by a mobile user is matched with nearby buildings, whose images are from on-line street-view data. We use SIFT descriptors for feature matching to get an invariance of scale and rotation between images taken at different shooting conditions [2].

A rough localization of the mobile user is obtained by triangulation or fingerprinting of the received Wi-Fi signal strength. Results from our work on Wi-Fi based localization and radio map interpolation are presented. The proposed method is demonstrated on buildings with Wi-Fi access points (AP) in vicinity.

The organization of this paper is as follows: In Sect. 61.2, Wi-Fi based localization technology is described. Results on tracking of moving devices are presented in Sect. 61.3. In Sect. 61.4, a technique on radio map interpolation is presented. In Sect. 61.5, we present results on SIFT-based building recognition. Finally, we draw a conclusion in Sect. 61.6.

61.2 Wi-Fi Based Localization

For a mobile user with a smart phone, an easy way of localization is either using the built-in GPS or Wi-Fi function. Wi-Fi based localization is an alternative technology for indoor localization. Smart phones with Wi-Fi access functions can be localized, with reference to nearby AP [3].

Fig. 61.2 Images used in SIFT matching test

Received signal strength (RSS) gets weaker as a mobile device (MD) moves away from an AP. RSS is used to estimate the distance from the signal source. In indoor propagation, RSS is modeled to have a linear decrease in (dBm) with reference to the logarithmic distance [4]. $P(r)$ and $P(r_0)$ in (61.1) are the signal power at distance r from the radio source and at reference distance r_0. W_A and w are the wall attenuation factor and the number of walls, respectively.

$$P(r) = P(r_0) - 10 \cdot n \cdot \log\left(\frac{r}{r_0}\right) - w \cdot W_A \tag{61.1}$$

Real radio signal propagation in a building is perturbed due to the reflections and scattering by the obstacles in the propagation path, which results in a complicated attenuation pattern. Location fingerprinting has been introduced to overcome this instability [3]. Radio map is a database of RSS at many reference points (RP) from nearby access points. The position of a MD is identified by matching its RSS pattern (fingerprint) $[v_1, v_2, \ldots v_n]$ with the RSS vector $[s_1, s_2, \ldots s_n]$ in the radio map.

A RP is placed at a position surrounded by several APs. The RSS pattern of a RP from neighboring APs is monitored. Each wireless link between a RP and an AP is regarded as a distinct channel. Hence, an RP holds K estimates of separate channels for K neighboring APs. These estimates are being continuously updated.

Let proxy(i) represent the RP which is the closest to MD_i. If the channel characteristics between AP_j and proxy(i) has changed for some reason and this caused the RSS on proxy(i) from AP_j is stronger than *normal*, it is reasonable to assume that MD_i will have a similar change of RSS from AP_j. Hence, the path loss parameters for channel ($MD_i : AP_j$) in (61.1) are inherited from (proxy(i):AP_j). This way, the fluctuations of an indoor wireless channel are efficiently compensated.

Usually, Euclidean distance is used for the similarity metric. The distance between the RSS vectors of a MD and each RP is computed. Then RP_i with the shortest distance is selected as the proxy RP. In (61.2), N_{AP} and RSS^j represent the number of APs and the RSS from the *j*th AP, respectively.

$$i = \arg\min_{i} \sum_{j}^{N_{AP}} \left(RSS^{j}_{MD} - RSS^{j}_{RP_i}\right)^2 \tag{61.2}$$

61.3 Tracking of a Moving MD

Point-by-point decision of the proxy RP entirely depends on the similarity of RSS vectors and thus p_e, probability of error, is far from zero (3–10 %). Most of these errors may be removed by hidden Markov model (HMM) tracking if the previous findings for the proxy RP in the trajectory are considered in the decision of the present one.

61.3.1 Simulation Results by MD Tracking

Tests were performed for several moving paths. Figure 61.3 shows one of the paths which traverse several rooms and a hallway. A comparison of the error rate by HMM and Euclidean distance is shown in Fig. 61.4.

The area along the hallway walls in Fig. 61.3 are the most vulnerable since the RSS patterns on these grid points are much similar to each other and thus the p_e is as high as 23 % in point-by-point decisions.

With HMM tracking, p_e decreases to as low as 3 %. Most of the erroneous findings along the hallway are corrected. Point-by-point decision is prone to errors when the RSS vector is more similar to that of non-proxy RPs. On the contrary, HMM tracking, which considers all the previous decisions in the trajectory, is less susceptible to accidental matching of RSS vectors.

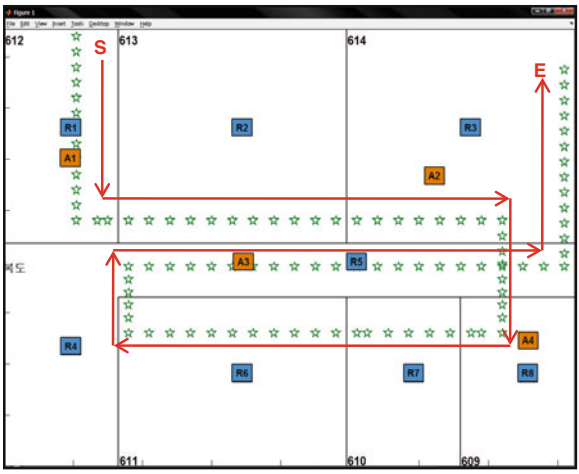


Fig. 61.3 Moving path of MD (*square* AP, *circle* RP)

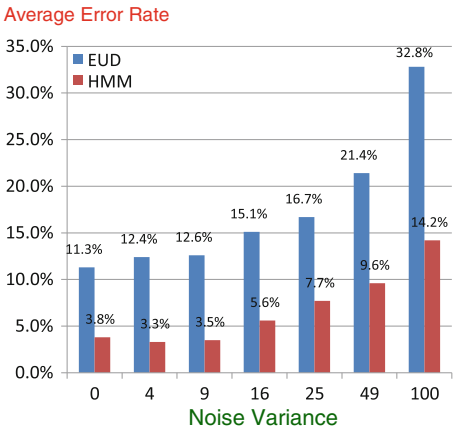


Fig. 61.4 Performance comparison of ED decision and HMM tracking

61.3.2 MD Tracking on Real RSS Measurements

We tested for real environment, which is similar to the structure used for simulation test. Four APs and four RPs are deployed over six rooms and a hallway. RSS vectors are sampled at 101 points along the moving path. An MD moves from one point to another point every 10 s. Tests were performed along four different moving paths, each of which passes through regions susceptible to accidental matching of RSS vectors.

We present the result for path 3 in Fig. 61.5. Star marks represent the trajectory of MD. Circles and triangles stand for errors in point-by-point decision and in HMM tracking, respectively. The arrow points to the erroneously matched RP. Summing the results for four paths, we have six matching errors with Euclidean

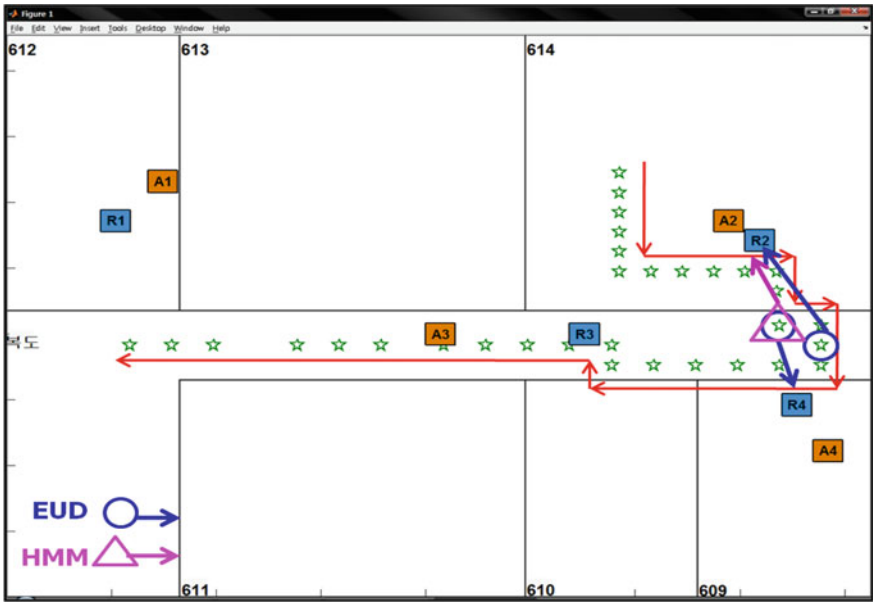


Fig. 61.5 Tracking errors by HMM and by EUD (Path 3)

distance and just a single matching error with HMM matching. In real environment as well as in computer simulation, HMM-based tracking algorithm proved to have much better performance.

61.4 Interpolation of Low Density Radio Map

Interpolation of a coarse radio map into a dense one may reduce the cost of a radio map. An interpolated map, however, has low accuracy near a wall. We present a method of constructing a high-density radio map by discontinuity preserving smoothing (DPS). For the DPS functionality, we used an adaptive smoothing (AS) [5], which preserves the discontinuity by localized smoothing in accordance with the wall layout of indoor building, instead of simply fitting the measured data into a parametric curve.

In adaptive smoothing, which is often used in image processing, the filtering strength of signal is controlled by the magnitude of signal gradients. For image signal, areas with large gradient represent edge area (object boundary). By weakening the filtering strength around edges, blurring of object boundaries in filtering can be avoided. Walls in radio map correspond to edges in image signal.

Figure 61.6 shows the advantage of adaptive smoothing over conventional interpolation. The small circles represent the measured RSS. The stars and diamonds stand for the interpolated values and DPS-reconstructed values,

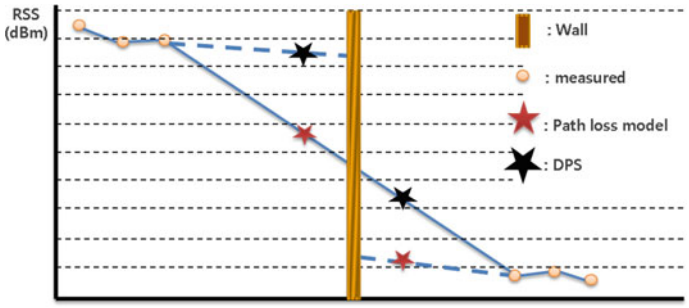


Fig. 61.6 Adaptive smoothing can accommodate the wall barrier

respectively. In the DPS, smoothing is almost separately performed on both sides of the wall. Hence the estimate of RSS is affected mainly by those which lie on the same side of the wall and those beyond the wall have less impact on the estimation.

The performance of DPS-reconstructed radio map is compared with IDW-based interpolation and Voronoi-interpolated map [6]. The location is identified by matching its RSS pattern with radio map and finding the best match. The test was performed in three office rooms and along an elongated hallway. Each RP is a grid point of 1.2 m by 1.2 m. The number of RPs used is 145.

The fluctuation of RSS from a Wi-Fi AP is quite large. In the initial gathering, the RSS at each RP is measured 100 times during 200 s and the average is recorded to reduce the effect of random fluctuation. To prevent a short-term measurement from dominating the channel characteristics, the interval is kept to be 2 s between measurements. RSS vectors of all 145 RPs are randomly selected with a sampling ratio varying from 10 to 90 %. This way, we constructed a series of low density radio map to find the lower bound of sampling ratio for a reasonable performance of localization. Figure 61.7 shows the average error in localization. As expected, the radio map reconstructed by DPS outperforms both IDW-interpolated map and Voronoi-interpolated map.

Besides, we find two observations: First, when the sampling ratio is higher than 20 %, the accuracy in localization is approaching that of a genuine full density radio map. This means that the cost can be reduced to as low as 20 %.

Second, with sampling ratio higher than 60 %, the performance of DPS-reconstructed radio map is *even better* than that of the original full density map. In general, the average error increases as the sampling ratio decreases. In common sense, 90 % map is expected to have a performance equal to, but not higher than, the original full density radio map, even at the best case.

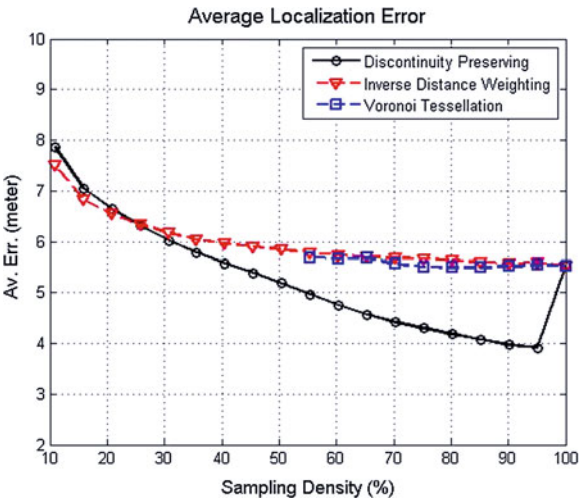


Fig. 61.7 Comparison of average error in localization

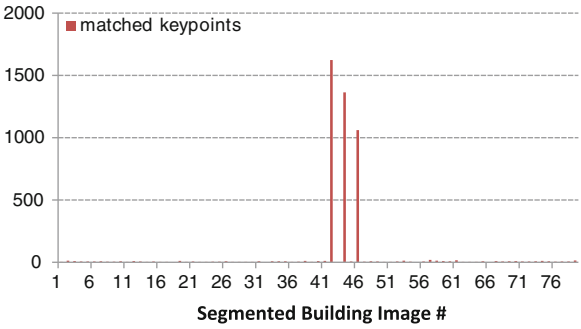


Fig. 61.8 Result of SIFT matching with query image

61.5 SIFT-Based Building Matching

For each of the twenty five buildings, three images are taken from varying angles and then they serve as the street view DB. Figure 61.8 shows the number of matched SIFT key points between the query image and each of the DB images. The result that DB images (numbered as 41, 44 and 47) correspond to the query image coincides with the ground truth.

61.6 Conclusion

We presented a method of building recognition on mobile device images assuming that an approximate location is computed by other means such as GPS or Wi-Fi based localization. The proposed method is demonstrated on buildings with Wi-Fi

access points in a campus area. The query image taken by a mobile user is matched with nearby buildings, whose images are from on-line street-view data. SIFT descriptors provided a robustness of invariance of scale and rotation between the query image and street view DB images.

We presented a HMM-based method of tracking the proxy RP for a moving MD and examined the performance of the proposed HMM tracking algorithm for a moving MD. Simulation results show that the error rate with HMM achieves a considerable decrease, when compared with point-by-point decision by Euclidean distance. We also presented a method of constructing a high-density fingerprinting radio map by DPS of a low-density radio map. We found that the cost of radio map can be reduced to 20 %.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (grant number NRF-2013R1A1A2012745).

References

1. Hutchings R, Mayol-Cuevas W (2005) Building recognition for mobile devices: incorporating positional information with visual features, CSTR-06-017. Computer Science, University of Bristol, Dec 2005
2. Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
3. Bahl P, Padmanabhan V (2000) An in-building RF-based user location and tracking system. In: IEEE Infocom 2000, vol 2, pp 775–784, Mar 2000
4. Seidel S, Rapport T (1991) 914 MHz path loss prediction model for Indoor wireless communications in multi-floored buildings. IEEE Transactions on Antennas and Propagation, pp 207–217, Feb 1992
5. Saint-Marc P, Chen J (1991) Adaptive smoothing: a general tool for early vision. In: IEEE transactions on pattern analysis and machine intelligence, vol 13(6), June 1991
6. Lee M, Han D (2012) Voronoi tessellation based interpolation method for Wi-Fi radio map construction. Commun Lett IEEE 16(3):404–407
7. Li B, Wang Y, Lee H, Dempster A, Rizos C (2005) Method for yielding a database of location fingerprints in WLAN. IEE Proc Commun 152(5):580–586
8. Bong W (2012) Generation of fingerprint radio map by adaptive smoothing. MS. Thesis, University of Seoul, Feb 2012

Chapter 62
Context Aware Based Access Control Model in Cloud Data Center Environment

Yun Sang Byun and Jin Kwak

Abstract Due to the cloud environment to the development of smart devices and the development of IT technology recently, the interest about cloud data center which combined cloud with the existing data center has increased, and which makes users become easier to access to the resource such as database, web, and document with no regard to time and place. However, there're security problems as conflicting results like authority or policy for access of resource, verification of data integrity and data security. In particular, there's a drift about user's access control in cloud that is used by a lot of users, it is not enough to adopt existing access control method to cloud circumstance. Therefore, this paper proposes situation recognition technology based access control model which can control lots of cloud's users and access of device flexibly.

Keywords Cloud data center · Context aware · Context aware based access control · Access control

62.1 Introduction

The outflow of illegal information in the cloud data center has the possibility of exposure not only the user's personal information but also information of customers using the data center, to prevent this problem, various research and development of

Y. S. Byun
ISAA Laboratory, Department of Information Security Engineering, Soonchunhyang University, Asan, South Korea
e-mail: ysbyun@sch.ac.kr

J. Kwak (✉)
Department of Information Security Engineering, Soonchunhyang University, Asan, South Korea
e-mail: jkwak@sch.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_62, © Springer Science+Business Media Dordrecht 2014

security technology have proceeded consistently. Especially, research on the access control technology to prevent access by unauthorized users have been actively, but access control existing methods already is developed in consideration of only the internet environment existing IT environment, so it can't be applied to the cloud data center that many users and large-scale data is used.

Also, it is possible to provide cloud service in anytime and anywhere by development of smart device and supply. For these cases, due to access to the cloud data for users based on a variety of devices, the access control method considering the status of each user is required. Thus, in this paper proposes access control model based on situation recognition which can control access of data by playing a role properly after awareness of a lot of users' situations. This paper consists of 6 sections. The Sect. 62.2, we analyze related work that analyzes cloud data center and existing access control models, furthermore, studies information of context. At the Sect. 62.3, the problem of access control solutions proposed is analyzed, and the Sect. 62.4, makes a suggestion access control model based on situation recognition. Section 62.5, studies access control model proposed, and At last, it is made a conclusion at the Sect. 62.6.

62.2 Related Work

62.2.1 Cloud Data Center

Cloud Data Center is created by applying cloud computing method to existing IDC (Internet Data Center) to prevent IT resources such as storages from Internet technology and outside threat. That means data center based on offering services in cloud method. Thus, cloud data center makes saved data and IT intra not be in corporate, and saves data in the outside data center. It can solve the problems of cost to provide the expensive equipment that is different from the IDC existing cloud data centers, it is possible to provide services faster and more efficient.

Recently, in work related cloud data center, various studies that takes into account the environmental issues, problem of low power consumption, protection of data from threats that cannot be perceived, such as natural disasters of various security and hacking, and improving the efficiency of data center performance and research on such protection and management, are underway. Further, studies to reduce the power consumption to be consumed in the cooling system of various devices associated with the construction of data centers has been actively. The cloud data center, there is an advantage that may be in the unconstrained time · space, accesses to the computing environment, and takes advantage of the necessary data, however, variety of security threats to data stored in the data center is existed, it has been researched to solve this problem steadily [1, 2].

62.2.2 Existing Access Control Model

Because an existing access control model works mostly from the object of access control's point, it's not suitable to cloud circumstance changed frequently. Access controls recently invented are Discretionary Access Control, Mandatory Access Control (MAC), Role Based Access Control (RBAC), Activity Based Access Control (ABAC), Task-Role Based Access Control (T-RBAC).

62.2.2.1 Role Based Access Control

Role based access control (RBAC) is the method that is used for reducing users and resources management, and it has different feature compared with the existing way, which makes users judge the grant of authority by role in the organization, not identifier or justified rules. Also, it can work flexibly the particular security policy and easy authority management due to possibility to use lower layer [3, 4].

62.2.2.2 Activity Based Access Control

Activity Based Access Control (ABAC) is the access control mechanism that has been studied for use in smooth collaborative environment is expressed in the workflow, and is defined as the gathering of activity combined to work for common aim. The existing access control model; DAC, MAC, and RBAC are performed for permission from the time of granting access, so it is possible to use the access permission from any point. However, in case of access control model of action-based, the allocation of access rights of the user is separated from the activation of permissions, the use of grant is limited according to the progress of the workflow even when it is authorized [5, 6].

62.2.2.3 Task-Role Based Access Control

Task-Role Based Access Control (T-RBAC) is united existing one and Activity Based Access Control. The biggest difference of Task-Role Based Access Control and existing one is a way to grant access to a user, if the traditional access control model was granted directly access rights by role of each, but task-role based access is granted through the challenges that can access to perform its role [7].

62.2.3 Context Information

Context Information can be defined by changes in the environment, including people and things which can be distinguished users or user identity characteristic. In

addition, it also can be defined by emotional state, place, or date and time in the environment which the user belongs as the definition of granularity [8].

Context Toolkit of Georgia Tech is a status information collection tool that has been developed for the easy production of applications related to situational awareness. Context Toolkit provides information to the application by collecting status information, so that the application can use in accordance with the respective characteristics. In addition, components of Context Toolkit communicate with based on XML over HTTP to make sure for the application aware of status information. Context Toolkit is consisted of Widget, Aggregator, Interpreter, each part is as follow:

- Widget—It is responsible for relaying between the application and the operating environment, and is controlled to be able to access context information. Furthermore, it can hide the content of the information collection mechanism of the situation to abstract the status information appropriately by the needs of the application.
- Aggregator—It owns all the functions of the Widget, at the same time, puts together the state information of the actual object, performs the role of a gateway between the Widget and basic applications.
- Interpreter—It abstracts or re-interpretates at a high level status information from low-level, that is, It binds different contextual information or translates information to unify status information represented by different way [9].

62.3 Analysis of the Problem

As research on the access control has been studied actively, various access control techniques have been developed, however, there are difficulties to apply to the cloud data center environment the virtualization technology and IT technology various mechanisms of access control that takes into account only the environment of the PC-based. To make this solution, it has studied about various access control model in similar environment, however, there's a problem to control the access by varied device users.

62.3.1 MAC

Mechanism of the mandatory access control, as a method of controlling access to objects belong based on security rating is assigned to the object principal and principal use. Mandatory access control scheme is subject and the subject is given to the use of the objects based on the security level for the object belonging to a technique to control access. Subject is increased, the amount of the object using a large amount of the objects given a security rating, it is often difficult, because the access was somewhat difficult to apply the cloud is present.

62.3.2 DAC

Discretionary access control scheme [10] is subject and another subject is a member of the group based on the authorization for access to the object that you want to use a method of control. This is shared by the objects that are often used in environments where. But it does not comply the rule of integrity and the principle of least that Users should be granted only the minimum necessary. The provision of related functions is difficult to secure. So, essentially required integrity is not suitable for a cloud environment.

62.3.3 CA-TRBAC

CA-TRBAC [8] Actively access control model is constantly changing to cope with the situation, depending on the configuration of the policies have been developed to enable. However, this is not considered a situation that can occur. Because the state information is often changed to instantly apply the policy difficult.

In particular, many users can use without restrictions of time and space in the cloud environment, the immediate application of the policy difficult.

62.3.4 RBAC

RBAC Access Control Model [3, 4] is the number of users that can arise due to the management of resources in order to reduce the development and use of the access control method on the privileges granted to the user it is possible to inherit. So, easy rights management and security policy to implement flexibility may provide advantages. However, the dynamic activation and application-level access to the oath required by the constraints of the workflow because it does not take into account difficult to use in a cloud environment.

62.3.5 ABAC

Access Control Model [5, 6] of the existing access control model with a different access privileges and rights assignment activation has the advantage that the separation. Restrictions according to the application level according to the specifications for providing support for the implementation of integrity rules. However, the workflow does not belong to the large amount of work, such as cloud cover, because the workload is difficult to apply in many cases.

62.3.6 T-RBAC

T-RBAC [7] is a feature which granted access to the direct role does not have access to that role is given to the task of performing. In addition, it can resolve the limitations of action-based access control model and role-based access control model. However, over time due to constraints restricting the use of resources to fulfill the principle of least privilege does not provide a way to be difficult to apply to the cloud.

62.4 The Proposed Method

62.4.1 The Proposed Model

In this paper, Status information processing system, the policy management system, authentication system, is composed of rights management system, the database mainly, each system, access control model based on situational awareness, such as those proposed in this paper, as a sub-module is reconstructed. For more information for each system, it can be summarized as follows Table 62.1.

Access control model proposed in this paper is organized as Fig. 62.1, each system is configured as follows.

62.4.2 Course of Action

1. The user data is used for user information (ID/PW, PKI, device information, and the like) transmitted together with the context information request permission to access.
2. Access Control module distinguishes user's connection environment, device, and network state based on the information of user access request.
3. After receiving the admission request, the server will transmit the data stored in the received information and comparing the existing user and device authentication. (Denial of certification and permit access decisions based on the results).
4. Extract information from the user information, than the situation information categorized by the information stored in the collector. Classified data is transmitted to a preprocessor in accordance with the environment information and time information from each module performs a pre-processing to be recognized. The collected context information is also sent at the same time to the destination inference and behavior inference. So, it can be deduces User location information and user behavior. Then, the context information and transmits the policy management system.

Table 62.1 Component of the proposed method

Component	Meaning
User	The number of users that can request access to use the existing resources in the CDC
Authentication system	System that performs authentication of the user who requests access to the CDC
Policy management system	System on the basis of the status information sent by the user, to grant the appropriate policy
Context process system	System on the basis of the status information sent by the user, to grant the appropriate policy
Role management system	System in a system to manage permissions, to grant the appropriate authority in conjunction with other systems
Database	Database, which holds all the information access control model is required

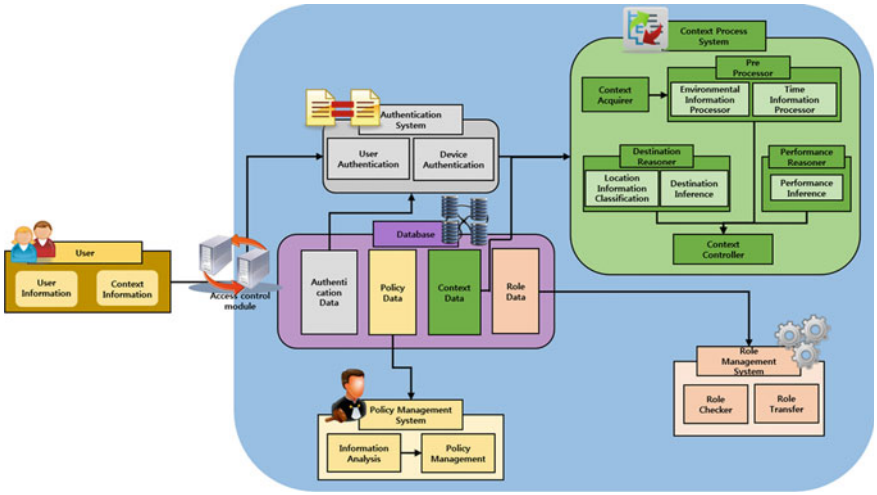


Fig. 62.1 Proposed access control model

5. Policy management system analyzes the received context information. Analysis results are sent to the policy management module according to the situation of the user to determine and select the policy. The selected policy is transmitted to Context Controller.
6. Context Controller on policy and status information is transmitted to the controller synthesis. Through the receiving policies data from the policy management system, it determined the suitability of the conditions and policies which is for user. Then, depending on the determination result suitability, the data transfer to Role Management System or to notify the result to determine whether it will access.
7. Role Management System is based on data received from the Controller to grant users the right to confirm that. In addition, the rule conversion module performs

the appropriate permissions according to the user's situation. (In this case, the authority to change again by Role Checker Review).

8. Step by step through each system to perform the determination and the review process, and finally, as appropriate for the user is authorized, access steps to determine whether or not is determined.

62.5 Comparative Analysis

Before the proposed model, comparing and analyzing the access control mechanism of the existing analyzed the problem can be summarized as follows: Table 62.2. The cloud data center environment, the access control model proposed in this paper, an access control model to determine if access is to recognize the situation in response to changes in the user's environment.

- Applicability: applicability was analyzed the considerations processing capacity corresponding to the amount of objects and the surrounding environment. MAC gives to security rating to the object. So, the object increases make the problem of management ability. And it does not consider any user's environment. So, it is hard to apply cloud environment. However, the proposed model, though the user's environment, location, and action deduction, it gives the stored rights. So, regardless of the amount of the object, it is possible authorize. Therefore, it is possible to apply the cloud environment.
- The application of the policy: it was possible to access control mechanisms of all the analysis. However, pre-registration policies can only be applied to the similar environment. Thus, the feedback is made or due to rapid change in the environment is difficult to apply the policy. It is difficult to apply the cloud. In the case of access control model that was proposed in this paper, by utilizing the resources of the cloud, they are saving by setting the DB in the management system. According to the status of the user to navigate through the application of appropriate policies, it is possible to grant.
- Immediate response: Immediate response is needed. However, the access control technique is an immediate response is not possible. Access control techniques are pre-registered method to the policy to be applied. The proposed access control model is communication with the user's real time. So environmental information, behavior information, location information, and receiving and often detects a change of circumstances. Then, it gives the appropriate authorization. It is possible to use the cloud environment.
- Rights flexible assignment: Access control scheme based on defined rules and policies. So this can give the authority to user or take back the authority from user. Many users, however, often move in a cloud environment, and to advance communication, the fixed authority allocation may be cause security problems.

Table 62.2 Comparison with existing access control models

구분	MAC	DAC	CA TRBAC	RBAC	ABAC	T-RBAC	제안 기법
Applicability	X	△	△	△	△	△	O
Application of the policy	△	△	△	△	△	△	O
Immediate response	X	X	O	X	X	△	O
Rights flexible assignment	X	X	△	X	△	△	O

X: Impossibility; △: But insufficient; O: Available

According to the proceeds to communicate from time to time depending on changes, this model determines the status information collected from the policy. And then, it performs the assignment of permissions. In result, the proposed access control model is flexible can be assigned privileges.

62.6 Conclusion

Cloud data centers stores a large amount of data, because an unauthorized user to access the data occurred, so it is possible to occur accidents such as use the data illegally frequently. Thus, secure management of data stored in the data center is necessary and access control's need has been increased.

Therefore, in this paper, an access control model is proposed to take into account the status of the user that changes fluidly in the cloud data center to grant access to information associated with it. Access control model based on situational awareness proposed in this paper is expected to be foundation of the development of access control technology on the basis of the status information in every circumstance changed frequently as well as cloud center.

It is planned to study a detailed mechanism to grant permission in accordance with the role and role management and analysis about the collection and process of status information.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2012-010886).

This work was supported by the Soonchunhyang University Research Fund.

References

1. Lee BY, Park JH, Yoo JS (2012) Distributed file systems architectures of the large data for cloud data services. J Korea Contents Assoc 12(2):30–39
2. Lee AR, Cho DE, Lee JY (2012) A study on the protection of user data in the cloud system. J Digit Policy Manage 10(11):12
3. Sandhu RS, Coyme EJ, Feinstein HL, Youman CE (1996) Role-based access control method. IEEE Comput 29(2):38–47

4. Ferratiom D, Cugini J, Kuhm R (1995) Role based access control (RBAC): features and motivations. In: Proceedings of 11th annual computer security application conference 1995
5. Dagstull GC, Dollimore J (1994) A security model for cooperative work: a model and its system implications. Positions paper for ACM European SIGOPS workshop 1994
6. RK Thomas, RS Sandhu (1997) Task-based authorization controls (TBAC): a family of models for active and enterprise-oriented authorization management. In: Proceedings of the IFIPWF11.3 workshop on database security 1997
7. Oh S, Park S (2000) Task-role based access control (T-RBAC): an improved access control model for enterprise environment. In: Proceedings of the 11th international conference on database and expert systems applications, pp 264–273
8. Moyer MJ, Ahamad M (2001) Generalized role-based access control. In: Proceedings of IEEE international conference on distributed computing systems (ICDSC 2001), pp 391–398
9. Dey AK, Abowd GD (2000) The context toolkit: aiding the development of context-aware applications. In: Workshop on software engineering for wearable and pervasive computing, June 2000
10. Pfleeger CP (1997) Security in computing, 2nd edn. Prentice-Hall International Inc, Upper Saddle River

Chapter 63
Whitelist Generation Technique for Industrial Firewall in SCADA Networks

Dong-Ho Kang, Byoung-Koo Kim, Jung-Chan Na and Kyoung-Sun Jhang

Abstract The use of SCADA systems has been increased since the 1960s as a need arose to more efficiently monitor and control the status of remote equipment. But, they are becoming more and more vulnerable to cyber-attacks due to utilize standard protocols and increase connectivity. The purpose of this paper is to introduce our work and confirm the validity of our approach for identifying network and application protocol attacks in SCADA networks.

Keywords Whitelist · Industrial firewall · SCADA · Cyber-attacks

63.1 Introduction

In general, SCADA (Supervisory control and data acquisition) systems include systems, software, and protocols used to monitor and control various physical infrastructures or facilities in industrial environments. In the past these systems were completely physically isolated environments from external networks and used proprietary hardware and protocols [1, 2]. But modern SCADA systems have

D.-H. Kang (✉) · B.-K. Kim · J.-C. Na
Convergence Security Research Section, ETRI (Electronics and Telecommunications Research Institute), Daejeon, South Korea
e-mail: dhkang@etri.re.kr

B.-K. Kim
e-mail: bkkim05@etri.re.kr

J.-C. Na
e-mail: njc@etri.re.kr

K.-S. Jhang
Department of Computer Engineering, Chungnam National University, Daejeon, South Korea
e-mail: sun@cnu.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_63, © Springer Science+Business Media Dordrecht 2014

distributed architecture and are connected to the corporate network and to the Internet. As well, these systems use general-purpose operation systems and industry-standard communication protocols such as Modbus, DNP3 for communication between a SCADA System and field devices such as PLC, RTU. The increased connectivity and the use of standard protocols can help to optimize manufacturing and distribution processes, but it also exposes these networks to the myriad security problems of the Internet [3].

The objective of this paper is to introduce our work and confirm the validity of our approach for preventing network and application protocol attacks in SCADA networks. For that, this paper is organized as follows. In the next section, we address the general SCADA architecture and protocols. Section 63.3 presents cyber-attacks characteristics of SCADA networks and defense approaches. Section 63.4 introduces our approach and Sect. 63.5 is given conclusion.

63.2 SCADA Architecture and Protocol

63.2.1 The SCADA Architecture

The SCADA architecture usually three different areas [4]. A typical SCADA system layout is shown in Fig. 63.1.

A control center includes HMIs, SCADA servers and historian systems for process control, the gathering of data in real time from field devices in order to control sensors and actuators. A field site includes multiple field devices that send commands to actuators and provide the data received from sensors to SCADA servers. Physical infrastructure consists of many different types of sensors and actuators that are monitored and controlled by a field device.

63.2.2 Modbus Protocol

Modbus is an application layer messaging protocol which provides client/server communication between devices in SCADA systems and offers services specified by function codes [5] (Fig. 63.2).

The client (master) that initiates a Modbus transaction builds the Modbus ADU (Application Data Unit). The Modbus ADU is consists of the MBAP (Modbus Application Protocol) header and the PDU (Protocol Data Unit). The PDU has a function code and function parameters. The function codes indicate to the server (slave) which kind of action to perform. The Modbus TCP/IP uses the TCP/IP stack for communication and extends the PUD with an IP header.

There are no security elements in the Modbus. Any attacker that can reach a Modbus server will be able to read and write to the field device as well as reboot

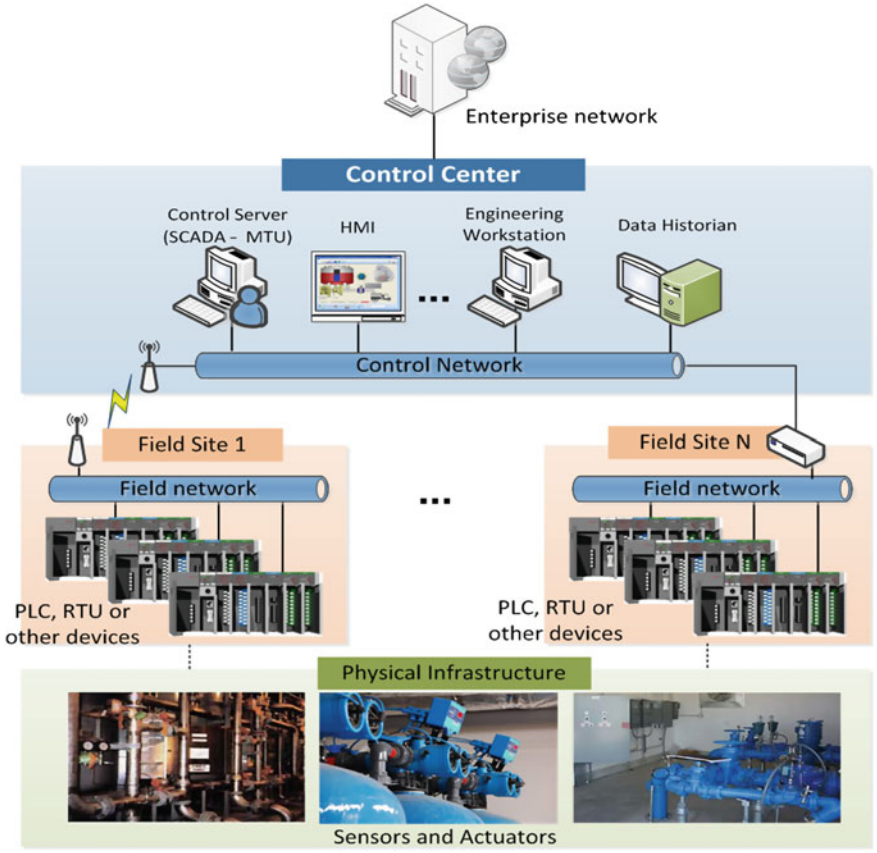


Fig. 63.1 SCADA system general layout

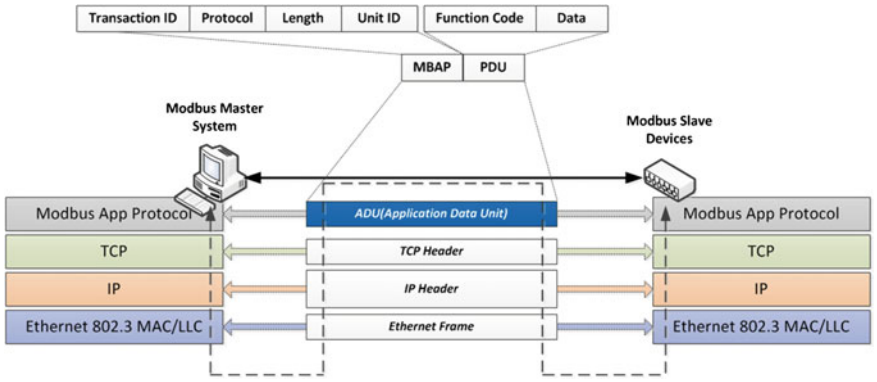


Fig. 63.2 The format of Modbus TCP/IP ADU

the device and run diagnostic commands. The simplicity of the Modbus protocol and widespread availability of free Modbus clients makes it relatively simple to attack a Modbus server [6].

63.3 Cyber Threats and Defense Approaches

63.3.1 Cyber Threats

We surveyed vulnerability assessment tools, Metasploit [7], Nessus [8], and Modscan [9] for the classification of Cyber-attacks in SCADA networks. These tools are commonly available to find known and newly discovered vulnerabilities on SCADA systems. And we surveyed some reports that were released by Digital Bond's Project Basecamp [10]. As a result of our survey, we describe that various type of attacks on SCADA systems can be grouped into two categories: Network Protocol attacks, Application Protocol attacks

Network Protocol attacks.

These types of Attacks use weak points of network protocols such as TCP/IP suite that have a number of serious security flaws. Therefore, most network protocol based attacks happened in Internet environment may be caused in SCADA networks were adopted IP network technologies. We introduce some types of Network protocol attacks (Table 63.1).

Host Discovery is the process for gathering information about each host, such as its operating system and version to verify whether they can be accessed or not. Using the information gathered about each target hosts in the host discovery step attackers launches scan to conform what ports are open, with listening services on target systems. Host Discovery and Scan attack are the commonly type of passive attacks to collect the fundamental information of vulnerabilities on target systems. DoS (Denial-of-Service) attack is active attack to make systems or network resource unavailable. Network protocol attacks have two characteristics as the following.

- ***Random access***: Host Discovery or Scan attacks generally send packets with the sequential or random destination addresses and ports to target networks or systems for obtaining the list of target systems and their services.
- ***Source address spoofing***: DoS attack does not consider about receiving responses to the attack packets. Therefore, Attackers can send packets with a forged source IP address for obscuring the true source of the attack.

To prevent unauthenticated access to field devices in a field network connected to an external network it is necessary to implement packet filtering technique between two networks. Packet filtering technique will make it harder for network protocol attacks to enter into field devices.

Table 63.1 Network protocol attacks

Attack type	Attacks
Host discovery	OS fingerprinting
Scan	TCP SYN/ACK scan
	TCP connect() scan
	TCP FIN Stealth Scan
	Xmas tree stealth scan
	TCP null stealth scan
	Windows scan
	RPC scan
	Version detection scan
DoS attack (Denial-of-service)	TCP/UDP flooding
	Smurf attack

Application Protocol attacks.

In our work, we only surveyed Modbus as an application protocol. Application Protocol attacks can cause damage to field devices being controlled by sending out improper commands, because they don't support integrity checking and authentication mechanism. Like network protocol attacks, these attacks also preceded by a step of gathering information about devices for finding vulnerable targets in a network. Table 63.2 shows generally types of application protocol attacks.

Application protocol attacks have a characteristic.

- ***Unpredictable Command***: SCADA systems generally produce predictable sets of command used for communication between a SCADA server and field devices. On the contrary, Application protocol attacks tend to use unconventional commands at irregularly interval.

Intrusion Detection System that is capable of recognizing SCADA traffic could detect suspicious or unconventional SCADA commands by inspecting a command value in the payload of all packets.

63.3.2 Defense Approaches

IP packet filtering.

As we described above, IP packet filtering has been recommended as an effective way to protect field devices from network protocol attacks. This technique used to control network access by monitoring outgoing and incoming packets and allowing them to pass or deny based on filtering rules using 3,4 layer on OSI model. A rule is composed to the source and destination IP addresses protocols and ports. Therefore, it is important to define filtering rules to block unauthenticated access originating from the other networks. This means blocking the majority of

Table 63.2 Application protocol attacks

Attack type	Attacks
Application scan	Modbus version scanner
	PLC Modbus mode identification
	PLC IO scan status
	Report slave ID
	Function code scan
Improper command execution	Force listen only mode
	Read/write request to a PLC
	Slave device busy exception code delay
	Acknowledge exception code delay
	Broadcast request from an client

incoming Internet traffic like http, ftp, telnet into the field site. And, IP packet filtering must permit only the normal connectivity between the SCADA server and field devices. Consequently, for securing field devices in SCADA systems from unauthenticated access the most important challenge is that filtering rules are completely defined. A flow whitelisting approach [11] addresses one of packet filtering techniques to enhance SCADA network security by identifying illegitimated access using whitelisted flows.

Network Intrusion Detection.

Intrusion Detection is the most well-known security approach in cyber security techniques. Misuse and anomaly based methods are used for intrusion detection. Misuse based methods are sometimes referred to as signature-based detection trying to detect abnormal behavior by analyzing the given traffic and matching several rules. While misuse detection is effective in known attacks, an obvious limitation of these methods is that they cannot detect new attacks whose signatures are unknown. As the example of research projects on misuse based detection for SCADA systems, Digital Bond developed SCADA IDS preprocessors and provide several attack signatures that have been defined for the open source IDS system Snort [12]. These signatures include some rules for detecting application protocol like DNP3, Modbus based attacks. Therefore, misuse based method could detect abnormal control commands using the predefined SCADA signatures. This method has to sometimes update new signatures for detecting the newly known attacks from external network. But, SCADA systems should be operated in the isolated environment as much as possible from any other network. Because of the reason, IDS based on misuse based method for SCADA is the unsuitable approach.

Anomaly based methods build models of normal data and detects any deviation from the normal model in incoming traffic. Anomaly detection methods have the advantage that they can detect attacks that are not known beforehand. But, anomaly based method have the problem that they often generate false alerts. This problem is especially very critical in a SCADA environment.

As an alternate to reduce the threat of cyber attack on SCADA, Network Intrusion Detection might become a proper approach. But, as we addressed above, Network traffic Intrusion Detection on the SCADA network should solve the critical problems.

63.4 Our Approach

Our goal is to automatically generate whitelists for industrial firewall in SCADA networks. For this purpose, we designed and implemented an industrial firewall with whitelists generation technique. Our system analyzes legitimate SCADA traffic on communication of SCADA servers and field devices to generate whitelists based on specific portions of each packet in the traffic. This system performs some normalization and preprocessing in order to extract appropriate information in each packet for generating whitelists. We assume that there are no any malicious packets in all captured traffic during the process of whitelists generation. Figure 63.3 shows the system architecture with whitelists generator and multi-layered filters.

Whitelists generator builds three types of lists during a given period of time that are used by multi-layered filter for identifying attack packets. System whitelist is the list for system authentication. Flow whitelist includes all legitimate flows between SCADA server and field devices. Modbus command whitelist defines the set of regular Modbus commands. Using these whitelists filters decide whether to pass the packet or drop it. We argued network and application based attacks may be occurred in SCADA environments have three characteristics in Sect. 63.3. Our system is able to identify these attacks in SCADA traffic using whitelists.

63.4.1 System Authentication

The system pattern analyzer extracts the source MAC/IP address of each packet in incoming traffic. This extracted list is an entry of system whitelist and is referred by the system filter. The filter treats system whitelist as the list of authenticated systems. The filter can identify all network based attacks that have the characteristic of spoofing using system whitelist. Figure 63.4 demonstrates an example of system authentication.

In this example, the compromised server (SCADA Server-1) sends packets with forged a source IP address to field devices to gathering information about field devices. Because the SCADA server-2 as an authenticated system is only stored in system whitelist, these packets is dropped by the system filter using system whitelist.

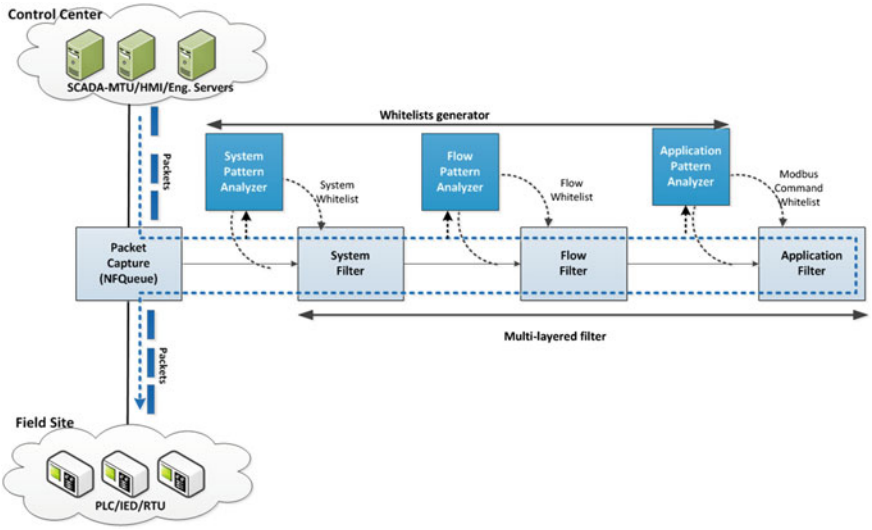


Fig. 63.3 The architecture of our system

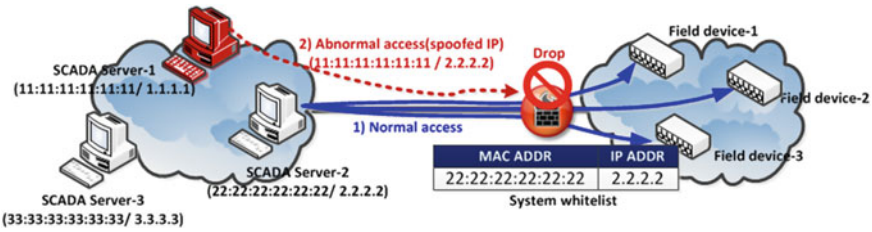


Fig. 63.4 System authentication using system whitelist

63.4.2 Flow Authentication

The flow pattern analyzer collects 5-tuple information (source/destination IP address, source/destination port, and Protocol) of each packet in incoming and outgoing traffic and keeps track of the state of network connection. This information is stored in flow whitelist and is treated as a legitimated flow by the flow filter. The filter can drop all network based attacks that have the characteristic of randomly access using flow whitelist. Figure 63.5 shows an example of flow authentication.

The compromised server (SCADA server-2) try to scan to conform what ports are open, with listening services on field devices. The system filter will forward all packets sent by the server to field devices due to treat the SCADA Server-2 as an authenticated system by the filter in Fig. 63.4. But the flow filter drops a serious of

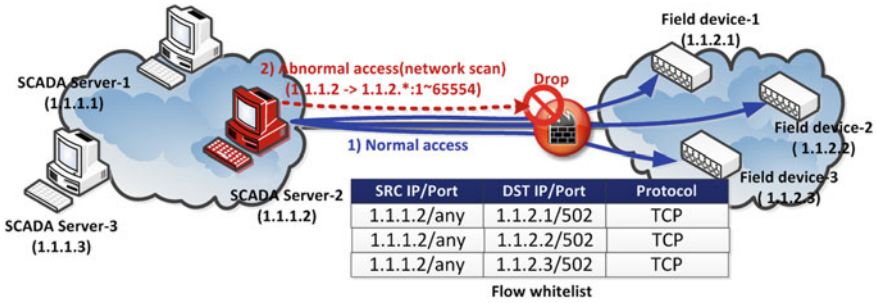


Fig. 63.5 Flow authentication using flow whitelist

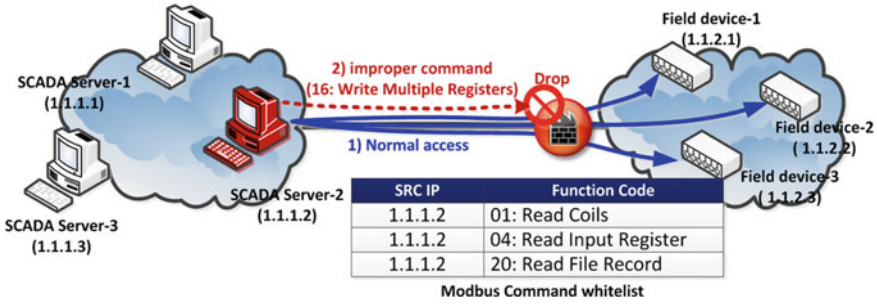


Fig. 63.6 The function code verification using Modbus command whitelist

packets using flow whitelist due Flow whitelist is registered only 502 port as an allowable port.

63.4.3 The Verification of Modbus Commands

The application pattern analyzer gathers function code of each message in Modbus traffic. This gathered list is an entry of Modbus command whitelist and is treated as permitted command by the application filter. The filter can identify all Modbus attacks that use inappropriate commands. Figure 63.6 demonstrates an example of the verification of Modbus commands.

If the compromised server (SCADA server-2) sends the command (Function code: 16) to field devices, the system and flow filter will pass the command using whitelists in Figs. 63.4 and 63.5. But because Modbus command whitelist contains only function code 01, 04, 20 as legitimated commands, the application filter will drop the packet with function code 16.

63.5 Conclusion

SCADA systems are facing the threat of cyber-attacks due to utilize standard open protocols and increase connectivity to external networks. We are working on an industrial firewall research project to decrease various cyber threats on SCADA systems. This paper described cyber-attacks and the existing defense mechanisms. And then, we introduced our approach. We have proposed the technique of whitelist generation to solve some problems faced by the existing IP packet filtering and Intrusion detection mechanisms for securing SCADA devices. Using whitelists multi-layered filters decide whether to pass the packet or drop it. In our work, the automatic build of whitelists is relatively easy. But, it is very hard to confirm whether the authenticated traffic is only routed in the internal network using our whitelists in the real world. We will address the evaluation of the proposed approach in our future work.

Acknowledgments This work was supported by the IT R&D program of MSIP/KEIT. [010041560, A development of anomaly detection and a multi-layered response technology to protect an intranet of a control system for the availability of pipeline facilities].

References

1. Stouffer K, Falco J, Scarfone K Guide to industrial control systems (ICS) security. Special publication NIST 800(82)
2. Galloway B, Hancke GP (2013) Introduction to industrial control networks. IEEE Commun Surv Tutorials 15(2), Second Quarter
3. Igure VM, Laughter SA, Williams RD (2006) Security issues in SCADA networks. Comput Secur 25(7):498–506
4. Stouffer K, Falco J, Kent K (2006) Guide to supervisory control and data acquisition (SCADA) and industrial control systems security. Special Publication NIST-SP-800-82-2006, National Institute of Standards and Technology (NIST)
5. Modbus Application Protocol V1.1b3, Modbus Organization, Inc. Retrieved 2 Aug 2013
6. http://www.digitalbond.com/scadapedia/protocols/modbus-2/
7. http://www.metasploit.com/
8. http://www.tenable.com/products/nessus
9. https://code.google.com/p/modscan/
10. http://www.digitalbond.com/tools/basecamp/
11. Barbosa R, Pras A, Sadre R (2013) Flow whitelisting in SCADA networks. In: Seventh annual IFIP working group 11.10 international conference on critical infrastructure protection
12. http://www.digitalbond.com/tools/quickdraw/

Chapter 64
Design of Industrial Firewall System Based on Multiple Access Control Filter

Byoungkoo Kim, Dong-Ho Kang, Jung-Chan Na, Beom-Hwan Chang and Tai-Myoung Chung

Abstract Although existing control systems have been developed and operated independently, the development of information technology has made access to control systems easier. Note, however, that such advancement of control systems gave rise to many security vulnerabilities; thus, the threat of cyber-attack is increasing as well. Since availability is the most critical factor in a control system, independent network security technology is most appropriate. The leading security device in such form is the firewall system. Therefore, this paper discusses the design and application of the IndusCAP-Gate (*Industrial Cyber Attack Prevention—Gate*) system, an industrial firewall system that fundamentally prevents unauthorized access to a control system. It applies access control filters of various levels to provide flexible and clear access control functionality. Most of all, the proposed system—unlike an IT firewall that implements access control to an unspecified mass—has a structure that facilitates the provision of security policy specific to each zone of the control system intranet used for access control of the specific system and service.

B. Kim (✉) · D.-H. Kang · J.-C. Na
Convergence Security Research Section, Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea
e-mail: bkkim05@etri.re.kr

D.-H. Kang
e-mail: Dhkang@etri.re.kr

J.-C. Na
e-mail: njc@etri.re.kr

B.-H. Chang
Cyber Investigation Police, Howon University, Gunsan, South Korea
e-mail: bchang@howon.ac.kr

B. Kim · T.-M. Chung
Department of Electrical and Computer Engineering, Sungkyunkwan University, Seoul, South Korea
e-mail: tmchung@ece.skku.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_64, © Springer Science+Business Media Dordrecht 2014

Keywords Control system · Unauthorized access · Multiple-access filter · Fire wall

64.1 Introduction

A control system is a computer based system that is widely used in typical factories having an automated production system and in national infrastructures such as power generation and resource management. It generally consists of various control devices such as sensor and actuator as well as the controller that controls the devices. Depending on the size and purpose, it may also have HMI (Human Machine Interface) and SCADA (Supervisory Control and Data Acquisition) server to monitor efficiently the remote device from a central position. For communication between such control systems, fast and efficient technologies such as Ethernet are applied more widely. Furthermore, it is evolving into an open system that enables interface to the Internet. Still, alongside the continuously increasing convergence of the control system and IT is the growing possibility of cyber-attacks on control systems. Although cyber-attack on control systems has been thought to be unlikely because of the network separation and use of closed control protocol, recent cyber-attacks on the control system—such as Stuxnet, Duqu, and Flame—have become a key issue. In other words, although the advancement of control systems by applying IT has many benefits in terms of efficiency, it also poses more security threats to a control system because of increased vulnerability.

Availability is the most essential factor for a control system since service should not be interrupted even for a moment. Therefore, the application of an independent form of security technology is preferred to the security software installed in the control device. The firewall system is the leading security device of such form [1]. Note, however, that the existing IT firewalls do not support dedicated protocols such as Modbus, ICCP, and DNP3 of control systems and characteristics of the control network; hence the need for the development and application of network security technology customized to the control system. Moreover, existing IT-based firewalls must perform broad access control targeting an unspecified number of systems and services, whereas industrial firewalls control access to specific systems and services. The IndusCAP-Gate system proposed in this paper applies various levels of access control filters to provide flexible and clear access control functions to specific system, service, and control command. It is also designed to conform to the concept of "Zone and Conduit" model as the ANSI/ISA-99 international standard; thus offering the benefit of having an effective structure to control access to each zone of the control system intranet.

The rest of this paper is organized as follows: Sect. 64.2 describes the security threats to the control system and the result and trend of existing studies to cope with such threats; Sect. 64.3 presents the design of the proposed IndusCAP-Gate system; Sect. 64.4 shows the result of implementation of the proposed technology; Sect. 64.5 discusses the conclusion and future plan.

64.2 Related Work

Since control systems are operated in closed environments, they were deemed safe from cyber-attack; hence the insufficient countermeasures to cyber-attacks. Even when IT security systems were installed, they were limited to some zones, and protection of the control systems such as PLC (Programmable Logic Controller) was inadequate. Moreover, control system security products that do not reflect the use of control system communication protocols such as Modbus and DNP3—which were designed without considering any authentication and encryption mechanism—and security coding have posed a serious threat to control system availability [2]. Figure 64.1 shows the main types of network attack that may occur in the control system intranet environment [3–6].

A firewall controls the traffic flow between the networks, and there are more attempts to apply the existing IT firewall technology to control systems [7]. Although the existing IT firewalls have relatively low level of control protocol analysis (DPI, deep packet inspection), its defense technology against DDoS attacks, etc., on control systems in TCP/IP environment is high. On the other hand, control system security devices feature outstanding analysis of control protocols, but the TCP/IP session analysis (SPI, Stateful Packet Inspection) and defense against DDoS attack are relatively weaker than the IT firewall technology. As such, although there is not much difference between the IT firewall and control system firewall, applying the IT firewall designed for the IT environment in the special control system environment—where availability is the first priority—has limitations.

Since the invasion of control systems can occur through not only malicious threats like cyber-attack but also an unintended mistake, the development of security technology suitable for the control system intranet environment is needed to protect the control systems. As such, many control system firewalls have been studied and developed for commercial purposes. They include Modbus DPI Firewall [8] from Tofino, SCADA Firewall [9] from Bayshore Network, Scalance [10] from Siemens, and Eagle mGuard [11] from Innominate. Of these, Modbus DPI Firewall (Modbus) and SCADA Firewall (DNP3) analyze the control protocols, whereas others perform TCP/IP-based access control. Other products have been developed and introduced in the market, but it is difficult to know the detailed operations of the commercial products; hence the need for security technology that supports the limited control protocol in the control system intranet and access control to each zone of the fixed network.

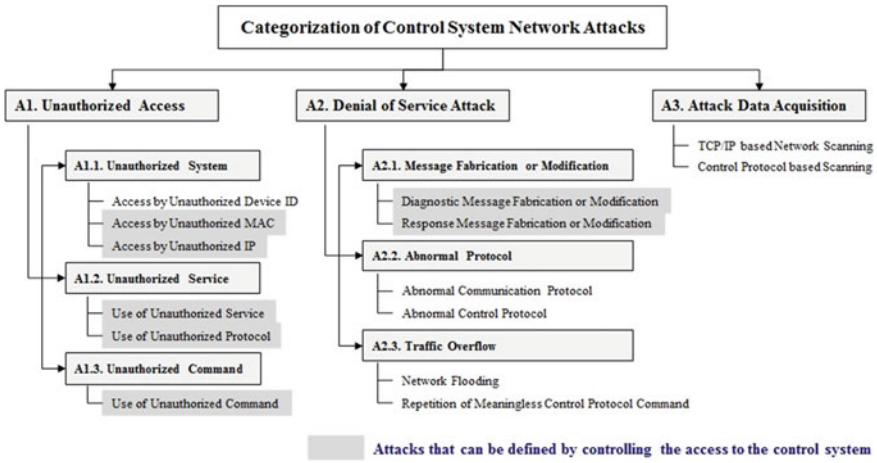


Fig. 64.1 Categorization of network attacks against control systems

64.3 Design of Industrial Firewall

64.3.1 Architecture and Application of the IndusCAP-Gate System

The IndusCAP-Gate system proposed in this paper provides a multiple-access control filter to block unauthorized access to the control system intranet; it is positioned between the controlled zones in the control system intranet to control traffic flow between the terminals. Figure 64.2 shows the overall architecture of the IndusCAP-Gate system, and Fig. 64.3 illustrates its application.

As shown in Fig. 64.2, the control system intranet can be mainly divided into SCA network and control network. The IndusCAP-Gate system can be installed in various positions according to the size and purpose of the control system intranet. Its architecture enables users to check unauthorized access through the user interface. The system is organized into four function blocks, each of which is described as follows:

- Packet Collection and Control Block: The block collects incoming packets through the network interface card (NIC) and delivers them to the access control blocks. It can perform the function of delivering or blocking packets according to the result of multiple-access control filter application. It is operated in inline mode—which blocks the packets if needed according to the processing result of the control network/application layer—or monitoring mode, which delivers them regardless of the processing result. It can also perform the default access control filter (default filter) function in interface unit depending on the existence of the access control filter policy of the access control block in the control network/application layer.

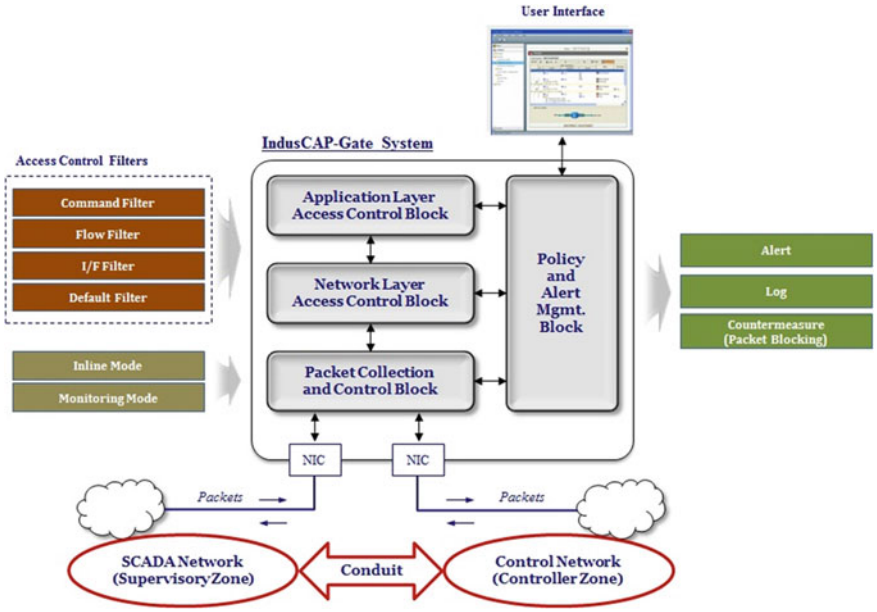


Fig. 64.2 IndusCAP-gate system architecture

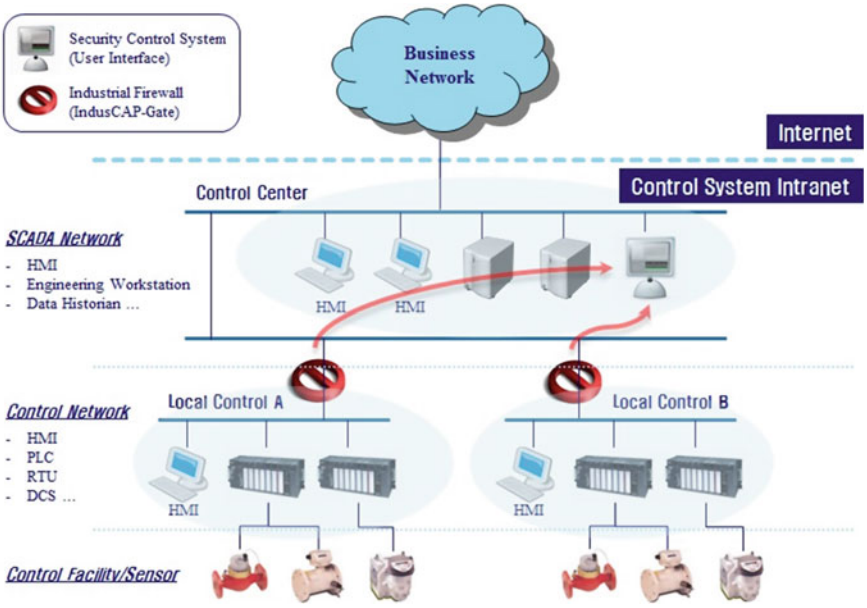


Fig. 64.3 IndusCAP-gate system application

- Network Layer Access Control Block: The block performs network-level access control in the control network layer. It analyzes incoming packets using the interface access control filter (I/F filter) and communication flow access control filter (Flow Filter), and then blocks unauthorized service, system, and protocol from accessing the interface or communication flow.
- Application Layer Access Control Block: The block performs application-level access control in the control application layer. It analyzes incoming packets using the control command access control filter (Command Filter), and then blocks unauthorized access to the control command.
- Policy and Alert Management Block: The block manages the access control filter policies of each access control block and measures to process (alert, log, and countermeasures) the results of packet collection, execution mode of processing blocks, and application of access control filter policies. It also connects the external user interface.

Using the blocks above, the IndusCAP-Gate system blocks unauthorized access to the control system. The access control function is executed by organically combining the default filter, I/F filter, flow filter, and command filter. Moreover, the interfacing of the SCADA network and control network in inline mode makes it suitable for the multilevel access control of the communication layer unit during communication between different security zones.

64.3.2 Multiple Access Control Filter Based Blocking of Unauthorized Access

As described above, the IndusCAP-Gate system's multiple access control filter based blocking of unauthorized access is performed by 4 filters. The leading control protocol, Modbus TCP [12], was the highest priority of system design.

Figure 64.4 shows the simple organization of processing and policies of multiple-access control filters in the communication layer unit. Each filter can be described as follows:

- Default Filter: As shown in Fig. 64.5, a default filter is enabled according to the existence of policy of other access control filters (disabled if there is at least one policy of other access control filters for each interface). It only decides whether the incoming packet will be allowed or denied. Since such enables total access control of incoming packets into a specific interface, it can be useful for special-purpose access control.
- I/F Filter: The policies of MAC/IP pair are applied for each interface. Only those packets conforming to the applied policies are selected and delivered to the opposite interface. In other words, it performs whitelist-based access control reflecting the authentication system data in the control system security domain connected to each interface.

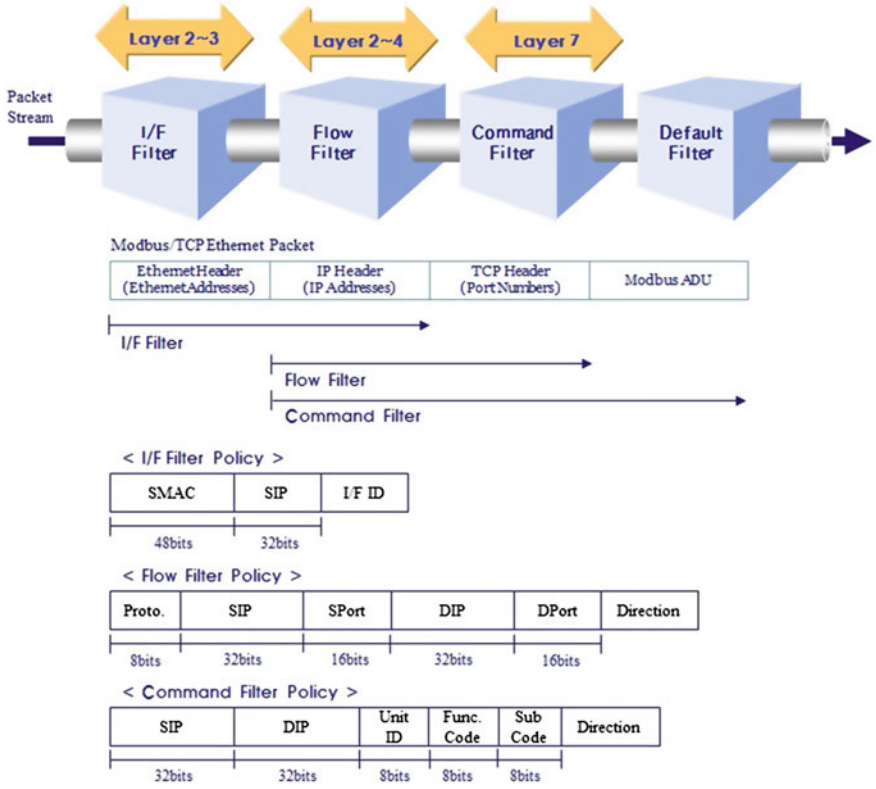


Fig. 64.4 Processing and policy of multiple-access control filter

- Flow Filter: The filter performs network-level, 5 tuple-based access control. It controls access to the service, system, and protocol units by packets that have passed the I/F filter.
- Command Filter: The filter performs application-level access control and analyzes the control protocol. It controls access to the Modbus control command by packets that have passed the I/F filter and flow filter. It can also apply some of the Modbus invasion signatures published by DigitalBond [13].

Figure 64.6 shows the overall packet processing flows of multiple-access control filters. As shown in the figure, processing of incoming packets into the interfaces is the same except those branching into each interface. Only the packets allowed through a filter can be delivered to the next filter. In other words, only those packets allowed through all filters are delivered to the opposite interface. The process allows the IndusCAP-Gate system to block unauthorized access to the control system and apply access control policies efficiently according to the size and nature of the control system intranet.

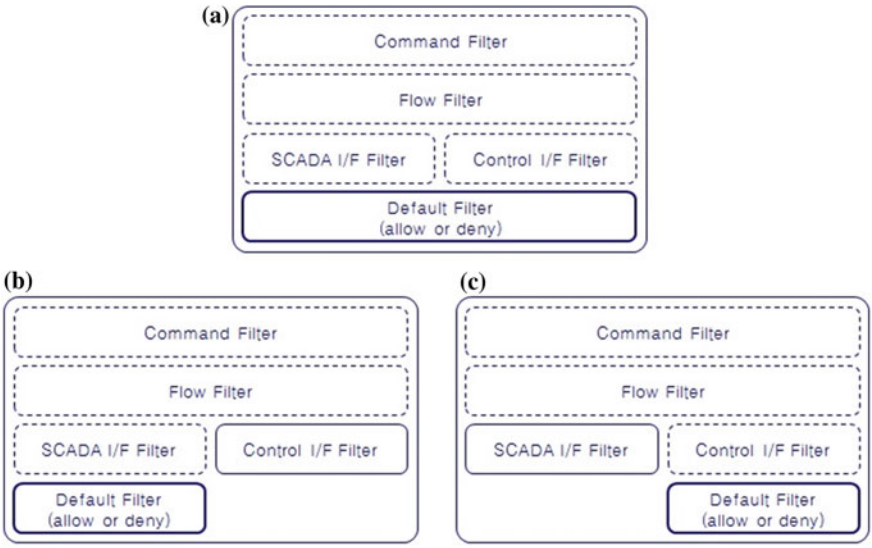


Fig. 64.5 Enabling default access control filter

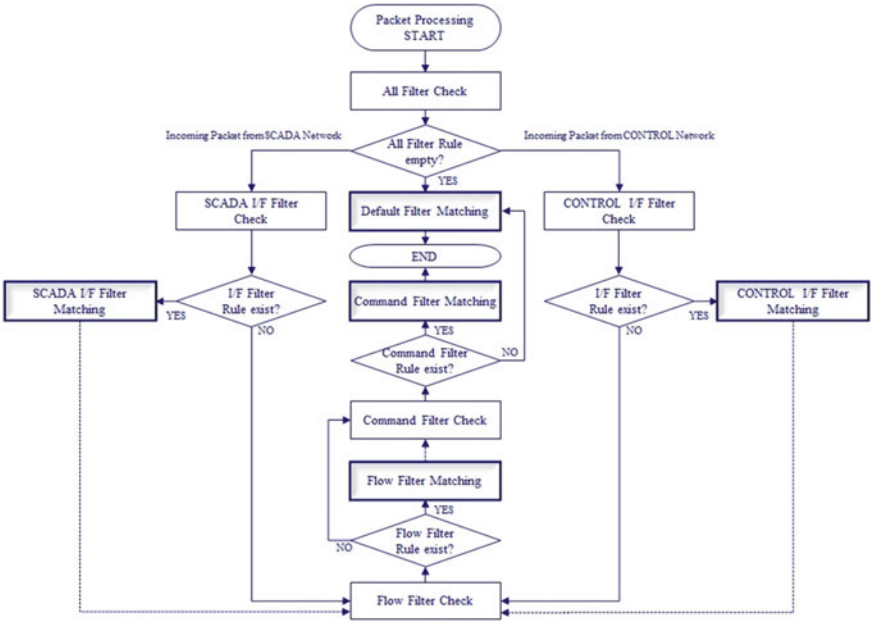


Fig. 64.6 Overall packet processing flow through multiple-access control filters

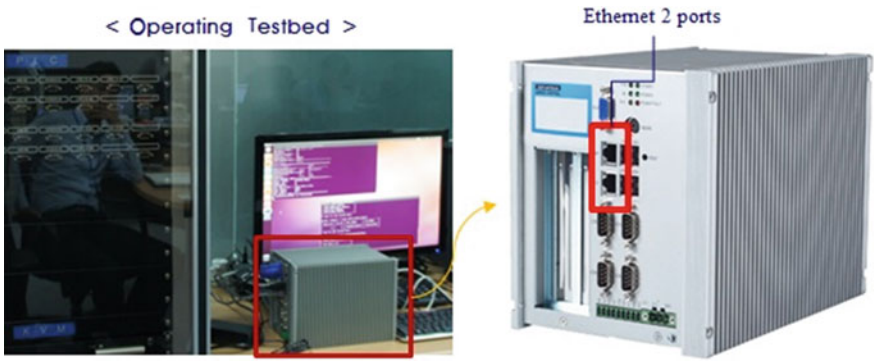


Fig. 64.7 IndusCAP-gate system prototype platform

64.4 Implementation and Evaluation

The multiple access control filter based blocking of unauthorized access as the technology proposed in this paper was developed to provide the access control technique suitable for the control system intranet. It can be installed between different zones in the control system structure to provide effective access control. To prove it, we developed a prototype of the IndusCAP-Gate system as industrial firewall. The IndusCAP-Gate system was implemented to run in Linux OS, adopting the UNO-3072L platform—which is a no-fan system—to suit the nature of the control system operating environment. Figure 64.7 illustrates the prototype platform and test-bed operation of the IndusCAP-Gate system. The packet processing performance of the IndusCAP-Gate system was tested using the IXIA traffic generator. Since the control system intranet environment generally has limited traffic, up to 20 Mbps packets transfers were tested. The test result showed that the system was able to process 100 % of incoming packets and was suitable for the control systems environment where availability is the first priority.

In addition, to secure software reliability through code safety checking, all software programs installed in the IndusCAP-Gate system were tested through the international standard code verification required by IEC 61508 as the standard for the procedure for verifying the safety of control system functions, IEC61511 as the standard for the functional safety of safety systems in the process industry, and CERT C language Secure Coding. High-level software safety and security were also assured by static testing and systematic dynamic testing using commercial tools such as Code Inspector and Sparrow.

64.5 Conclusions

This paper has described the design and implementation of the IndusCAP-Gate system, the multiple access control filter based industrial firewall system that blocks unauthorized access to control systems. The proposed system—unlike an IT firewall that implements access control to an unspecified mass—has a structure that facilitates the provision of the security policy specific to each zone of the control system intranet used for access control of the specific system and service. It applies access control filters of various levels to provide flexible, clear access control functions to specific system, service, and control command. Most of all, it is designed to conform to the concept of "Zone and Conduit" model as the ANSI/ISA-99 international standard; thus offering the benefit of having an effective structure to control access to each zone of the control system intranet.

In the future, we plan to address the problems identified by various tests and study the techniques to provide clearer and lighter access control functions. In addition, we will verify the safety of the system by installing and operating it in an actually operating control system intranet and expand the functionality to support the DNP3 protocol in addition to the Modbus protocol.

Acknowledgments This work was supported by the IT R&D program of MSIP/KEIT. [010041560, A development of anomaly detection and a multi-layered response technology to protect an intranet of a control system for the availability of pipeline facilities].

References

1. Yoon J, Kim W, Seo J (2012) Control system network security technology trend. 22(5):22–27
2. Yoo H, Yoon J, Sohn T (2013) Whitelist-based technique to detect abnormal situation for control system security. J Korea Inf Commun Soc 38B(08):641–653
3. Yoon J, Kim W, Seo J (2012) Study on technology requirement using the technological trend of security products concerning the industrial control system. J Korea Inst Inform Secur Crytology 22(5):22–26
4. Zhu B, Joseph A, Sastry S (2011) A taxonomy of cyber attacks on SCADA systems. In: Proceedings of IEEE international conference on internet things (iThings/CPSCom), pp 308–388. Dalian, China, Oct 2011
5. Guide to Industrial Control Systems (ICS) Security. NIST Special Publication 800(82), Jun 2011
6. Huitsing P, Chandia R, Papa M, Shenoi S (2008) Attack taxonomies for the Modbus protocol. Int J Crit Infrastruct Prot 1:37–44
7. Cheon Y (2009) Network design and architecture for industrial control system security. J Inf Process 19(5):60–67
8. Tofino Security Appliance. http://www.tofinosecurity.com/products
9. SCADA Firewall. http://www.bayshorenetworks.com/singlekey-scada-firewall.php
10. Scalance S. http://www.automation.siemens.com
11. mGuard. http://www.innominate.com/en/products
12. Modbus IDA (2004) Modbus application protocol specification v1.1a. North Grafton, Massachusetts
13. DigitalBond's Basecamp Project. http://www.digitalbond.com/tools/basecamp

Chapter 65
The Improvement for Performance of Inter-VTS Data Exchange Format Protocol in VTS System

Namje Park

Abstract This paper suggests the library which supports different parsers depending on applications in order to make up for the defects of open source. Using the library, systems have been developed to simulate the IEVF service on the wireless environment, and time and memory used for the V-145 protocol for each parser have been measured. Android-based smart phones were used to emulate the vessels, and Windows7-based desk tops were used to emulate the VTS service center.

Keywords VTS · Vessel traffic service · IVEF · Performance · Maritime

65.1 Introduction

e-Navigation is the next-generation marine navigation system which is based on electronic technologies and IT technologies. It includes various systems, standard protocol and related software programs which provide stability of vessel navigation and efficient route information, and protect the marine environment. e-Navigation consists of 7 sectors (VTS, AIS, ECDIS, IBS/INS, ARPA, LRIT and GMDSS). VTS means the vessel traffic system. International Association of Lighthouse Authorities (IALA) is developing the standard Inter-system VTS Exchange Format (IVEF) protocol for exchange of navigation and vessel

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013 R1A1A4A01013587).

N. Park (✉)
Department of Computer Education, Teachers College, Jeju National University,
61 Iljudong-ro, Jeju-si, Jeju 690-781, Korea
e-mail: namjepark@jejunu.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_65, © Springer Science+Business Media Dordrecht 2014

information between VTS systems and between VTS and vessels. The output of standardization is distributed as the IALA recommendation V-145, and the protocol is implemented with an open source. The IVEF open source, however, is the code used to check the functions of standard protocols. It is too slow to be used in the field, and requires a large memory.

This paper describes how the DOM-parser-based EtriIVEF library is converted into the XmlPullParser-based library. The XmlPullParser library is an intermediate form between SAX and DOM. It has the advantages of SAX (requiring less memory) and DOM (easy to design the code). Therefore, the XmlPullParser-based V-145 protocol is expected to enhance the speed and memory use when compared with the existing library. This paper summarizes the theory of SAX, DOM and Pull parser, and describes how to implement the EntriVEF Pull Parser. This paper also compares the performance between the parsers when they are used in the V-145 protocol, and provides the login method using the MacAddress of the phones.

This paper suggests the library which supports different parsers depending on applications in order to make up for the defects of open source. Using the library, systems have been developed to simulate the IEVF service on the wireless environment, and time and memory used for the V-145 protocol for each parser have been measured. Android-based smart phones were used to emulate the vessels, and Windows7-based desk tops were used to emulate the VTS service center.

65.2 IVEF (Inter-VTS Exchange Format) Service

The IVEF service is the framework for exchange of information for vessel traffic, and is being developed by the e-Navigation Working Group of IALA. The IALA committee recommends V-145 as the communication protocol between the VTS systems, and provides the downloading service of the V145-protocol-related SDK and document on the OpenIVEF site [3]. The IVEF service works as the server/client model to exchange traffic conditions between the VTS systems. Messages are roughly divided into the control information messages and the real-time information messages. Control information messages are subdivided into the user authentication messages, shutdown messages, service requests to servers, response messages, and the messages providing the server status. Real-time information message consists of the current location of a vessel, the expected route, the place anchored, and other physical information on the vessel.

65.3 XML Parser in VTS Platform

XML parsers are classified into pull-based parsers and push-based parsers depending on the operation method. DOM parser is a typical pull-based parser, and SAX parser is a typical push-based parser. They have conflicting performance in

terms of memory use, node searching speed and random access speed, and have merits and demerits in convenience of development.

65.3.1 Pull-Based Parser

A pull-based API receives data directly from a parser if there is any event or node a user wishes to handle. Since the structure of the parsing source code is similar to the schema of XML documents, developers can easily analyze the code and use the parser API. Normally, a pull-based parser ① resets the parser, ② conducts searching from the outermost nodes of XML to the sub-nodes, and then, ③ performs appropriate action if it finds a desired node or attribute.

65.3.2 Push-Based Parser

In a push-based API, a user registers a callback function to handle desired events, and if the desired event occurs, the callback function is invoked in the library or the system. The application program cannot control other events than the ones given by the parser. The push-based API requires less memory and is faster in searching. However, since it works based on the state machine to control events, however, it is difficult for developers to use the parser API.

65.4 Comparison of Performance Time

65.4.1 Performance Time on PC (ms)

The performance time on PC is fastest in EIVEF_PullParser. It was expected that SAX parser would take less time, but in reality, it was faster in the Pull parser. It is because all the node events are handled for SAX, and no more events are handled for Pull parser once the required data are extracted (Fig. 65.1).

65.4.2 Performance Time on PC (ms)

Memory usage was the highest in the DOM memory as expected. It was lowest in the Pull parser, and higher than expected in the SAX parser. It is because of the structures required internally for generation of event and frequent invocation of

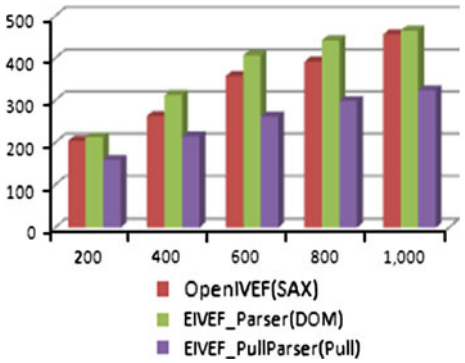


Fig. 65.1 Graph on performance time on PC for each parser

functions. It seems that dummy memory which is not actually used is also measured because garbage collection is not performed in time (Fig. 65.2).

65.4.3 Performance Time on Android (ms)

The tests on the Android phone showed that the Pull parser has better performance in general. It was unusual that, unlike the tests on PC, the DOM parser showed better performance than the SAX parser. It is because functions are invoked frequently in the SAX parser, and the function call processing is delayed on the Android platform (Fig. 65.3).

V-145 message protocol was performed in XMLPullParser faster by about 70 % than in the SAX-based open source, and by about 20 % than in the DOM parser. It was unusual that it was slower in the SAX parser than in the DOM parser, which is because it took time to process frequent function calls to process SAX parser due to the characteristics of the Android system. Also, it is because all the node events are handled for SAX, and no more events are handled for XmlPullParser once the required data are extracted.

65.4.4 Comparison of Memory Usage

The memory usage on the Android phone for the DOM parser was increased to about 15.6 MB when there were 1,000 objects. For XmlPullParser, the memory usage was about 6 M for 1,000 objects. Contrary to expectations, the memory usage increased continuously for SAX, and it is because the memory used to manage the status was not properly cleared. Since it is a garbage memory, SAX is expected to use the same amount of memory as the Pull parser (Fig. 65.4).

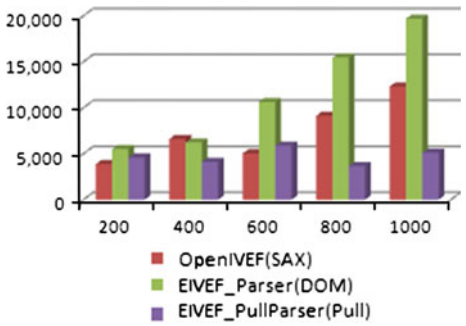


Fig. 65.2 Graph on memory usage on PC for each parser

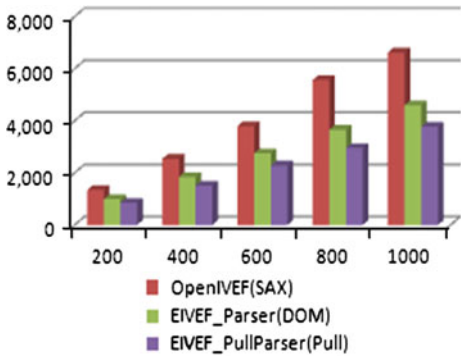


Fig. 65.3 Graph on performance time on phone for each parser

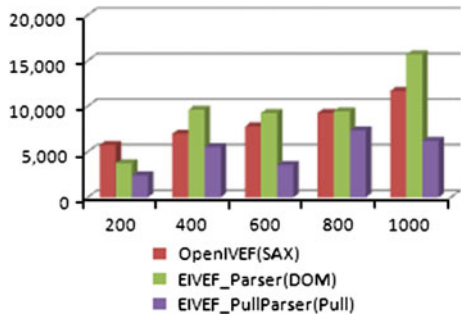


Fig. 65.4 Graph on memory usage on phone for each parser

65.5 Conclusion

This paper describes the wireless IVEF protocol system using XmlPullParser which is the intermediate form between the DOM parser and the SAX parser. The performance of handling V-145 protocol message was analyzed with the developed system. The experiment showed that XmlPullParser processes messages faster and requires less memory space than the DOM parser or the SAX parser. Since the V-145 protocol messages generate a large number of small messages, it

is required to select the parser depending on whether the protocol is received by vessels or the VTS control center. In the test, all the nodes and attributes of the V-145 messages were measured. In the future study, it is required to measure performance for random access of a specific message to a specific node.

References

1. Laxhammar E, Falkman G, Sviestins E (2009) Anomaly detection in sea treffic—a comparison of the gaussian mixture model and the kernel density estimator. In: ISIF 2009
2. Kawaguchi A et al (2009) Towards the development of intelligent navigation support systems gor group shipping and global marine traffic control. IET Intell Transp Syst 3(3):257–267
3. Lee BG, Han JW, Jo HS (2010) Design of situation awareness and aids to navigation structure of VTS for maritime safety. J Korea Inf Commun Soc 35(7):1073–1080
4. Park N, Kwak J, Kim S, Won D, Kim H (2006) WIPI mobile platform with secure service for mobile RFID network environment. In: Shen HT, Li J, Li M, Ni J, Wang W (eds) APWeb workshops 2006. LNCS, vol 3842. Springer, Heidelberg, pp 741–748
5. Park N (2010) Security scheme for managing a large quantity of individual information in RFID environment. In: Zhu R, Zhang Y, Liu B, Liu C (eds) ICICA 2010, vol 106., CCISSpringer, Heidelberg, pp 72–79
6. Park N (2011) Secure UHF/HF dual-band RFID: strategic framework approaches and application solutions. ICCCI 2011., LNCSSpringer, Heidelberg, pp 488–496
7. Park N (2011) Implementation of terminal middleware platform for mobile RFID computing. Int J Ad Hoc Ubiquitous Comput 8(4):205–219
8. Park N, Kim Y (2010) Harmful adult multimedia contents filtering method in mobile RFID service environment. In: Pan J-S, Chen S-M, Nguyen NT (eds) ICCCI 2010, vol 6422., LNCS(LNAI)Springer, Heidelberg, pp 193–202
9. Park N, Song Y (2010) AONT encryption based application data management in mobile RFID environment. In: Pan J-S, Chen S-M, Nguyen NT (eds) ICCCI 2010, vol 6422., LNCS(LNAI)Springer, Heidelberg, pp 142–152
10. Park N (2011) Customized healthcare infrastructure using privacy weight level based on smart device. In: Lee G, Howard D, Ślęzak D (eds) Communications in computer and information science, vol 206. Springer, Heidelberg, pp 467–474
11. Park N (2011) Secure data access control scheme using type-based re-encryption in cloud environment. In: Katarzyniak R et al (eds) Studies in computational intelligence, vol 381. Springer, Heidelberg, pp 319–327
12. Park N, Song Y (2010) Secure RFID application data management using all-or-nothing transform encryption. In: Pandurangan G, Anil Kumar VS, Ming G, Liu Y, Li Y (eds) WASA 2010. LNCS, vol 6221. Springer, Heidelberg, pp 245–252
13. Park N (2010) The implementation of open embedded S/W platform for secure mobile RFID reader. J Korea Inf Commun Soc 35(5):785–793
14. Park N, Ko Y (2012) Computer education's teaching-learning methods using educational programming language based on STEAM education. In: Park J, Zomaya A, Yeo S et al (eds) NPC 2012, vol 7513. Springer, Heidelberg, pp 320–327
15. Ko Y, An J, Park N (2012) Development of computer, math, art convergence education lesson plans based on smart grid technology. In: Kim T, Stoica A, Fang W et al (eds) Computer applications for security, control and system engineering, vol 339. Springer, Heidelberg, pp 109–114
16. Ko Y, Park N (2011) Experiment and verification of teaching fractal geometry concepts using a logo-based framework for elementary school children. In: Kim T, Adeli H, Slezak D et al (eds) Future generation information technology, vol 7105. Springer, Heidelberg, pp 257–267

17. An J, Park N (2012) The effect of EPL programming based on CPS model for enhancing elementary school students'creativity. In: Park JJH, Jeong Y, Park SO et al (eds) Embedded and multimedia computing technology and service, vol 181. Springer, The Netherlands, pp 237–244
18. An J, Park N (2011) Computer application in elementary education bases on fractal geometry theory using LOGO programming. In: Park JJ, Arabnia H, Chang H et al (eds) IT convergence and services, vol 107. Springer, The Netherlands, pp 241–249
19. Kim Y, Park N (2012) Development and application of STEAM teaching model based on the rube goldberg's invention. In: Yeo S, Pan Y, Lee YS et al (eds) Computer science and its applications, vol 203. Springer, The Netherlands, pp 693–698
20. Park N, Cho S, Kim B et al (2012) Security enhancement of user authentication scheme using IVEF in vessel traffic service system. In: Yeo S, Pan Y, Lee YS et al (eds) Computer science and its applications, vol 203. Springer, The Netherlands, pp 699–705

Chapter 66
Abnormal Behavior Detection Technique Based on Big Data

Hyunjoo Kim, Ikkyun Kim and Tai-Myoung Chung

Abstract Nowadays, cyber-targeted attacks such as APT are rapidly growing as a social and national threat. As an intelligent cyber-attack, the cyber-targeted attack infiltrates the target organization or enterprise clandestinely using various methods and causes considerable damage by making a final attack after long-term and through preparations. Detecting these attacks requires collecting and analyzing data from various sources (network, host, security equipment) over the long haul. Therefore, this paper describes the system that responds to the cyber-targeted attack based on Big Data and a method of abnormal behavior detection among the cyber-targeted attack detection techniques provided by the proposed system. Specifically, the proposed system analyzes faster and precisely various logs and monitoring data that have been discarded using Big Data storage and processing technology; it also provides integrated security intelligence technology through data correlation analysis. In particular, abnormal behavior detection using MapReduce is effective in analyzing large-scale host behavior monitoring data.

Keywords Cyber-targeted attack · Big data · Abnormal behavior detection · Feature · Correlation analysis · MapReduce · Security intelligence technology

H. Kim (✉) · I. Kim
Network Security Research Team, Cyber Security Research Department,
Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea
e-mail: hjookim@etri.re.kr

I. Kim
e-mail: ikkim@etri.re.kr

H. Kim · T.-M. Chung
Computer Engineering Department, SungKyunKwan University, Suwon, Korea
e-mail: tmchung@ece.skku.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_66, © Springer Science+Business Media Dordrecht 2014

66.1 Introduction

Unlike the existing attack targeting unspecified systems, the cyber-targeted attack is an organizational attack with a clear target; nowadays, it is used as an attack method for cyber terror, hacktivism, and cyber warfare. The cyber-targeted attack is rapidly growing as an intelligent, persistent social and national threat designed to steal industrial secrets or military secrets from major government agencies or enterprises and customer information, paralyze the industrial control system and consequently cause astronomical physical damages, or wage an act of war [1]. APT (Advanced Persistent Attack)—such as the SK Communications hacking and 3.20 Cyber Incident—is an intelligent cyber-targeted attack designed to steal confidential information or seize control over major facilities after infiltrating the network of the target organization (e.g., enterprise or organization). It is difficult to detect and respond to an attack in advance because the attack is made over a long period of time (not a one-time attack), and various malicious codes and attack roots are used. Detecting an attack requires collecting the organization's large-scale data from various sources (network, host/server, security equipment, etc.) over the long haul and analyzing the behavior history. In addition, correlation analysis is important in understanding the meaning of each individual attack, since confidential attacks are attempted continuously until a final attack is successful.

With the development of cloud computing base technology, we now have enough computing power to process large-scale data (e.g., Big Data). Therefore, we can integrate and analyze the numerous attack data that was difficult to do in the past. As a result, security intelligence technology based on Big Data analysis emerged. This paper proposes the cyber-targeted attack detection and response system and an abnormal behavior detection method using MapReduce, the security intelligence technology based on Big Data that integrates and analyzes enormous amounts of security logs, network and host information, and application transactions.

The rest of this paper is organized as follows: Chap. 2 introduces the security intelligence technology and security technology research using Big Data as related studies; Chap. 3 describes the proposed system and abnormal behavior detection method using MapReduce; Lastly, Chap. 5 presents the conclusion and future works.

66.2 Related Research

66.2.1 Security Intelligence Technology

Security intelligence technology is integrated security management technology that configures network and system security events as interrelated security infrastructure to defend against a cyber-targeted attack. Research studies on internal behavior surveillance technology and product development are conducted in full

scale using Big Data processing/analysis technology. With the Big Data processing/analysis technology expanded to various application areas, the technology is utilized as security event analysis technology. SIEM (Security Information and Event Management) leaders are introducing and utilizing Big Data analysis technology to apply intelligent security and respond to new attacks. SIEM provides a function that recognizes the security context by integrating statistical information and security event information from various security equipment and servers and network equipment and analyzing correlations among such kinds of information [1].

Related overseas projects include the CINDER (Cyber-INsiDER) and cyber genome programs that have been implemented by the US's DARPA since 2010. The CINDER program focuses on the analysis of internal staff's behavior, whereas the cyber genome project expresses the correlation and properties among application software, data flow, and users based on the formal analysis of abnormal behavior [2, 3]. IBM QRader is a representative product that provides security intelligence technology. Having acquired Q1 Labs, which provides strong SIEM technology, IBM provides the function of collecting various sources and analyzing the network and application behavior using QRader. For this, Netflow data is processed, with all events collected from sources analyzed in detail to present a solution for security incidents [4]. Nonetheless, QRader is not mature enough to make an analysis pattern for events collected from various sources. Splunk provides predefined functions to support enterprise security monitoring and analysis and use case. For this, large-scale data and real-time correlation are analyzed flexibly using the Splunk application. Note, however, that the Splunk application has shortcomings, e.g., needs to be customized for enterprise security, not suitable for the security organization that lacks professional knowledge.

66.2.2 Security Research Using Big Data

As described earlier, Big Data storage and processing technology can be utilized in the security area. Currently, theses related to security using Big Data are introduced besides products providing security intelligence technology. Y. Lee et al. explained the technology that analyzes Internet traffic using Hadoop MapReduce. To analyze the IP, TCP, HTTP, and Netflow data, MapReduce was implemented for each, with the I/O format defined. According to the research result, 14 Gbps throughput was obtained in analyzing a 5 TB file on 200 test nodes [6]. Another research proposed a MapReduce-based security log analysis system that collects and analyzes large-scale heterogeneous security logs in an integrated manner. The mixing ratio and execution time were compared among the firewall, intrusion detection system, and web logs [7]. Despite yielding meaningful results, these studies are not suitable for the area requiring the correlation analysis of various-kinds information such as the detection of the cyber-targeted attack, since only one kind of data is analyzed using Big Data technology.

66.3 Abnormal Behavior Detection Based on Big Data

The APT attack intrudes upon the internal network and goes through a latent period to prepare a final attack (confidential information leak, system breakdown, etc.). The cyber-targeted attack response system proposed by this paper aims to cope with the cyber-targeted attack by detecting abnormal behavior in advance before the final attack is made. This paper focuses on the cyber-targeted attack response system (SINBAPT: Security Intelligence techNology for Blocking APT) and Big Data processing/analysis method of the SINBAPT system to detect the abnormal behavior.

66.3.1 Cyber-Targeted Attack Response System

The system proposed in this paper provides a function that detects abnormal behavior by analyzing the correlation between security events occurring in the existing security equipment, network traffic and statistical information, and all behavior information occurring at the host to detect the cyber-targeted attack. In addition, the system provides the function traces back the attack source based on the analysis results and security visualization for administrator to see and understand the information conveniently. Figure 66.1 shows the functions of SINBAPT system in layers.

Data agent in the data and event sensor layer is installed in the target system for collecting the data; it senses the data generated by each system and transfers the data to the data collection layer. On the other hand, the data collection layer receives the data from data sensors, saves large-scale accumulated data in the Hadoop file system (HDFS: Hadoop Distributed File System), and transfers the real-time processing data to the real-time big data processing part. The large-scale data storage/processing layer is composed of HDFS and Hbase (NoSQL) to save the large-scale data, and MySQL Cluster (in-memory DBMS for real-time processing). The large-scale accumulated data and real-time processing data will be processed by MapReduce and Esper/Storm, respectively. The targeted attack analysis layer consists of real-time data analysis, accumulated data analysis, and correlation analysis. Real-time data analysis is composed of LSAE (Legacy Security event Analysis Engine), which analyzes security logs and alerts generated by security equipment, NBAE (Network-Based Analysis Engine), which analyzes the data related to network traffic, and HBAE (Host-Based Analysis Engine), which analyzes the data related to behavior occurring at the host. For accumulated data analysis, LBAE (Large-scale Batch Analysis Engine)—which accumulates all data generated by security equipment, network, and host over a long period of time—is provided; it performs analysis periodically or upon request. The real-time and accumulated data analysis engine detects abnormal behavior using the mining-based analysis model and rule-based analysis method. CRAE (Co-Relational

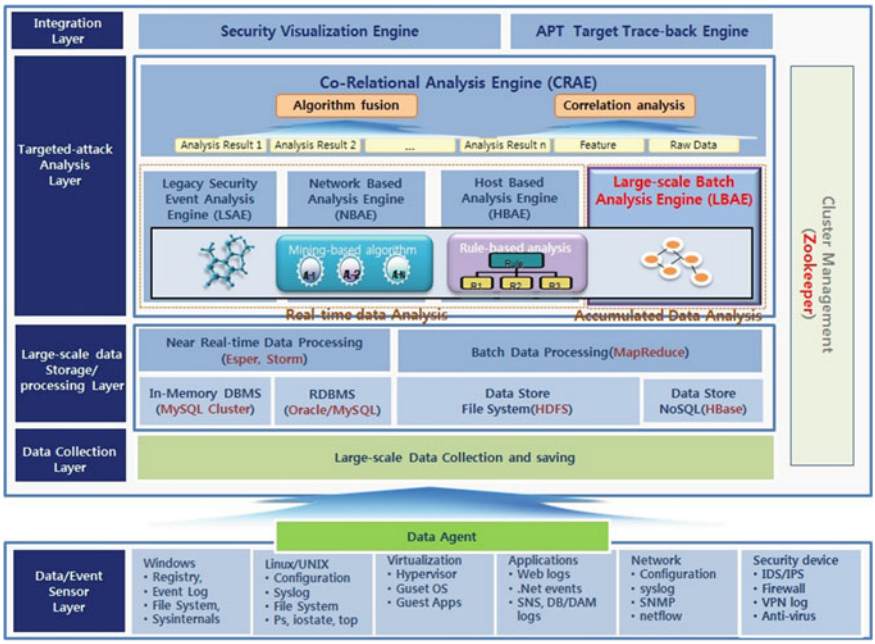


Fig. 66.1 Structure of the cyber-targeted attack response system (*SINBAPT*)

Analysis Engine) finally detects abnormal behavior by analyzing in detail the data in other areas and which are related to the abnormal behavior analysis result to identify correlations among the data. Lastly, the integration layer is composed of the security visualization engine that provides the analysis result generated by the targeted attack analysis layer to users—using visualization technology—so that they can understand the security context more intuitively and the APT track-back engine that traces back the attack source based on the analysis results using the APT tracking mechanism.

66.3.2 Large-Scale Data Analysis Engine

This paper describes the large-scale accumulated data analysis method performed by the large-scale data analysis engine of the SINBAPT system. Although the SINBAPT system takes all large-scale data (security event, network, host, etc.) collected from multiple sources as analysis target, this paper will focus on the analysis of the large-scale accumulated data generated by the host. This is because traces of malicious code or hacking—which are frequently used for a cyber-targeted attack—are mainly left in the host.

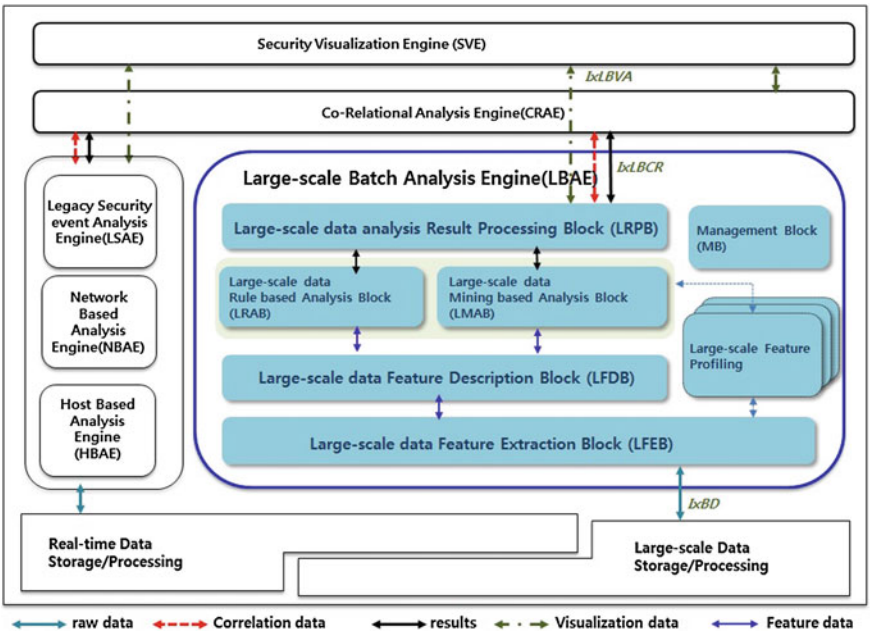


Fig. 66.2 Structure of the large-scale data analysis engine

Figure 66.2 shows the structure of the large-scale data analysis engine, which has the IxBD (Interface-BigData) interface (used to access the large-scale data needed for large-scale data storage/processing), IxLBCR (Interface-LB to CR) interface (used to send and receive the data needed for correlation analysis), and IxLBVA (Interface-LB to VA) interface with the visualization engine (to visualize large-scale data analysis results). Table 66.1 summarizes the functions of the large-scale data analysis engine with their I/O data.

Feature profiling is the function that defines a feature needed for rule and mining-based large-scale accumulated data analysis as well as its properties and relations. Table 66.2 shows the definition of major host-based features. Currently, the host targets the Windows OS system.

All program behaviors including malicious code are performed in the OS as one of four categories. (In other words, behavior related to the file system, registry, process, and network.) Based on this classification, suspicious behavior and information are defined as a feature, and features are extracted from the large-scale accumulated data (related to the host) saved in HDFS for each process using the feature extraction function. The extracted feature data is presented in a form needed for analysis using the description function, as shown in Fig. 66.3.

The described data is created for each process. At this time, the process is analyzed by combining the parent process and child process into a single process to identify the subject and flow of overall behavior. This is because most malicious codes perform abnormal behavior by creating a new process.

Table 66.1 Functions of the large-scale data analysis engine

Function name	I/O data	Description
Feature profiling	Defined in advance according to the analysis algorithm or rule	Defines the feature needed for large-scale accumulated data analysis and profiles its properties and relations
Feature extraction	Large-scale accumulated data/feature information	Extracts a feature from the large-scale accumulated data
Feature description	Feature information/described feature information	Processes the information of the extracted feature into a form used for analysis
Rule-based analysis	Described feature information/ analysis results	Rule-based, large-scale accumulated data analysis
Mining-based analysis	Described feature information/ analysis results	Mining-based, large-scale accumulated data analysis
Analysis result processing	Analysis results/data type used by the correlation and visualization engine	Processes analysis results into the data requested by the correlation and visualization engine
Management	–	Overall management function for the analysis engine (e.g., monitoring interval setting)

Table 66.2 Major host-based features

Type	Major host-based features	Code	Risk level
File	File deletion in the system folder	F1	H
	File renaming in the system folder	F2	H
	File creation in the system folder	F3	H
	File creation in the temporary folder	F4	H
	Executable file creation	F5	H
	File creation in the temporary folder	F6	M
	File creation	F7	M
Registry	Registry deletion	R1	H
	Service deletion	R2	H
	Adding automatic execution	R3	H
	Registry registration	R4	H
	Service registration	R5	H
	Adding a BHO item	R6	M
Process	Other process creation	P1	H
	Other process termination	P2	H
	Other process search	P3	H
	DLL code injection	P4	H
	Thread creation	P5	M
Network	Port opening	N1	M
	Port binding	N2	M
	Network connection	N3	M
	Network disconnection	N4	M
	Data sending	N5	M
	Data receiving	N6	M

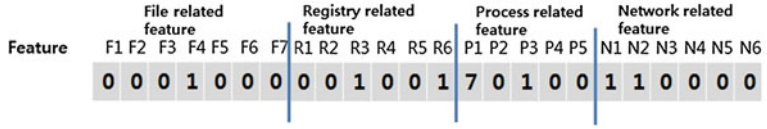


Fig. 66.3 Feature description

66.3.3 Abnormal Behavior Detection Method Using MapReduce

In this paper, MapReduce is used to extract and describe a feature from the large-scale data saved in HDFS and to perform initial analysis. MapReduce improves the processing and analysis speed by collectively handling the big data accumulated over the long haul in a distributed manner.

Figure 66.4 shows the MapReduce handling process that analyzes large-scale accumulated data to detect host-based abnormal behavior. MapReduce receives the host-based large-scale accumulated data and divides it to perform a map task. Mappers performing the task read the large-scale accumulated data based on host behavior sequentially. If behavior defined as a feature occurs, structs are created for each process (a, b, c in Fig. 66.4 are the process identifier) as shown in Fig. 66.3, with the feature value changed to 1. The interim result value generated by the Map() function is entered as input value for the Reduce() function in a format wherein values are combined by process after going through the Partition and Shuffle phase. By taking the process as key, the Reduce() function can identify the type and frequency of abnormal behavior performed by the process during a certain period of time (analysis interval) by executing the sum operation for each bit of the feature struct. The abnormal behavior detection result using MapReduce is analyzed by the abnormal behavior detection rule defined in feature profiling or used as input for the data mining algorithm. The result can also be used as trigger point for correlation analysis.

Map(): (key, value) = (file offset, one line of the file)
⇨ list(process identifier, struct of feature)
Reduce(): (key, value) = (process identifier, list(feature struct))
⇨ list(process identifier, struct of feature)

66.4 Test Environment

The proposed system is implemented and tested on the Big Data platform composed of 12 nodes as shown in Fig. 66.5. First, 7 systems (Intel Xeon E5620 2.4 GHz 8-core CPU, 30G RAM, 2.5T HDD) are clustered to build a Hadoop platform, and 2 systems (2.5 GHz 6-core CPU, 32G RAM, 500G HDD) are used

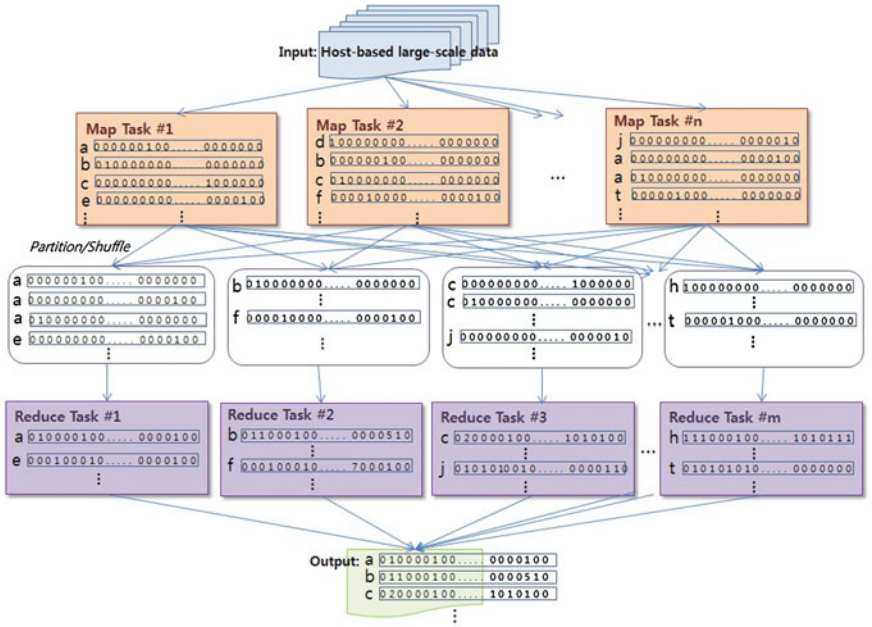


Fig. 66.4 MapReduce handling process to detect abnormal behavior in the host

➢ SINBAPT BigData Platform Rack #1 (12Nodes)

	System	Specification	Count	Storage
Real-time	MySQLCluster (IMB x3550M4)	6C(2.5GHz) 64G RAM 500G	2EA	128G
	Esper (IMB x3550M4)	6C(2.5GHz) 40G RAM 500G	2EA	40G
	Storm (APPRO1624)	8C(2.4GHz) 20G RAM 500G	2EA	40G
Large-scale	HDFS (APPRO1624)	8C(2.4GHz) 20G RAM 2.5T	7EA	17.5T

Fig. 66.5 SINBAPT big data platform specification

as MySQL cluster to save the data in real time. A total of 6 systems (excluding Hadoop master system) provide physical space (15T) to save the accumulated big data. Since Hadoop replicates data three times, the total storage space is 5T.

Currently, up to 64G real-time processing data can be saved. To extend the Big Data platform configured as described above, 19 more systems are currently being prepared. As a result, when the platform is successfully extended, the total physical space of the accumulated data will be 55T, and the total storage space of the real-time data will be 192G. A total of 7 systems used as HDFS are running on CentOS 6.4, with Hadoop 1.2.1, HBase 0.94.11, and Zookeeper 3.4.5; MySQL Cluster 7.3 version is installed as in-memory database.

This study is designed and implemented under this test environment.

66.5 Conclusion and Future Works

Like the APT attack, the cyber-targeted attack—which has a certain latent period after infiltrating the system secretly to achieve the final attack objective—attempts phased hacking using various methods over a long period of time to hide its abnormal behavior. Therefore, the attack cannot be detected without analyzing large-scale accumulated data. This paper has discussed the SINBAPT system, which detects and responds to the cyber-targeted attack in advance, and the abnormal behavior detection method using Hadoop's MapReduce. Likewise, features were defined to detect the cyber-targeted attack, and abnormal behavior was detected using MapReduce. Detection results were used for rule and data mining-based analysis according to feature profiling. The correlation between the result and real-time data analysis result was also analyzed to detect the final cyber-targeted attack more accurately. Currently, analysis engines are in the implementation phase after passing through the design phase. We will prove that the abnormal behavior detection method using MapReduce is effective by measuring the data processing speed and throughput of the large-scale data analysis engine through testing.

Acknowledgments This research was funded by the MSIP (Ministry of Science, ICT and Future Planning), Korea in the ICT R&D Program 2013 [Cyber-targeted attack recognition and traceback technology based on the long-term historic analysis of multi-source data].

References

1. Kim J, Lim S, Kim I, Cho H, Noh B (2013) Trend of cyber security technology using big data. Electronic communication trend analysis (3rd edn.), vol 28
2. R&D Support of DARPA Cyber Genome Program (2010) General dynamics. http://publicintelligence.net/hbgary-general-dynamics-darpa-cyber-genome-program-proposal/
3. Wikipidia Cyber genome project. http://wiki.echelon2.org/wiki/Cyber_Genome_Project
4. IBM. http://www-01.ibm.com/software/tivoli/products/security-operations-mgr/
5. Splunk. http://www.splunk.com/

6. Lee Y, Lee Y (2013) Toward scalable Internet traffic measurement and analysis with Hadoop. ACM SIGCOMM Comput Commun Rev 43(1):6–13
7. Choi D, Moon G, Kim Y, Noh B (2011) Large quantity of security log analysis using MapReduce. J Korean Inst Inf Technol (8th edn.) vol 9
8. Dean J, Ghemawat S (2008) MapReduce: simplified data processing on large clusters. Commun ACM 51(1):107–113

Chapter 67
Rotational Bridge Priority Masking for Building Equal-Cost Multi-tree

Changjin Suh and Uijin Jang

Abstract The routing in ethernet backbone networks chooses large number of shortest paths on purpose and establishes multiple shortest path trees for the better load balancing and faster protection by using the bridge priority masking method. This simple and nice method, however, does not permit bit position change, resulting in bitwise unfairness in the choice of paths. This paper raises an open mathematical bridge priority masking problem that allows any kind of bit permutation, and gives two solutions under our proposed 8-way rotation with inversion, and proves that only two solutions exist. According to the performance evaluation using total path inspection, a proposed rotational solution enhances the fairness in choosing paths by 8–43 % than the current one.

Keywords Equal-cost multi-tree (ECMT) · Bridge priority masking · Rotational bridge priority masking · Priority matrix · Shortest path bridging (SPB)

67.1 Introduction

The equal-cost multi-path technology or shortly ECMP can utilize multiple shortest paths originally developed for IP/MPLS networks [1, 2] for load balancing and fast protection. ECMP provides the extended link-state routing's per-hop decision to choose many next hop routers.

C. Suh
School of Computer Science and Engineering, Soongsil University, Seoul, Korea
e-mail: cjsuh@ssu.ac.kr

U. Jang (✉)
Korea Copyright Commission, Soongsil University, Seoul, Korea
e-mail: neon7624@gmail.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_67, © Springer Science+Business Media Dordrecht 2014

ECMP has been inherited to Ethernet provider backbone networks which are empowered by the link state control plan protocol—the "Provider Link State Bridging" [3], in which network multiple shortest paths can be used from multiple trees.

The 802.1ad allows 16 different shortest path trees for each bridge. The term 'equal-cost multi-tree' (ECMT) [4] is used in Ethernet instead of ECMP in IP networks. To support unicast as well as multicast, ECMT requires the path symmetry indicating that a backward path is the one reversed to the forward path. The path symmetry is a strict requirement that nullifies most of IP/MPLS routings and some Ethernet routing [5]. So far, no research has been tried to basically enhance ECMT. The research in [6] uses more careful selection rule to have extra tie-break with additional link weight. But the benefit requires manual jobs to network operators, and does not go well with ECMT.

This paper tries the first research to modify the bridge priority masking, which is a key idea for establishing ECMT. This paper discusses over the bit priority masking that provides sixteen priority rules in choosing 16 equal-cost multi-paths mentioned in the standard 802.1aq Shortest Path Bridging [7] used in ethernet backbone networks and ethernet provider backbone networks.

This paper consists as follows. Section 67.2 mentions a mathematical model and proposes new terms such as the priority matrices and the priority words. Section 67.3 defines a new research problem and Sect. 67.4 gives a proof and a solution. Section 67.5 evaluates the performance, and the conclusion follows.

67.2 Modeling the Standard and the Priority Matrix P_S

The standard 'Shortest Path Bridging' selects the best path according to the bridge priority masking in 16 different ways. Our proposed 16 × 16 'priority matrix' can store all these information. We transform this 16 ways of bridge priority masking in the current standard to the 16 × 16 priority matrix PS as shown in Table 67.1a. A column in PS is generated from an independent bridge priority masking. 16 columns are rearranged to make diagonal elements all 0 in PS. A lower index in a column includes a preferred bridge Id. Each column is used only to choose the best bridge and the best path.

Suppose there are two equal-cost shortest paths T_h1 and T_h2, whose path Ids are 1 and 2 consisting of two intermediate bridges whose Ids are {5, 6} and {7, 8} respectively. Table 67.2 gives the result of path selection in the first two rows after the heading. The upper row shows the selected bridge among 5, 6, 7 and 8 and the low row lists the selected path between 1 and 2. The selection varies for every c'th column or the c'th priority rule. In the rightmost column, the total selection counts are listed in the form of (S_{N5}, S_{N6}, S_{N7}, S_{N8}) or (S_{M1}, S_{M2}), where S_{Nk} represents the number of bridge selection for Nk (k = 5, 6, 7, 8) and S_{Mj} indicates the number of path selection for T_hj (j = 1, 2). The later rows in Table 67.2

Table 67.1 Priority matrices. **a** Standard p_s. **b** Proposed p_x.**c** Proposed p_p

(a)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	0	3	2	5	4	7	6	9	8	11	10	13	12	15	14
2	2	3	0	1	6	7	4	5	10	11	8	9	14	15	12	13
3	3	2	1	0	7	6	5	4	11	10	9	8	15	14	13	12
4	4	5	6	7	0	1	2	3	12	13	14	15	8	9	10	11
5	5	4	7	6	1	0	3	2	13	12	15	14	9	8	11	10
6	6	7	4	5	2	3	0	1	14	15	12	13	10	11	8	9
7	7	6	5	4	3	2	1	0	15	14	13	12	11	10	9	8
8	8	9	10	11	12	13	14	15	0	1	2	3	4	5	6	7
9	9	8	11	10	13	12	15	14	1	0	3	2	5	4	7	6
10	10	11	8	9	14	15	12	13	2	3	0	1	6	7	4	5
11	11	10	9	8	15	14	13	12	3	2	1	0	7	6	5	4
12	12	13	14	15	8	9	10	11	4	5	6	7	0	1	2	3
13	13	12	15	14	9	8	11	10	5	4	7	6	1	0	3	2
14	14	15	12	13	10	11	8	9	6	7	4	5	2	3	0	1
15	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1	0

(b)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	8	2	12	4	10	9	14	1	6	5	11	3	13	7	15
1	1	0	6	8	5	2	11	12	3	4	13	10	7	9	15	14
2	2	9	0	4	12	14	8	10	5	7	1	3	11	15	6	13
3	3	1	4	0	13	6	10	8	7	5	9	2	15	11	14	12
4	4	10	3	13	0	8	1	6	9	14	7	15	2	12	5	11
5	5	2	7	9	1	0	3	4	11	12	15	14	6	8	13	10
6	6	11	1	5	8	12	0	2	13	15	3	7	10	14	4	9
7	7	3	5	1	9	4	2	0	15	13	11	6	14	10	12	8
8	8	12	10	14	6	11	13	15	0	2	4	9	1	5	3	7
9	9	4	14	10	7	3	15	13	2	0	12	8	5	1	11	6
10	10	13	8	6	14	15	12	11	4	3	0	1	9	7	2	5
11	11	5	12	2	15	7	14	9	6	1	8	0	13	3	10	4
12	12	14	11	15	2	9	5	7	8	10	6	13	0	4	1	3
13	13	6	15	11	3	1	7	5	10	8	14	12	4	0	9	2
14	14	15	9	7	10	13	4	3	12	11	2	5	8	6	0	1
15	15	7	13	3	11	5	6	1	14	9	10	4	12	2	8	0

(c)

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	0	8	4	12	2	5	9	14	1	6	10	13	3	11	7	15
1	1	0	6	8	10	4	13	12	3	2	11	5	7	9	15	14
2	2	9	0	4	3	7	1	10	5	14	8	12	11	15	6	13
3	3	1	2	0	11	6	5	8	7	10	9	4	15	13	14	12
4	4	10	12	13	0	1	8	6	9	7	14	15	2	3	5	11
5	5	2	14	9	8	0	12	4	11	3	15	7	6	1	13	10
6	6	11	8	5	1	3	0	2	13	15	12	14	10	7	4	9
7	7	3	10	1	9	2	4	0	15	11	13	6	14	5	12	8
8	8	12	5	14	6	13	11	15	0	4	2	9	1	10	3	7
9	9	4	7	10	14	12	15	13	2	0	3	1	5	8	11	6
10	10	13	1	6	7	15	3	11	4	12	0	8	9	14	2	5
11	11	5	3	2	15	14	7	9	6	8	1	0	13	12	10	4
12	12	14	13	15	4	9	10	7	8	5	6	11	0	2	1	3
13	13	6	15	11	12	8	14	5	10	1	7	3	4	0	9	2
14	14	15	9	7	5	11	2	3	12	13	4	10	8	6	0	1
15	15	7	11	3	13	10	6	1	14	9	5	2	12	4	8	0

additionally include the number of selections under a same path condition but with two proposed priority matrices P_X and P_P defined in Table 67.1b and c respectively.

A priority matrix P_S can be simply defined by 16 4-bit vectors. We represent the 4 bit-vector at the k'th (k = 0, 1, 2, …, 15) column as $w_{k,}$ $x_{k,}$ $y_{k,}$ z_k. A bit-vector can be selected among 4 constant bit-vectors in (67.1a) or their 4 masked ones in (67.1b) (Table 67.3).

Table 67.2 The chosen bridge Id/path Id at all columns with given two best paths $T_h\,1 = \{5, 6\}$, $T_h\,2 = \{7, 8\}$

rule	selection	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	total
P_S	node Id	5	5	6	7	5	5	6	7	8	8	8	8	8	8	8	8	(4,2,2,8)
	path Id	1	1	1	2	1	1	1	2	2	2	2	2	2	2	2	2	(6,10)
P_X	node Id	5	8	6	8	5	6	8	8	5	6	5	7	7	8	7	8	(4,3,3,6)
	path Id	1	2	1	2	1	1	2	2	1	1	1	2	2	2	2	2	(7,9)
P_P	node Id	5	8	6	8	8	5	5	8	5	6	8	5	7	7	7	8	(5,2,3,6)
	path Id	1	2	1	2	2	1	1	2	1	1	2	1	2	2	2	2	(7,9)

Table 67.3 Priority matrices expressed with 4-bit priority words

column / rules	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
P_S	abcd	abcd	abcd	abcd	abcd	abcd	abcd	abcd	abcd	abcd	abcd	abcd	abcd	abcd	abcd	abcd
P_X	abcd	dabc	adcb	cdab	cbad	dcba	badc	bcda	bcda	badc	dcba	cbad	cdab	adcb	dabc	abcd
P_P	abcd	dabc	bcda	cdab	dabc	abcd	cdab	bcda	bcda	cdab	abcd	dabc	cdab	bcda	dabc	abcd

$$\begin{cases} a = [0000000011111111]^T, \\ b = [0000111100001111]^T, \\ c = [0011001100110011]^T, \\ d = [0101010101010101]^T. \end{cases} \tag{67.1a}$$

$$\begin{cases} \underline{a} = [1111111100000000]^T, \\ \underline{b} = [1111000011110000]^T, \\ \underline{c} = [1100110011001100]^T, \\ \underline{d} = [1010101010101010]^T. \end{cases} \tag{67.1b}$$

The superscirpt ^{T} in (67.1a, 67.1b) is an operator which transposes a 1×16 matrix to be 16×1 in dimension. The masked bit vector $\underline{v}$ is a bit vector which reverses 16 elements in v from t to $(1 - t)$ for $t \in \{0, 1\}$. In this paper, $\underline{vr}$ is equal to $\underline{v}\,\underline{r}$. Also this paper separately uses a, b, c and d as constant bit-vectors and w, x, y and z as variables. A 4-bit priority vector is also called a priority word vector or a priority word.

16 kinds of 4-bit vectors w_k, x_k, y_k, z_k in the standard priority matrix are determined in the boundary that satisfies (67.2).

$$w_k \in \{a, \underline{a}\},\ x_k \in \{b, \underline{b}\},\ y_k \in \{c, \underline{c}\},\ z_k \in \{d, \underline{d}\} \quad \text{for} \quad k = 0, 1, \ldots, 15 \tag{67.2}$$

67.3 Definition of Problems

The objective of this paper is to find statistically fairer priority matrices. We require four rules in defining a priority matrix or 16 priority words to achieve our goal.

- (R1) A column includes each of 0, 1, 2, … and 15 exactly once.
- (R2) A row includes each of 0, 1, 2, … and 15 exactly once.
- (R3) Priority words allow any bit-vector permutation, if R1 is satisfied.
- (R4) A fair rotation rule has to be used and it must produce all 16 priority words.

R1 is a basic requirement for a selection rule. R2 is useful in filtering out low-quality matrices. R3 and R4 are rotation rules. P_S satisfies R1 and R2 but not R3 nor R4. The rule R3 can be mathematically described as

$$\begin{aligned} &\text{for } k = 0, 1, \ldots, 15 \\ &\begin{cases} \{w_k,\ x_k,\ y_k,\ z_k\} \subset \{a, \underline{a},\ b, \underline{b},\ c, \underline{c},\ d, \underline{d}\}, \\ \{a,\ b,\ c,\ d\} \subset \left\{w_k, \underline{w_k},\ x_k, \underline{x_k},\ y_k, \underline{y_k},\ z_k, \underline{z_k}\right\}. \end{cases} \end{aligned} \tag{67.3}$$

Due to (67.3) any two bit-vectors in $\{w_k, x_k, y_k, z_k\}$ cannot be (a, a), (a, $\underline{a}$), ($\underline{a}$, $\underline{a}$), (b, b), … or ($\underline{d}$, $\underline{d}$). The non-rotational bit assignment constraint in (67.2) is relaxed to (67.3) in the rotational assignment.

67.4 Cyclic Priority Matrix P_P and P_X

This section searches and defines two cyclic priority matrices P_P and P_X that satisfy R1, R2, R3 and R4 in Sect. 67.3. As an answer to R4, we propose the rotation with inversion (R_I) that transforms a priority word wxyz into xyz$\underline{w}$ or $\underline{z}$wxy in Fig. 67.1.

In both ways, R_I generates 8 different priority words with 7 consecutive rotations and relocates to the original place after 8 rotations. For simplicity we list words rotating bits to the right, which is the counterclockwise rotation in Fig. 67.1. R_I also keeps R1 and R3. The only unsolved requirement is R2. R2 is proved with the 'rotation group.' The rotation group G_R(wxyz) is defined as a set,

$$G_R(\text{wxyz}) = \{\text{wxyz},\ \text{xyz}\underline{\text{w}},\ \text{yz}\underline{\text{wx}},\ \underline{\text{zwxy}},\ \underline{\text{wxyz}},\ \underline{\text{xyzw}},\ \underline{\text{yzwx}},\ \underline{\text{zwxy}}\}. \tag{67.4}$$

The set G_R(wxyz) includes 8 priority words generated by the rotation with inversion R_I starting with wxyz. Note both a word and its masked one belong to a same rotation group.

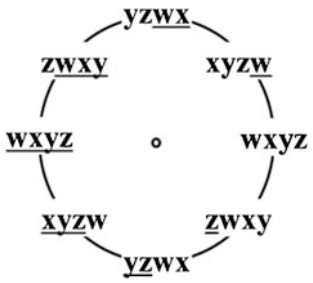


Fig. 67.1 $G_R(wxyz)$ using the rotation with inversion R_I

Generate two G_R's with two initial words 0000 and 0010, as shown in (67.5a) and (67.5b), and call them O and I respectively. O and I are disjoint and union of them covers all 4-bit numbers.

$$\begin{aligned} O &= \{0000_2,\ 0001_2,\ 0011_2,\ 0111_2,\ 1111_2,\ 1110_2,\ 1100_2,\ 1000_2\} \\ &= \{0,\ 1,\ 3,\ 7,\ 15,\ 14,\ 12,\ 8\}, \end{aligned} \tag{67.5a}$$

$$\begin{aligned} I &= \{0010_2,\ 0101_2,\ 1011_2,\ 0110_2,\ 1101_2,\ 1010_2,\ 0100_2,\ 1001_2\} \\ &= \{2,\ 5,\ 11,\ 6,\ 13,\ 10,\ 4,\ 9\}. \end{aligned} \tag{67.5b}$$

It is useful to represent a priority word just using I's and O's, called IO representation. For example the priority word abcd is represented as [OOIO IIIO OIII OIOO]^T. It is important to learn that any word in a G_R has a same IO representation. So IO representation will be used as a unit of G_R. For example,

$$G_R(abcd) = [OOIO\ IIIO\ OIII\ OIOO]^T. \tag{67.6}$$

Another special feature of G_R is that 8 priority words in a G_R keep R2, which is simply proved as next. If we collect every j'th ($0 \le j \le 15$) member from 8 priority words in a G_R to a set, the set will be (67.5a) or (67.5b). Both cases store all distinct numbers.

To fill in a whole priority matrix we need two G_R's. Without loss of generality, the first G_R is $G_R(abcd)$. Let us determine the second G_R–$G_R(wxyz)$ that keep R2 with $G_R(abcd)$. To be the solution, $G_R(wxyz)$ should have the IO representation opposite to $G_R(abcd)$ in (67.6). So

$$G_R(wxyz) = [IIOI\ OOOI\ IOOO\ IOII]^T. \tag{67.7}$$

Define the second word group—the masked word group $G_C(wxyz)$, which is a set of masked priority words from wxyz, by masking any subset of bits such as {wxyz, wxy$\underline{z}$, wx$\underline{y}$z, wx$\underline{yz}$, w$\underline{x}$yz, …, $\underline{wx}$yz, $\underline{wxyz}$}. For example, $G_C(abcd)$ is P_S.

We are going to find all solutions of wxyz that satisfies (67.7) in the search list $G_C(W_{ord})$ for every non-masked word W_{ord}. Because bits rotate, we can set the first bit w arbitrarily. Fixing w = a, we have different 3! ways. The set of every non-masked W_{ord} is {abcd, abdc, acbd, acdb, adbc, adcb}. For any axyz in this set, probe $G_C(axyz)$ whether some a′x′y′z′ in $G_C(axyz)$ satisfies (67.7). As a result of

total search, we found that there exist only two solutions that satisfy (67.7) and they are $G_R(a\underline{b}c\underline{d})$ and $G_R(\underline{d}c\underline{b}a)$. The discussion so far is abstracted in the next proposition.

Proposition *Only two G_R pairs satisfy the requirements R2 and R3, if the first $G_R(abcd)$ is pre-assigned. They are $\{G_R(abcd), G_R(a\underline{b}c\underline{d})\}$ and $\{G_R(abcd), G_R(\underline{d}c\underline{b}a)\}$. The priority words in three G_R's are listed below.*

$$G_R(abcd) = \{abcd,\ bcd\underline{a},\ cd\underline{ab},\ d\underline{abc},\ \underline{abcd},\ \underline{bcd}a,\ \underline{cd}ab,\ \underline{d}abc\},$$
$$G_R(abcd) = \{a\underline{bcd},\ \underline{bcd}a,\ c\underline{dab},\ \underline{dabc},\ \underline{abcd},\ b\underline{cda},\ \underline{cdab},\ da\underline{bc}\},$$
$$G_R(dcba) = \{\underline{dcb}a,\ \underline{adc}b,\ b\underline{adc},\ \underline{cbad},\ d\underline{cba},\ ad\underline{cb},\ \underline{badc},\ c\underline{bad}\}.$$

With collecting 16 words from $\{G_R(abcd), G_R(a\underline{b}c\underline{d})\}$ and from $\{G_R(abcd), G_R(\underline{d}c\underline{b}a)\}$, P_P and P_X are established as shown in Fig. 67.1b and c respectively.

67.5 Fairness Evaluation

This section discusses over how much fairly the priority matrices P_S, P_X and P_P choose paths among T_h N_d-bridge paths. To focus into priority matrices, we assume 16 bridges have the same Id but different in the leading 4 bits. This is a same assumption that 16 bridges have all different 4-bit Id's. Test types are classified by the pair (T_h, N_d), which represents that a selection test is repeatedly for T_h paths, each of which consists of N_d distinct intermediate bridges whose Id are chosen among 16 4-bit different binary numbers. The standard deviation is used as an attribute to measure the unfairness of path selection.

A test type (T_h, N_d) consists of $16!/(T_h! \cdot (N_d!)^{Th} \cdot (16\text{-}T_h \cdot N_d)!)$ test units. At the i'th test unit c_i, 16 selection is done among T_h paths based on 16 column rules in P_q with $P_q \in \{P_S, P_X, P_P\}$. The path having the highest priority bridge Id is a winner. Let $h_j[c_i]$ the number of selections for the j'th ($j = 1, 2, \ldots, T_h$) path in the test unit c_i. The vector $H[c_i]$ defined by $(h_1[c_i], h_2[c_i], \ldots, h_{Th}[c_i])$ stores the number of selections in the c_i'th test unit. A standard deviation σ_i is calculated from $H[c_i]$ using (67.8) during a test unit, and is averaged to be in σ throughout C test units during a test type according to (67.9).

$$\sigma_i = \sqrt{\sum_{j=1,T_h} \left(h_j[c_i]\right)^2 - \left(\frac{16}{N_d}\right)^2}, \tag{67.8}$$

$$\sigma = \frac{1}{C} \cdot \sum_{i=1,C} \sigma_i,\ C = \frac{16!}{T_h! \cdot (N_d!)^{T_h} (16 - T_h \cdot N_d)!}. \tag{67.9}$$

Table 67.4 includes a pair of results $[\sigma_q, E_q]$ under a test type (T_h, N_d) and three priority matrices. The left parameter σ_q is termed the unfairness with a suffix q

Table 67.4 Unfairness and relative enhancement [σ_q, E_q] in the path selection test among T_h N_d-bridge paths (T_h, N_d)

results \ (T_h,N_d)	(3,1)	(4,1)	(6,1)	(8,1)	(4,2)	(2,4)	(3,2)	(3,3)
[σ_S, E_S]	[1.882,1.000]	[1.779,1.000]	[1.299,1.000]	[0.956,1.000]	[1.251,1.000]	[1.445,1.000]	[1.743,1.000]	[1.261,1.000]
[σ_X, E_X]	[1.452,1.296]	[1.386,1.284]	[1.093,1.188]	[0.845,1.131]	[1.106,1.131]	[1.278,1.131]	[1.466,1.189]	[1.120,1.126]
[σ_P, E_P]	[1.314,1.432]	[1.261,1.411]	[1.052,1.235]	[0.884,1.081]	[1.110,1.127]	[1.281,1.128]	[1.411,1.235]	[1.133,1.113]

from P_q. The standard deviation σ_q gets larger if selection is more biased and shrinks to 0 if all paths are selected with identical times. σ_q has little meaning alone but is good for comparison. The relative fairness enhancement E_q defined by σ_S/σ_q ($q \in \{S, X, P\}$) is attached in the cell. The best and worst E_T and E_P are underlined in Table 67.4. From Table 67.4, we deduct the followings properties.

- $\sigma_S \geq \sigma_X$, $\sigma_S \geq \sigma_P$: This means the two proposed P_P and P_X is always better than P_S.
- generally $\sigma_P < \sigma_X$: The case of $\sigma_X < \sigma_P$ still occurs but their difference is all negligible. So we can say P_P is better than P_X.
- $1.12 \leq E_X \leq 1.30$: P_X enhances the fairness compared with P_S by 12–30 %.
- $1.08 \leq E_P \leq 1.43$: P_P enhances the fairness compared with P_S by 8–43 %.

67.6 Conclusion

This Paper establishes the mathematical modeling of the current bridge priority masking that generates the equal-cost multi-trees and new mathematical terms such as priority matrix, 4-bit vectors and priority words, and raises an open problem how to relocate bits in harmony to upgrade the fairness in path selection

This Paper proposes a new bridge priority rotational masking based on our suggested 'rotation with inversion' rule. Utilizing some mathematical features in the rotation with inversion and proposed terminologies, we proved only two solution sets exist. The better priority matrix mentioned as P_P in our paper shows 8–43 % fairness enhancement in choosing paths compared with the current standard throughout all possible path cases. This enhancement is easily achieved by adding bit-shifting operation of bridge Id's to the current bridge priority masking hardware.

Acknowledgments This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the Convergence-ITRC (Convergence Information Technology Research Center) support program (NIPA-2013-H0401-13-1004) supervised by the NIPA.

References

1. IETF Std. rfc2992 (2000) Analysis of an equal-cost multi-path algorithm. http://tools.ietf.org/html/rfc2992
2. Mérindol P, Pansiot JJ Cateloin S (2009) Low complexity link state multipath routing. In: INFOCOM workshops 2009, pp 19–25
3. Allan David, Ashwood-Smith Peter, Bragg Nigel, Fedyk Don (2008) Provider link state bridging. IEEE Commun Mag 46(9):110–117
4. Allan David et al (2010) Shortest path bridging: efficient control of larger ethernet networks. IEEE Commun Mag 48(10):128–135
5. Carral JA et al (2010) TRE+: extended tree-based routing ethernet. ETRI J 32(1):157–159
6. Luo Z, Suh CJ (2011) An improved shortest path bridging protocol for ethernet backbone network. In: International conference on information networking (ICOIN), pp 148–153
7. IEEE Std. 802.1aq (2012) Shortest path bridging, IEEE 802, Feb 2012. http://www.ieee802.org/1/pages/802.1aq.html

Chapter 68
Linear SVM-Based Android Malware Detection

Hyo-Sik Ham, Hwan-Hee Kim, Myung-Sup Kim and Mi-Jung Choi

Abstract Important personal user information has become scattered in devices as mobile devices are now supporting various services and contents. Accordingly, attackers are expanding the scope of their attack not only in the existing PC and Internet environment but also to mobile devices. In this paper, we monitor the resource information of mobile devices to detect Android malware. Using the monitored information, we propose a method of detecting malware by applying linear SVM (support vector machine) that shows high classification performance in machine learning classifiers in order to automatically detect malware. The validity of the proposed methodology is verified through experiment results.

Keywords Android · Malware · Machine learning · SVM (support vector machine)

68.1 Introduction

Smartphone is accelerating the advancement of mobile industry. The number of smartphone users is also increasing exponentially. Modern people are using various services via smartphone such as Web search and SNS (social network service) and mobile banking service anytime anywhere. This is also changing the life

H.-S. Ham · H.-H. Kim · M.-J. Choi (✉)
Department Computer Science, Kangwon National University, Chuncheon, South Korea
e-mail: mjchoi@kangwon.ac.kr

H.-S. Ham
e-mail: manhhs@kangwon.ac.kr

H.-H. Kim
e-mail: hwanhee0920@kangwon.ac.kr

M.-S. Kim
Department Computer and Information Science, Korea University, Seoul, South Korea
e-mail: tmskim@korea.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_68, © Springer Science+Business Media Dordrecht 2014

pattern of modern people and industry paradigm. Although the advancement of smartphone has brought convenience and usefulness to modern people, a significant threat of security vulnerability is lurking on the other side. Since various types of important user information such as user location information, contacts and certificate are scattered in smartphone, there is a possibility of receiving serious damages from hacker's threatening attack. Currently, hackers are expanding their attack destination from existing PC environment to smartphone.

Figure 68.1 is a report by a Finnish security company F-secure that states that among 301 mobile malware samples that occurred in 2012, 238 malwares targeted Android platform [1]. While the number of malwares that target other mobile platforms is gradually decreasing as time went on quarter to 4th quarter, Android is showing a contrasting result. The reason for the increase in Android malware is because of its open source policy and Android platform's characteristic that does not make app verification of market difficult. In addition, it is because distributing malware in the market through the repackaging method of inserting it in normal application is easily performed.

In previous studies, there were various approaches to detect mobile malware such as signature-based detection) [2–5], behavior-based detection [6–10] and taint analysis-based detection) [11, 12]. This paper identifies the issues of previous studies and proposes a detection method through linear SVM (support vector machine) [13] that shows high performance among machine learning algorithms in order to effectively detect malware in Android platform with monitored resources during application running time.

68.2 Related Work

This section examines the trend of previous studies and explains the linear SVM method to detect mobile malware.

68.2.1 Mobile Malware Detection Trend

To detect abnormal behaviors occurring in existing mobile environment (malware, virus, worm, etc.), signature-based detection, behavior-based detection and taint analysis-based detection were used.

Signature-based detection [2–5] is a traditional method used to detect malware in PC environment. To define signature, static method and dynamic method are simultaneously used. Static analysis is a method of analyzing without actually starting program and it targets the source code and object code. It decompiles the source code of malware to discover vulnerabilities that occur in command, statement, etc. and defines them as signatures. Dynamic analysis is a method of finding certain patterns in memory leakage, traffic flow and data flow while actually running program, and defines them as signatures.

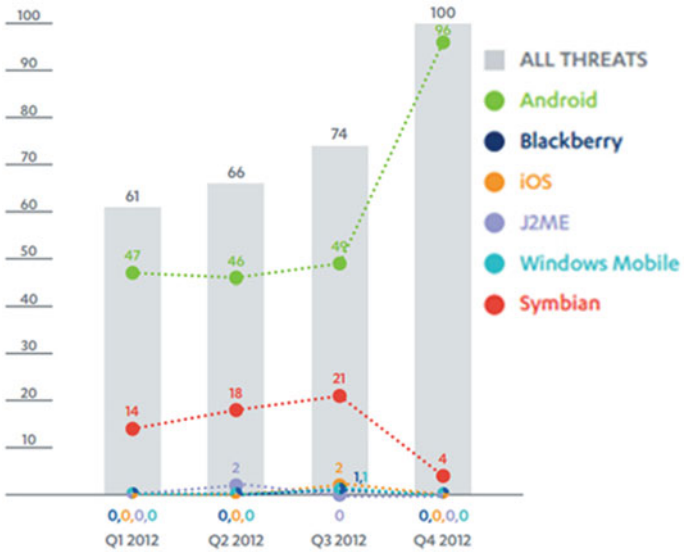


Fig. 68.1 Increase of android malware

Behavior-based detection [6–10] is a method of detecting invasion status by comparatively analyzing predetermined attack pattern and process behavior that occurs in a system. It is one of the studies that are receiving most attention in recent due to signature-based detection's limited detection of malicious behavior. It mainly monitors event information that occurs while using smartphone such as memory usage amount, SMS, battery consumption amount, etc. to detect abnormal pattern. Host-based detection for directly monitoring information inside device and network-based detection for gathering information via network are frequently used. Since host-based detection increases the usage of smartphone battery and memory, a detection method of collecting data inside device and transmitting data to outside analysis server is mainly used.

Dynamic analysis-based detection [11, 12] also called 'taint analysis' is a method of marking specific data and monitoring the process of data being sent in an application code to track the flow of data. Since a smartphone runs in a virtual machine, this method is considered as an appropriate one. However, its study is no longer being conducted currently due to the difficulty in applying it in actual environment because of the overhead to tracking data flow to a low level.

68.2.2 Malware Detection via SVM

In this paper, malware is determined based on the collected data by monitoring resource in Android environment. The behavior-based detection involves the inconvenience of having to determined malware infection status by examining data

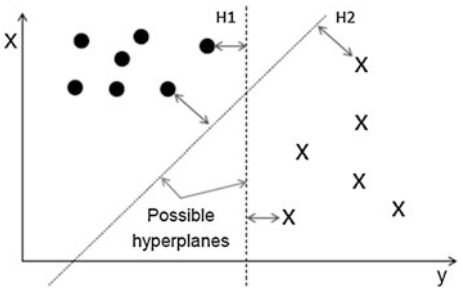


Fig. 68.2 Data classification using linear SVM

that consists of numerous features. Accordingly, the behavior-based detection uses machine learning method for automated malware classification and to ensure its identification and accuracy. Machine learning method is a method of entering the data collected from device as learning data to create a learning model and applying some of the other data to the learning model.

In previous studies [14, 15], there was an approach of detecting malware through RF (RandomForest) machine learning algorithm. RF is a combination of decision trees formed by random vectors independently sampled and it showed relatively high detection rate. In this paper, a linear SVM method is applied to detect malware. SVM is one of machine learning classifiers that are receiving most attention in recent and its various modifications and applications are occurring because of its high performance. In addition, SVM can solve the issue of nonlinear data classification and effective performance can be expected because it includes the function of filtering unnecessary features among the entered features.

Figure 68.2 is a method of coming up with hyperplanes that are the standards for classifying data by linear SVM through a learning process. The biggest advantage of the SVM is in maximizing the generalization capability by maximizing margin that is the distance among data. Accordingly, the classification performance of SVM is further maximized than that of other machine learning classifiers when data is positioned at ambiguous border. The purpose of this paper is to verify the superiority of our proposed method by comparing experiment results of SVM analysis method and RF applied in previous studies.

68.3 Collection of Resource Information for Malware Detection

In this section, a method of collecting resource information is presented to detect Android malware. It explains collected resource features and agents designed and implemented to collect resource information inside Android devices.

Table 68.1 Selected feature for malware detection

Resource type	Resource feature
Network	RxBytes, TxBytes, RxPacket, TxPacket
Telephone	Send/receive call
SMS message	Send/receive sms
CPU	CPU usage
Battery	Level, temperature, voltage
Process	Process ID, process name, running process, context switches
Memory	Native Total size, shared size, allocated size, physical page, virtual set size, free size, heap size, dirty page
	Dalvik Total size, shared size, allocated size, physical page, virtual set size, free size, heap size, dirty page

68.3.1 Resource Feature for Malware Detection

For detecting malware that is the target of analysis, resource information generated in a device is monitored when a user executes abnormal applications inserted with malware and normal applications. In previous study [10], every resource and events generated in an Android device were defined and these all features are used for analyzing malware. However, the number of features is too many with little correlation for the majority. It is not reflecting Android memory structure, either [10].

In this paper, 32 features that are highly related with targeted malware, as shown in Table 68.1, were defined by classifying them into seven categories according to resource types. Particularly, it does not simply monitor the total memory usage amount that simply changes through an application execution, but monitors memory usage amount classifying into native area and Dalvik Machine area by considering the memory characteristics of Android platform. The Dalvic machine memory is allocated when running each application.

68.3.2 Resource Information Monitoring Agent

To monitor the selected resource features, an agent is needed that can continuously monitor corresponding features inside a device. Figure 68.3 is the structure of agent for monitoring resource information.

Linux kernel manages OS and hardware at the very bottom of Android platform and it includes according d monitored data to device drivers. In addition, it becomes the starting point of monitoring resource information. Each collector independently gathers resource for each type in real-time and transmits the monitored data to a data management module. Resource information can be obtained through API calls and Linux commands. The data management module converts each data into vector form in order to apply collected resource data to a machine learning model and transmits to an external analysis server through a communication module. The reason for

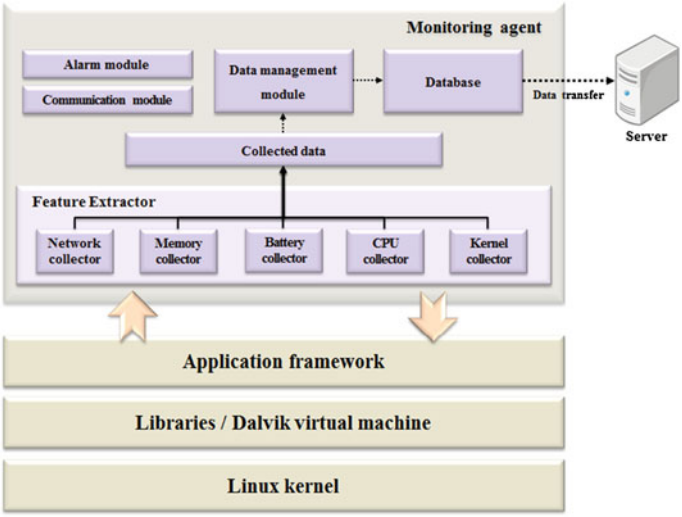


Fig. 68.3 Architecture of resource information monitoring agent

transmitting data to the external server for analysis is to reduce performance overhead that occurs to detect malware in a mobile device. The analysis server notifies to user the detection status of malware determined through a machine learning classifier via an alarm module. Lastly, in a user level, monitoring status and malware detection status are visualized for display.

68.4 Experimental Results

In this section, it compares the detection result of the RF classifier with that of the propose linear SVM classifier to verify malware detection status. The purpose is to verify the effectiveness of the malware detection of the SVM classifier through comparison.

68.4.1 Experiment Method

To detect malware, 14 normal applications and 14 abnormal applications inserted with malware are composed as data set. In addition, two types of experiment environment are composed to verify its applicability in actual mobile environment.

The first experiment environment is the method of using 80 % of data collected from 5 Android devices as learning data to create model through learning process and evaluating performance by entering 20 % of data as test data based on the

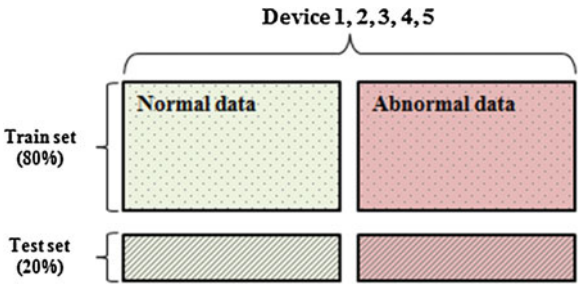


Fig. 68.4 Method of collecting data from same device (Experiment 1)

created model. As for the objective of this experiment, it is a general method of evaluating if each classifier can detect malware in the case of composing learning data and test data in the same device, and the accuracy of resource features collected can be verified through the result of Experiment 1. Figure 68.4 shows the method of Experiment 1.

The second experiment is the method of differentiating the train set and test set. It applies 5-folds cross validation evaluation for data collected from each device. Performance of the classifiers can be determined through cross validation of data collected from different devices for evaluation. As for the objective of this experiment, it is a method of composing learning data and test data with data collected from different devices and evaluating the performance of classifier by separating data set. Accordingly, the Experiment 2 is a method that considers the transplantation among devices (Fig. 68.5).

68.4.2 Experiment Results

Figure 68.6 is a graph that compared the performances of each classifier in evaluation index when collected data were divided into normal/abnormal applications. Although the two classifiers all show high performances, the RF classifier showed better performance from the standpoint of TPR (True Positive Rate) and Accuracy. However, the case of Experiment 1 underwent learning/test process based on data collected from the same device. Since learning set and test set need to be classified to detect malware, this is not a suitable experiment method in actual environment of use. Since the SVM classifier also showed better performance than RF classifier from the standpoint of FPR (False Positive Rate) and Precision evaluation index, the two classifiers can be said as effective in detecting malware.

Figure 68.7 is a graph that compares evaluation index of each classifier when using data collected from different devices as learning data and test data. In Experiment 1, RF showed high numerical value of above 0.99 for TPR/Precision/Accuracy but it showed low detection performance except for TPR in Experiment 2. In addition, FPR that works as important performance index when detecting malware also showed relatively high numerical value of 0.019 thereby showing not

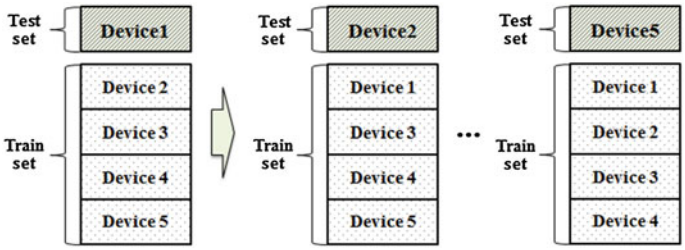


Fig. 68.5 Method of collecting data from different devices (Experiment 2)

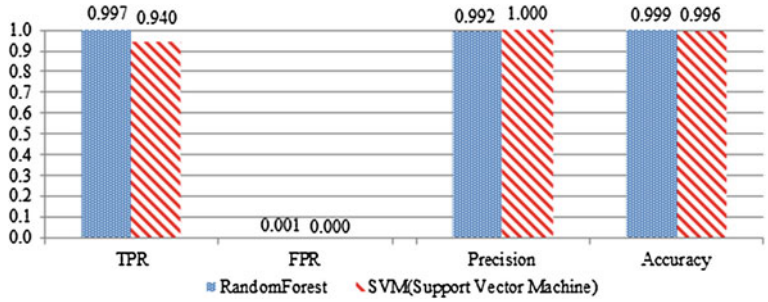


Fig. 68.6 Performance comparison of classifiers (Experiment 1)

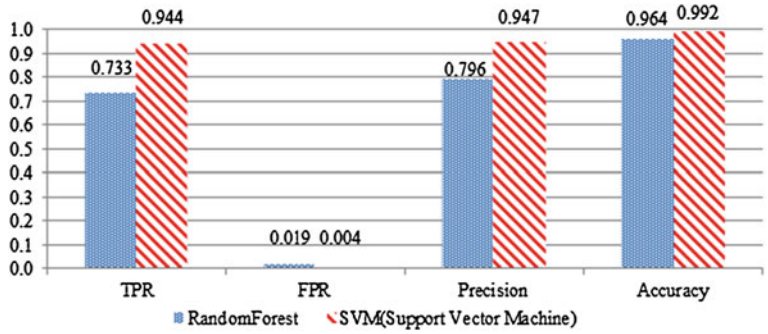


Fig. 68.7 Performance comparison of classifiers (Experiment 2)

good detection result. In the case of SVM, however, it maintained high performance even when experiment environment changed by showing the similar result of Experiment 1. Accordingly, the linear SVM classifier not influenced by environment can be considered as being more appropriate in detecting Android malware.

Table 68.2 shows the detailed performance index of Experiment 2. When RF and SVM classifiers were compared, the SVM showed about 21 % of improvement rate in TPR while showing about 1.5 % of reduction rate in FPR. In the case of

Table 68.2 Detailed performance indicator (Experiment 2)

Data type	TPR		FPR		Precision		Accuracy	
	RandomForest	SVM	RandomForest	SVM	RandomForest	SVM	RandomForest	SVM
Normal	0.998	0.989	**0.113**	**0.004**	**0.427**	**0.959**	**0.896**	**0.996**
Adrd.AQ	0.991	0.966	0.007	0.003	0.877	0.949	0.993	0.996
Anserver	**0.827**	**0.935**	0.007	0.001	0.909	0.993	0.981	0.995
Basebridge	**0.450**	**0.921**	**0.020**	**0.000**	**0.624**	**0.997**	**0.943**	**0.994**
DroidKungFu	**0.804**	**0.975**	0.000	0.002	0.998	0.981	0.985	0.996
FakeInst	**0.807**	**0.972**	0.010	0.021	0.856	0.779	0.977	0.979
Geimini	**0.818**	**0.870**	0.001	0.002	0.985	0.948	0.992	0.992
GoldDream	**0.682**	**0.994**	**0.014**	**0.004**	**0.792**	**0.945**	0.963	0.996
LightDD	**0.639**	**0.960**	0.001	0.000	0.984	0.996	0.973	0.997
Opfake	**0.475**	**0.764**	0.002	0.008	0.950	0.886	**0.959**	**0.975**
PjApps	**0.557**	**0.999**	**0.026**	**0.004**	**0.554**	**0.941**	**0.951**	**0.996**
RooterBT	**0.718**	**0.970**	**0.033**	**0.008**	**0.620**	**0.905**	**0.950**	**0.991**
SMSHider	**0.789**	**0.947**	0.002	0.002	0.971	0.972	0.983	0.994
Snake	**0.708**	**0.929**	0.002	0.002	0.964	0.968	0.975	0.993
Zitmo	**0.734**	**0.963**	**0.053**	**0.000**	**0.425**	**0.994**	**0.937**	**0.998**

Precision, it showed about 15 % of improvement rate while showing about 2.8 % of improvement rate in the case of Accuracy. The reason why OpFake showed low performance of 0.764 for TPR in SVM classifier is because OpFake was classified as FaskIst by showing similar pattern with malware expanded from FakeInst. In the overall result, however, it shows that the SVM classifier resulted in higher malware detection rate than RF.

68.5 Conclusion and Future Study

In this paper, we proposed a method of detecting Android malware through linear SVM by monitoring resource information. To increase the validity malware detection result, targeting malware was expanded and we performed monitoring by segmenting memory usage amount by classifying into Dalvik Machine area memory and native area. We considered the characteristics of Android platform rather than simply monitoring the total memory usage amount changed by running application. In addition, experiment result was compared with SVM and RF that showed good performance in the previous studies on malware detection to verify the superiority of our proposed method, and experiment results were deduced by composing two types of experiment environment. The result of analyzing the result of Experiment 1, 2 showed that the SVM classifier revealed the result of TPR = 0.944, FPR = 0.004, thereby nearly accurately detecting 14 targeting malware. Since machine learning method has the characteristic of its performance improving when much learning data has been given, there is a possibility of detecting malware excluded in the experiment through the proposed method.

Since numerous variations of Android malware and new malware are being discovered, additional study needs to be conducted in the future on ways to detect unlearned malware. We plan to also conduct studies on control according to malware detection. In addition, there is a plan to conduct studies on abnormal behavior detection such as attack besides malware.

Acknowledgments This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the MSIP (Ministry of Science, ICT and Future Planning) (2013R1A1A3011698).

References

1. F-Secure (2012) “Mobile threat report” Q4 2012
2. Schmidt AD, Camtepe A, Albayrak S (2010) Static smartphone malware detection. In: Proceedings of the 5th security research conference (future security 2010), p 146. ISBN: 978-3-8396-0159-4
3. Bläsing T, Schmidt AD, Batyuk L, Camtepe SA, Albayrak S (2010) An android application sandbox system for suspicious software detection. In: 5th international conference on malicious and unwanted software (MALWARE’2010), Nancy, France

4. Kou X, Wen Q (2011) Intrusion detection model based on android. In: 4th IEEE international conference on broadband network and multimedia technology (IC-BNMT), pp 624–628
5. Bose A, Hu X, Shin KG, Park T (2008) Behavioral detection of malware on mobile handsets. In: Proceedings of the 6th international conference on mobile systems, applications, and services, MobiSys '08
6. Schmidt AD, Schmidt HG, Clausen J, Yüksel KA, Kiraz O, Camtepe A, Albayrak S (2008) Enhancing security of linux-based android devices. In: Proceedings of 15th international Linux kongress, Lehmann
7. Cheng J, Wong SHY, Yang H, Lu S (2007) SmartSiren virus detection and alert for S-martphones. In: Proceedings of the 5th international conference on mobile systems, applications and services, MobiSys '07
8. Liu L, Yan G, Zhang X, Chen S (2009) Virusmeter preventing your cellphone from spies. Recent advances in intrusion detection lecture notes in computer science, vol 5758/2009, pp 244–264
9. Burguera I, Zurutuza U, Nadjm-Tehrani S (2011) Crowdroid behavior-based malware detection system. In: Proceedings of the 1st ACM workshop on security and privacy in smartphones and mobile devices, SPSM '11
10. Shabtai A, Kanonov U, Elovici Y, Glezer C, Weiss Y (2012) "Andromaly" a behavioral malware detection framework for android devices. J Intell Inf Syst 38
11. Fuchs AP, Chaudhuri A, Foster JS (2011) ScanDroid: automated security certification of android applications
12. Enck W, Gilbert P, Chun BG, Cox LP, Jung J, McDaniel P, Sheth A (2010) TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones. In: Proceedings of the 9th USENIX conference on operating systems design and implementation, OSDI'10
13. Burgesm CJC (1998) A tutorial on support vector machines for pattern recognition. Submitted to data mining and knowledge discovery. http://svm.research.bell-labs.com/SVMdochtml
14. Ham HS, Choi MJ (2013) Analysis of android malware detection performance using machine learning classifiers. In: International conference on ICT convergence
15. Kim T, Choi Y, Han S, Chung JY, Hyun J, Li J, Hong JW (2012) Monitoring and detecting abnormal behavior in mobile cloud infrastructure. In: IEEE/IFIP international workshop on cloud management (CloudMan 2012), pp 1303–1310, Maui, Hawaii, USA, 20 Apr 2012

Chapter 69
A Study on Authentication Mechanism to Ensure the Reliability of Content Distribution Using Log Monitoring and Encoding of OPSs

Byoung-Soo Koh, SangHo Seo and ByungOk Jung

Abstract New paradigm digitalization in cultural content enables illegal copy and distribution of music and/or movie in the same quality as original at a rapid pace. This causes the situation that blames the illegal distribution of cultural content on illegal OSP (Online Service Provider). While OSPs observe the relevant copyright law and take filtering technological protection measures, they keep distributing illegal content, disturbing a market order and infringing rights and interests of users and copyright holders evading filtering technology, omitting and forging/falsifying distribution related log record. Existing regulations and filtering technological protection measures are not enough to prevent illegal distribution of content. This paper examines information protection technology with multiple-signature OSP and user access control as a specific solution and proposes a log data authentication system.

Keywords OSP (Online service provider) · Web-hard · Log · Authentication · PKI

69.1 Introduction

As network technology has advanced in recent years, an increasingly wider range of online services have become available that provide and share various content. Online service providers (OSPs) buy content with copyright from content providers (CPs) and provide services to users.

B.-S. Koh (✉) · S. Seo · B. Jung
DigiCAP Co., Ltd., Richmond, Australia
e-mail: bskoh@digicaps.com

S. Seo
e-mail: shseo@digicaps.com

B. Jung
e-mail: bojung@digicaps.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_69, © Springer Science+Business Media Dordrecht 2014

Recently, however, cases of copyright infringement have been on the rise, including forgery and the falsification of OSPs' account data, and there are a growing number of elements that threaten the trust between OSPs and CPs. Information on key management and authentication mechanisms related to copyrights needs to be provided by respective protocols of individual OSPs that only conduct parts of the authentication process.

It is important to ensure the integrity of OSPs' data, enhance the credibility of significant data with coding and authentication, and improve protocols for more reliable data transmission. Another task is to develop a public key infrastructure (PKI)-based authentication sub-system that can systematically certify data given by OSP systems to ensure efficient data processing and enhance the credibility and transparency of OSP services.

This paper examines information protection technology with multiple-signature OSP and user access control as a specific solution and proposes a log data authentication system.

Chapter 2 reviews currently used key management technologies and authentication mechanism, Chapt. 3 analyzes ISO authentication and authentication system infrastructure centered on PKI X.509. The objective is to analyze an access control model that supports authentication technology more effectively and combine it with PKI-based access control technology. Chapter 4 describes the design of a framework for a log data authentication sub-system and proposes an authentication processing model. Lastly, Chap. 5 is a summary of the findings and conclusions.

69.2 Related Studies

69.2.1 Key Escrow: Key Management Methodology

In this process, a user's long-term keys (parts thereof or other key-related information) are held in escrow with one or more institutions. Since the keys are directly kept by the institution, the protection of the user's privacy entirely depends on the institution. For this reason, the institution's credibility is critical, and to ensure this, "secret sharing" is widely used whereby two or more institutions are used to keep keys. Its advantages are that keys can be restored in emergency and that it is a convenient and safe way as long as the institution is trustworthy.

The key escrow is implemented when the user sets up the key system, as shown below.

69.2.2 Encapsulation: Key Management Methodology

Unlike the key escrow process, this process involves creating fields with information that can restore keys in each session of encryption and add the field to code messages. This does not involve a key escrow. Key restoration by a law-executing

agency uses confidential data of the institution to restore the fields added to code messages to retrieve keys. Since it is not a long-term key, but rather a session-specific key, it limits restorability by a wire-tapping agent. For this reason, it is more secure from the user's standpoint than a key escrow. Moreover, if the existing protocol has an extendable field, a restoration fields can be added to save costs and enhance compatibility.

However, since the user generates a restoration field, the user may also deny the field. Thus, it is essential to check the validity of the restoration field and the credibility of a restoring party just as in a key escrow.

69.2.3 TTP: Key Management Methodology

In this method, a trusted third party (TTP) is assumed and all keys are generated in the user's designated TTP and distributed to the user. The TTP is in direct possession of the user's long-term keys. Each TTP's trustworthiness is crucial, as the user's keys are generated and distributed by designated TTPs.

In generating keys, each user's TTP uses keys to generate the users' long-term keys. The users receive the keys to generate session keys for coded communication. Since TTPs have all the users' keys, keys can be fully restored when necessary, and cross-border compatibility can be ensured as long as TTPs adopt the same method of key generation. However, its drawbacks are that the user's privacy entirely depends on TTPs, a large number of TTPs are needed, and often there are bottlenecks between TTPs and users or among TTPs.

69.2.4 Comparison of Key Management Methodologies

The pros and cons of the aforementioned three methodologies may be summarized as below:

	Pros	Cons
Key escrow	• 100 % certainty for key restoration under certain conditions • Convenient and safe	• User's privacy relies on an escrow institution • The escrow institution must be trustworthy
Encapsulation	• Safer than a key escrow • Lower cost • High compatibility • A wire-tapping agency's limited ability to restore keys	• Validity of the restoration field should be checked • The restoring institution must be trustworthy
TTP	• 100 % certainty for key restoration under certain conditions • Excellent cross-border compatibility	• A user's privacy relies on TTPs • The need for a large number of TTPs • Bottleneck between TTP-users or TTP–TTP

69.3 Analysis of Authentication Mechanisms

69.3.1 Introduction to Authentication Technology

A dispersed system poses a serious security threat, as a number of hosts are interconnected on the network and many users share them. Resources allocated among the hosts are shared in the form of a network service provided by the server. A requesting party of an individual process that requires access to a particular resources may a request service from the respective server, which requires authentication for a cross-identification check.

Risks related to message transmission include wire-tapping on the network, unauthorized revision, insertion or deletion of a message, or a replay attack. An integrity service ensures that all messages are received in the same form as they are sent [1–3].

In a dispersed system, authentication may be executed based on information exchanged between a requesting party and an authorizing party in one of the following three formats:

- Message content: Match between received message and sent message.
- Message source: Match between the sender of the received message and the sender as recorded in the field for an identity check.
- Substantial authentication: Match between the communication subject of connection or session with the claimed identity for an identity check.

For the safe design and operation of a dispersed system, authentication might be required under the following circumstances.

- Host–Host: For host-level activities that require certain coordination among hosts, link data need to be exchanged for individual hosts to modify their internal connection configuration.
- User–Host: A user may log on to the host to acquire access to a dispersed system. Host systems are dispersed and provide an open, accessible environment and require cross-authentication between users and hosts.
- Client–Server: In a dispersed network environment, a server needs to provide services to only authorized clients, and clients are only interested in legitimate server transactions. For example, cross-authentication might be needed between a seller and a consumer.
- Territory–Territory: A dispersed system may consist of many sub-systems, each of which is operated under an individual information protection policy. Communication subjects from different territories might require authentication.

69.3.2 PKI Authentication Service

PKI is the fundamental technology that ensures the safety and reliability of activities that take place in network environments including the Internet. X.509 is an ITU-T standard for PKI, part of the X.500 series standard that defines directory services. A directory is a set of dispersed servers that manage database of servers and users. Its data include user information, other features, and corresponding relationships between user names and network addresses.

X.509 defines a structure to specify the authentication service of the X.500 directory for its users. A directory may also be a depository of public key certificates. Each authentication contains users of the public keys and signed with personal keys of the trusted authentication authority. In addition, X.509 defines authentication protocols based on public key authentication.

The core of the X.509 structure is public key certificates related to each user. The user authentication is assumed to be issued by a trusted CA (authentication authority), and located on a directory by either a user or CA. The directory server is not responsible for the generation of public keys or authentication. It merely provides an easily accessible venue for users to acquire authentication.

Figures 69.1 and 69.2 show a typical authentication with the following elements.

69.4 Log Authentication System

69.4.1 OSP System Environment

When an SW blackbox is installed and operated in a web-hard company's server room or data center, its authentication and integrity should be guaranteed for the copyright holder's trust. To do so, studies on third-party authentication are needed. As part of such endeavors, this research examines key management methodology, a safe key delivery algorithm, and TTP algorithms to develop a methodology to provide an integrated SW blackbox through third-party authentication. The major functions of third-party authentication include key management algorithms to ensure the integrity of log data, the generation and management of authentication for safe key delivery, and the development and management of a policy for regular key updates. These are developed and designed to be flexibly applicable to OSPs in charge of log generation and backup, content-filtering companies, and client applications (Fig. 69.3).

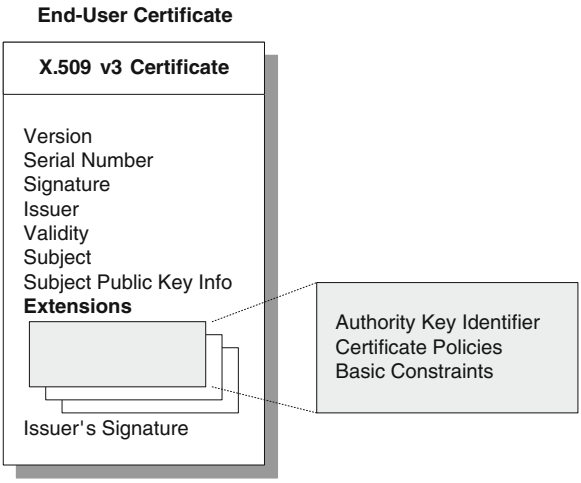


Fig. 69.1 X.509 user authentication format

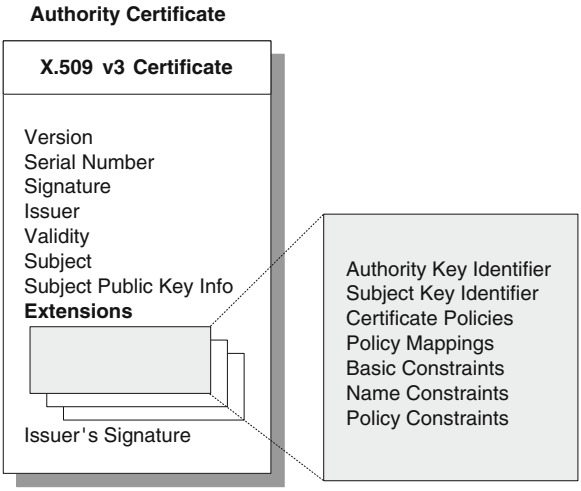


Fig. 69.2 X.509 institutional authentication format

69.4.2 Design of the Log Data Authentication System

A proper level of access control should be applied to log data safely stored in OSPs, and a process should be instituted to verify that log data are not forged or falsified even if they are original log or filtered logs tailored to each CP.

In an authentication system, source data, UTC, access control information, and the signature of each OSP are stored together. The integrity of source data can be verified without using a separate trusted hardware. In comparison, it is hard to ensure the credibility and integrity of data sources under the previous system environments, as log data are directly collected and processed within OSP systems.

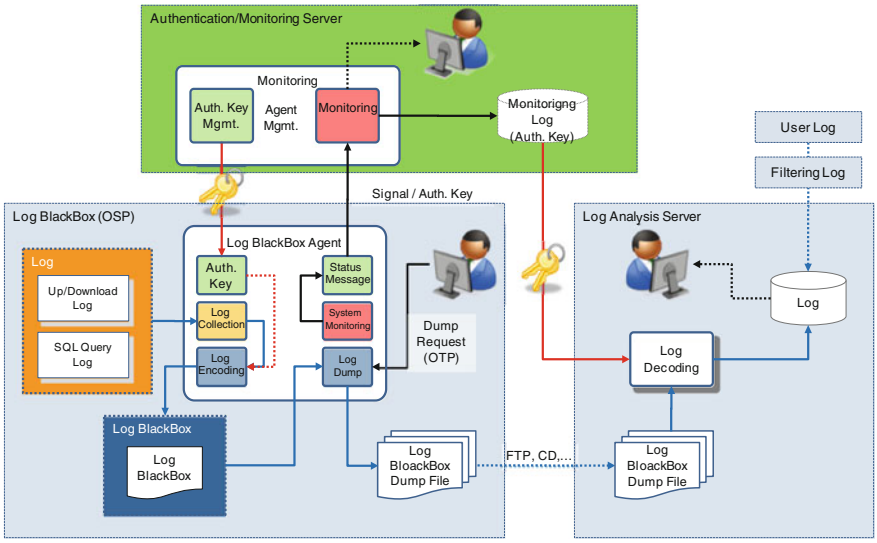


Fig. 69.3 Architecture of OSP (web-hard) system

For safe log management, the authentication system adds signature data to log data (account log, system log, DB log), and conducts log data authentication in a systematic way to enhance the credibility and integrity of log data collected in the OSP system.

69.4.2.1 Role of the Log Data Authentication System

This study proposes a log data authentication system under an OSP environment, which uses PKI-based authentication for authentication among OSP servers, CPs, and monitoring systems. The system is designed in a way that guarantees that log data are provided by expected clients and servers and information integrity by verifying that source log data collected by D/L (data logger) are not manipulated by the administrator or through a collusion with a content-filtering company. Users connect to the monitoring server through cross-authentication among authentication authorities to receive services.

69.4.2.2 Configuration of the Log Data Authentication System

1. System configuration

The log data authentication system consists of a front end processor (FEP) that receives source log data through communication with a D/L, the authentication processing part that conducts digital signatures on source log data, an NTP to

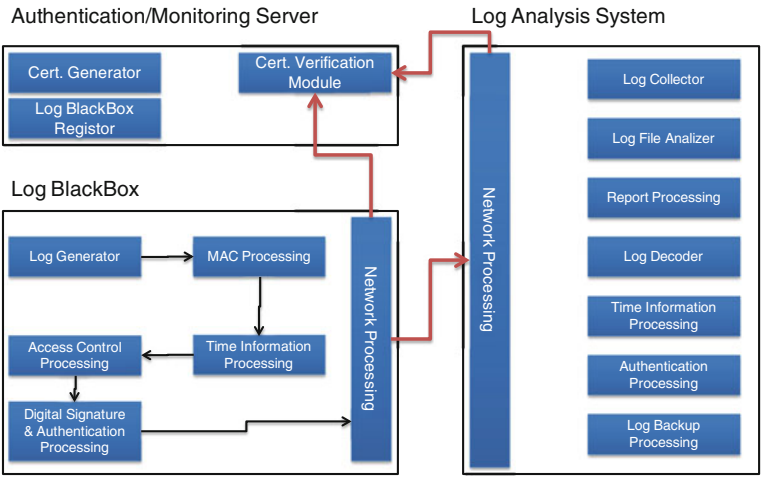


Fig. 69.4 Architecture of third-party authentication system

prevent time falsification by setting accurate world time, and the access control processing part that grants access rights to log data (Fig. 69.4).

The data field with digital signatures from the authentication processing part is illustrated below:

The field consists of average log data for 5 min and 1 h collected from D/L, the universal time coordinated (UTC) from the NTP server, and the signature of each system (OSP, CP). Signature fields may be expanded by the number of divisions that need authentication. Access control for average data is recorded on the Target ACI field. Five-minute average data and 1 h average data are each stored for 1 day and 7 days, respectively (Fig. 69.5).

2. Authentication processing

The UTC data are added to 5 min average log data from the D/L to generate a hash value, signed with personal keys on the log data authentication system, and the hash value and signature value are transmitted to the CP. Actual value is unknown to the CP, as it only receives the hash value of log data. The authentication system may be authorized with signatures from the log data authentication system, and each CP sends a hash value to the system through a signature. The log data authentication system verifies the signature, as illustrated in the work flow below, and stores signature of each division along with source log data (Fig. 69.6).

Data (10 minutes)	UTC (time info.	AC Information	Signature 1	Signature 2	Signature 3	.....	Signature n
Data (1 hour)	UTC (time info.	AC Information	Signature 1	Signature 2	Signature 3	.....	Signature n

Fig. 69.5 Example of a digital signed data field by the authentication processing module

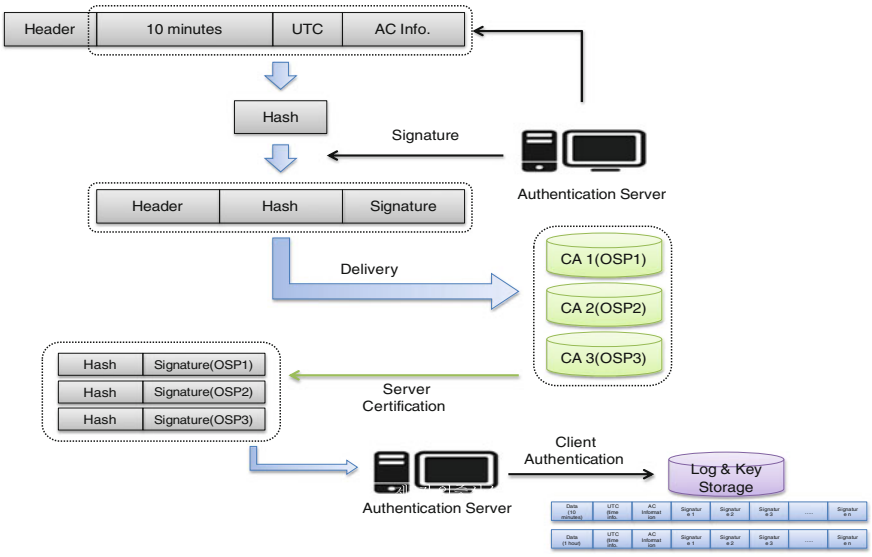


Fig. 69.6 Signature authentication flow chart

69.5 Conclusions

In this study, a log data authentication system embedded with information protection technology was proposed to enhance the credibility of various log data on an OSP system, with the goal of building a trusting relationship between OSPs and CPs regarding copyright protection. The proposed log data authentication system enables server/client authentication, guaranteed log data credibility, user access control, and audit records. The system adopts PKI-based public keys for authentication among OSPs, CPs, and third-party authentication systems, ensures that log data are provided by the expected client and server, and that the source log data collected from the D/L are not manipulated.

The earlier versions of log data monitoring systems only provided fragmented and local information and merely guaranteed the integrity of log data based on the log data collected within an OSP system. In comparison, the proposed authentication system can further open up log data through various channels, enhance its understandability through statistical and structural analysis, and meet user

demands for transparent information sharing by adopting a network environment. It was shown that the log data authentication system based on encrypted authentication mechanism is superior in terms of scalability and authentication efficiency and provides a wider range of security services with excellent confidentiality, integrity, non-repudiation, and availability.

Moving forward, it will be necessary to undertake more comprehensive research to ensure the integrity of authentication and audit records and to establish a safer communication system using the internet security protocols such as SSL, TLS, and IPSec.

Acknowledgments This study was conducted as part of research for the 2013 Copyright Technology Development project by the Ministry of Culture, Sports, and Tourism and Korea Copyright Commission.

References

1. Woo TYC, Lam SS (1992) Authentication for distributed systems, computer
2. www.iso.org
3. Mendes S, Huitema C (1995) A new approach to the X.509 framework: allow a global authentication infrastructure without a global trust model, symposium. On network and distributed system security
4. Seo S, Jung B-O, Ko B, Yum S (2012) Log-blackbox study on reliability guarantee of content distribution by observing events in OSP file system, ITCON2014 Summer 2012
5. Korea Copyright commission (2011) Study on evaluation of technical actions applied in OSP, policy research 2011-09
6. Korea Copyright commission (2013) Annul report for 2013 copyright

Chapter 70
Virtual Agent Learning for NUI/NUX-Based Games

Daxing Jin, Seoungjae Cho, Yunsick Sung, Kyungeun Cho and Kyhyun Um

Abstract Natural User Interface (NUI) embodies the concept of direct interaction with computers using the human body instead of using existing (intermediary) input devices. Further, Natural User eXperience (NUX) incorporates the concept of more natural and intuitive feedback in accordance with the user's command. This paper proposes an approach that enables virtual agents to interact with users playing games on the basis of NUI. Users have a more immersive gaming experience as a result of the NUX-based interactive virtual agent and the use of NUIs in games. To enable the virtual agent to learn automatically via interaction, we generate a virtual human using imitation learning. Then we apply Q-learning to enable the virtual agent to learn in a short time. The results of two experimental scenarios confirm that our proposed method is sufficiently able to instruct a virtual agent to support a virtual human.

Keywords Natural user interface · Natural user eXperience · Virtual agent learning

D. Jin
NC Soft, Seoul, Republic of Korea

S. Cho · K. Cho (✉) · K. Um
Department of Multimedia Engineering, College of Engineering, Dongguk University, Pildong-3-ga, Jung-gu, Seoul 100-715, Republic of Korea
e-mail: cke@dongguk.edu

Y. Sung
The Department of Game Mobile Contents, Keimyung University, Daegu, Republic of Korea

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_70, © Springer Science+Business Media Dordrecht 2014

70.1 Introduction

Natural User Interface (NUI) and Natural User eXperience (NUX) are two highly advanced user interface types that are actively being investigated. In addition, fields related to NUI/NUX-based gaming have been experiencing progressive advancement in their research and development efforts. This paper describes an approach that enables more immersion for NUI-based game users through a virtual agent. The virtual agent interacts with the virtual players controlled by users on the basis of NUI and enables the users to experience NUX beyond simple visual feedback. We attempt to substantially reduce the time required for a virtual agent to learn by applying an interaction learning approach. In our approach, the virtual agent learns together with a virtual human, resulting in the virtual human and virtual agent learning to interact more quickly.

The virtual human learns via imitation learning based on direct instruction data from users. In imitation learning, an entity observes another's behaviors and learns those behaviors by repeatedly imitating them. By means of this process, artificial intelligence that acts in similar ways to users can be implemented as virtual humans. Further, the virtual agent learns by applying the Q-learning algorithm based on the results of repeated interaction with the virtual human. This interactive learning enables the virtual agent to execute corresponding sub-tasks by ascertaining the tasks being performed by the virtual human.

The remainder of this paper is organized as follows. Section 70.2 gives an overview of related work. Section 70.3 outlines the learning processes utilized by the virtual human and agent, while Sect. 70.4 presents a number of task definitions. Sections 70.5 and 70.6 describe the virtual human and virtual agent learning processes, respectively. Section 70.7 discusses the experimental results obtained and Sect. 70.8 concludes this paper.

70.2 Related Work

Bicho [1] adopted Dynamic Neural Fields (DNF) as an interactive learning approach for interaction between human beings and robots. In the approach, DNF is used to recognize simplified human behavior data by pattern matching. Calinon and Billard [2] proposed an imitation learning framework that uses several datasets acquired from human beings who directly and repeatedly controlled a robotic arm. The framework uses Principal Component Analysis (PCA) to delete unnecessary data from the datasets and converts the PCA results into statistical distribution data using Independent Component Analysis (ICA). Finally, it applies Hidden Markov Model (HMM) to generate the learning model from the data. Thurau et al. [3] examined an approach to automatically generating artificial intelligence characters from learning datasets encapsulating the process of human beings playing games.

In the approach, the datasets are expressed in an arithmetic model based on Bayesian probability for imitation learning.

Other approaches apart from imitation learning are also used to generate artificial intelligence characters for games. Spronck et al. [4] proposed an approach to generating artificial intelligence characters that enabled reliable actions in a number of extraordinary situations more quickly and effectively by applying unsupervised online learning in real time. However, the artificial intelligence characters in his approach have relatively simple behaviors because they utilize script-based learning. Salem et al. [5] investigated an approach to teaching a robot in a virtual environment that reduces the time to determine optimum behaviors by making the robot learn the behaviors via virtual simulations. Yamaguchi et al. [6] proposed an approach that uses reinforcement learning to make a virtual agent learn. Like Salem's approach, the proposed approach uses a virtual agent that learns via simulation in a virtual environment. The learning results are then applied to a robot.

In this paper, we propose to generate a virtual human to aid in the learning of the virtual agent. Imitation learning is applied to the virtual human and Q-learning to the virtual agent for interaction with the virtual human. By generating a virtual agent that supports a user playing NUI-based games, the immersion experience of the user is enhanced.

70.3 Virtual Human and Virtual Agent Learning Processes

Figure 70.1 illustrates the learning processes utilized by the virtual human and virtual agent. First, a teacher directly describes a series of behaviors to be learned by the virtual human, which is then generated through imitation learning based on the series of behaviors. The virtual agent learns from interacting with the virtual human via Q-learning.

70.4 Predefinition

We define a Sub-task Table for the virtual agent to learn from interaction with the virtual human. The Sub-task table defines the sub-tasks of the virtual agent that can help the virtual human with its tasks. Table 70.1 shows examples of the Sub-tasks used in the table.

We also define a Task Success Probability Table to enable selection of the proper sub-task for the virtual agent when the virtual human tries to execute a specific task. Table 70.2 is a sample of the Task Success Probability Table.

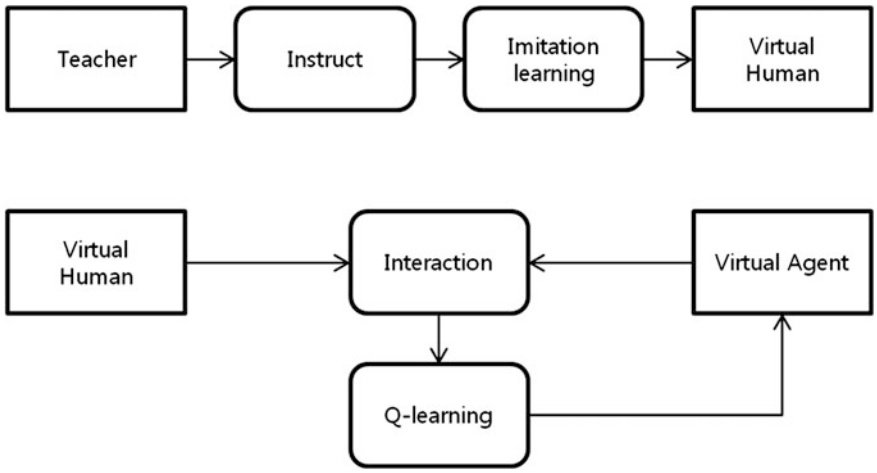


Fig. 70.1 Learning processes utilized by the virtual human and virtual agent

Table 70.1 Sample of the sub-task table

Tasks of virtual human	Sub-tasks of virtual agent
Sleep	Light off
Eat	Prepare food
…	…

Table 70.2 Sample of the task success probability table

Status	Behavior	Task 0	Task 1	…
Status 0	Behavior 0	14	0	…
Status 0	Behavior 1	3	32	…
…	…	…	…	…

70.5 Virtual Human Learning

We apply Bayesian probability-based imitation learning for the virtual human's learning. A teacher describes the series of behaviors to generate the data to be learned by the virtual human, which is then saved as one Behavior Vector. A behavior executed in a specific position is one element of a Behavior Vector. The teacher provides several independent Behavior Vectors by describing diverse behaviors.

The Behavior Frequency Table by Status is updated by accumulating the frequency of behaviors in a specific status in several Behavior Vectors executed by the teacher. The virtual human determines the next behaviors in the present status on the basis of the Behavior Frequency Table by Status. Figure 70.2 presents the process used to generate the Behavior Frequency Table by Status acquired by the teacher's instructions.

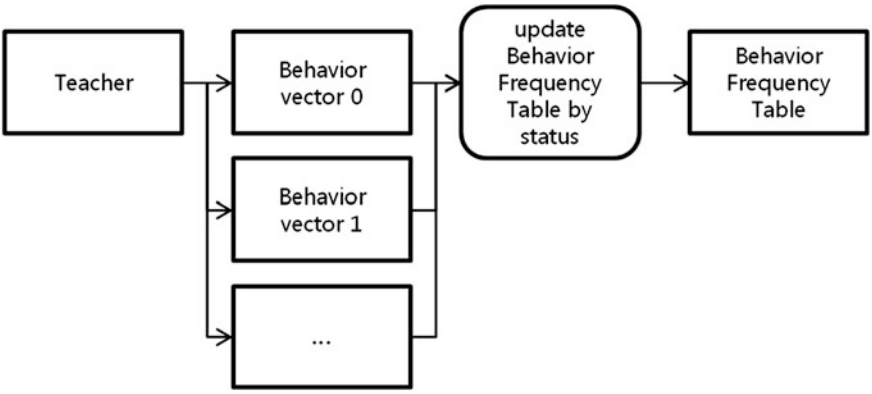


Fig. 70.2 Generating behavior frequency table by status via the teacher's instructions

70.6 Virtual Agent Learning

The virtual agent learns by interaction with the virtual human through Q-learning. This Section describes the Q-learning process utilized by the virtual agent and the updating of the Q-table with the results of interaction between the virtual human and the virtual agent.

The Q-table contains the total compensation value of the behaviors for each status that the virtual human and the virtual agent may have. Q-learning generates compensation values that are greater than zero at the time of task completion. This compensation value is added to the total compensation value of the behavior by status executed until the tasks are completed. When there is a task failure, a compensation value that is less than zero is added to the total compensation value of the behavior by status executed up to that point.

For learning interaction between the virtual human and the virtual agent, the virtual agent needs to ascertain the tasks of the virtual human. Figure 70.3 illustrates the process used by the virtual agent to select the sub-tasks while observing the virtual human.

The virtual agent ascertains the task being performed by the virtual human by identifying the task with the highest frequency of behavior executed in the present status in the Task Success Probability Table. Then, it determines the sub-tasks to use to aid the virtual human by referring to the Sub-task table.

70.7 Experiments and Analyses

We implemented a virtual environment to investigate the extent to which the virtual agent learns through interaction with the virtual human. Figure 70.4 shows the virtual environment used in our experiment.

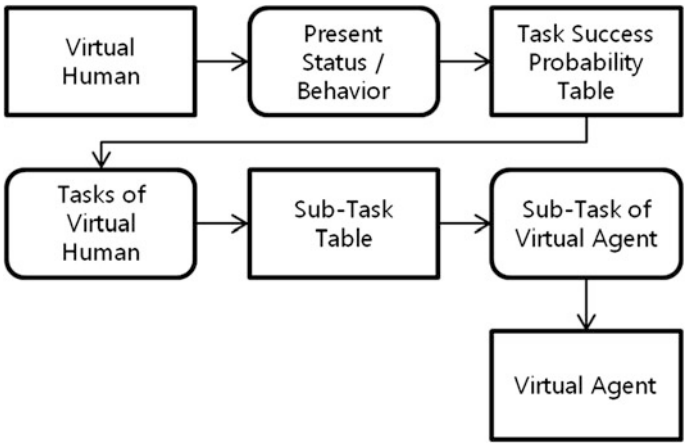


Fig. 70.3 Determination of sub-tasks by virtual agent while observing virtual human

The virtual environment used in the experiment is a space with a lattice form for application of imitation learning and Q-learning. It was implemented as a virtual residential space using the Unity3D engine. Each rectangle comprising each lattice form shows the location of the virtual human and the virtual agent as each status.

Figure 70.5 depicts the 3D models of the virtual human and the virtual agent used in the experiment.

Two learning objectives were defined for the experiment:

- the virtual agent should learn to turn off the light when the virtual human goes to bed; and
- the virtual agent should set the table when the virtual human desired to have a meal.

Table 70.3 shows the sub-task table used in the experiment.

In accordance with the learning objectives, a teacher instructed the virtual human using keyboard input. Next, the virtual human and the virtual agent learned via repeated interaction. The virtual human executed the tasks as instructed by the teacher after learning while the virtual agent observed the present status and behaviors of the virtual human and continuously updated the Task Success Probability Table through Q-learning. In accordance with the repeated interaction learning, the virtual agent was able to successfully complete the sub-tasks with a probability of 100 % after interacting with the virtual human approximately 1,000 times. The virtual agent learned very quickly because the experiment was implemented in the virtual environment.

Fig. 70.4 Virtual environment used in the experiment

Fig. 70.5 3D models of the virtual human and virtual agent

Table 70.3 Sub-task table used in the experiment

Tasks of virtual human	Sub-tasks of virtual agent
Sleep	Light off
Eat	Prepare food

70.8 Conclusion

This paper proposed a learning approach for a virtual agent that can provide interactive NUX to users in NUI-based games. The approach to teach interaction between human beings and robots was applied to learning by the virtual agent in a virtual environment. First, the virtual human learned via imitation learning. Then, the virtual agent learned by interaction with the virtual human through Q-learning. In the experiments conducted, proper interaction of the virtual agent with the virtual human was achieved very quickly.

We plan to develop content that utilizes a virtual agent that interacts with actual users by applying the results of this study to NUI games. Further, the results of this study can be applied to agent learning in virtual environments for senior citizens and physically challenged persons, as well as to NUI games.

Acknowledgments This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2013-H0301-13-4007) supervised by the NIPA (National IT Industry Promotion Agency). And this work was also supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0011266).

References

1. Bicho E (2009) A dynamic field approach to goal inference and error monitoring for human-robot interaction. AISB 2009 convention, pp 31–37
2. Calinon S, Billard A (2005) Recognition and reproduction of gestures using a probabilistic framework combining PCA, ICA and HMM. In: Proceedings of the 22nd international conference on machine learning (ICML), pp 105–112
3. Thurau C, Bauckhage C, Sagerer G (2004) Learning human-like movement behavior for computer games. In: Proceedings of international conference on the simulation of adaptive behavior 2004, pp 315–323
4. Spronck P, Sprinkhuizen-Kuyper I, Postma E (2003) Online adaptation of game opponent AI in simulation and in practice. In: Proceedings of the 4th international conference on intelligent games and simulation (GAME-ON 2003), pp 93–100
5. Salem M, Kopp S, Wachsmuth I, Joublin F (2010) Generating robot gesture using a virtual agent framework. In: International conference on intelligent robots and systems (IROS), 2010 IEEE/RSJ, pp 3592–3597
6. Yamaguchi T, Masubuchi M, Tanaka Y, Yachida M (1996) Propagating learned behaviors from a virtual agent to a physical robot in reinforcement learning. International conference on evolutionary computation, pp 855–860

Chapter 71
An Analysis of the Korea National DNS Using Big Data Technology

Euihyun Jung, Joonhyung Lim and Juyoung Kim

Abstract The DNS (Domain Name System) log contains useful information in terms of industry and society because it represents the host access information of Internet use. However, the DNS log analysis has not been studied except in the area of security until now because the size of the DNS logs is too huge and the logs are scattered in world-wide distributed servers. To resolve these issues, we propose the Localized Big Data Analysis (LBDA) with which data mining techniques are applied to the DNS log. We also perform an experiment using the real logs of the Korea national DNS servers and the result proved the validity of the LBDA. Using the LBDA, the DNS log is used to extract a lot of meaningful information for data-based decision-making by the government and enterprises.

Keywords DNS · Big data · Data mining

71.1 Introduction

The DNS (Domain Name System) is the core infrastructure of the Internet and it maps the domain and IP address mutually [1]. The DNS has a reverse tree structure beginning with a top-level domain and each domain is connected to a particular

E. Jung (✉)
Department of Computer Science, Anyang University, 602-14, Jungang-ro, Buleun-myeon, Ganghwa-gun, Incheon 417-833, Korea
e-mail: jung@anyang.ac.kr

J. Lim · J. Kim
Center of Korea Network Information, Korea Internet and Security Agency, Platinum Tower, 398 Seocho-ro, Seoch-gu, Seoul 137-857, Korea
e-mail: jhlim@kisa.or.kr

J. Kim
e-mail: ethan@kisa.or.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_71, © Springer Science+Business Media Dordrecht 2014

server, service, or resource [2]. Since the DNS query for a target host's IP address is the start point of Internet activities, the information related to all Internet use, in particular, host access, will be recorded in DNS logs. The DNS log has been attracting data analysts' interest [3, 4] as a result. There have been some studies regarding worm or botnet detection using the DNS in the security area [5–7]. However, to our regret, there has been no data mining research such as the access analysis on a particular host or country because DNS logs are scattered among DNS servers of many independent organizations, thereby it is difficult to collect DNS logs in an integrated manner for analysis.

Among DNS servers, the Korea national DNS is managed by the government as an authoritative DNS server to manage Korea country domains such as ".kr". The national DNS has a total of 15 servers, which are distributed at home and abroad, and it accumulates 1.5 billion query request logs in the size of 300 GB every day. Unlike other DNS servers, the national DNS is managed by the Korea Internet & Security Agency (KISA). The national DNS can provide the domain access status from the national viewpoint and it is suitable for data mining. However, currently, it takes several days to gather the logs scattered among 15 servers to the central server and to preprocess the logs for the one-day national DNS logs. The analysis method is also limited to the "maximum accessed domain" SQL query after this work because the SQL query is not effective for analyzing data in the various aspects. The conventional analysis programs such as R also have the limitation in the physical memory and they can only use a single core regardless of their running platforms' CPU capacity, thereby they cannot analyze the national DNS log due to its huge size (300 GB a day).

There are several technical issues that block the activation of data mining for the national DNS log. First, the DNS log is too huge to handle with the existing analysis tools. Second, DNS logs are dispersed among 15 DNS servers due to its characteristics. Unfortunately, two issues are conflict. The first issue is simply solved by applying the big data technology, especially Hadoop, which has been getting the spotlight these days. However, the second issue hinders to apply Hadoop because Hadoop assumes its HDFS (Hadoop Distributed File System) to maintain the whole data for analysis. The national DNS logs are dispersed worldwidely and it is very difficult to load the scattered logs into a single HDFS within a limited time.

This study proposes a method named "Localized Big Data Analysis (LBDA)" that is applied to the Korea national DNS log to solve these conflict issues. In order to prove the validity of the LBDA, we performed an experiment using the real logs on the Korea national DNS servers. The result showed the LBDA supported the various data mining techniques and had an advantage of the processing speed.

This paper is composed as follows. Section 71.2 describes the previous research, the structure of the Korea national DNS, and the technical issues of the data mining for the DNS log. Section 71.3 proposes the LBDA for resolving the issues and describes the structure of the analysis system. Section 71.4 explains the performed an experiment and evaluates the result of the experiment, and Sect. 71.5 draws a conclusion.

71.2 Previous Research and Technical Issues

71.2.1 Previous Research

Some studies have been conducted in the security area to detect worms or botnets by analyzing the access patterns in DNS logs [5–7]. The behaviors related to malicious code can be identified by monitoring DNS logs, because the domain query is the essential preceding action for all Internet connections. For instance, just before the attack, DNS queries tend to be increased for botnets because the bots have to access their Command and Control (C and C) server that issues attack commands. EXPOSURE [5] tried to understand the pattern related to a particular attack by applying 15 characteristics to DNS logs. Whyte et al. [6] proposed a method of applying the DNS for scanning worms. In addition, the Korea University research team proposed a method of detecting botnets by monitoring the characteristic that many distributed botnets make DNS queries at the same time. As described above, there has certainly been some degree of research about finding security problems and abnormal symptom in DNS logs. However, there are few studies that have applied the data mining techniques to extract socially and economically useful information from the DNS log because the data mining requires the understanding of the latest data trends by analyzing the huge amount of DNS logs periodically unlike the security area.

71.2.2 The Structure of the Korea National DNS

The DNS can be categorized into the cache DNS, the national DNS, and the root DNS [8]. An Internet user makes a query directly to the cache DNS. When the query is not resolved by the cache DNS, it is delivered to the national DNS and the root DNS. The national DNS and the root DNS are the authoritative DNSs that mange the country domain and the top level domain, respectively. A total of 15 servers (nine servers at home and six servers abroad) are running as the Korea national DNS deployed as shown in Fig. 71.1.

71.2.3 Technical Issues

It has been difficult to perform the data mining with the DNS log except in the security studies, because there have been no appropriate technical means to analyze the huge quantity of data dispersed world-widely. 300 GB data per day in DNS logs requires an excessive system specification for the existing in-memory analyzing tools like R. In addition, the in-memory tools take too much time to load the data into the memory, thereby a big data analysis platform like Hadoop [9]

Fig. 71.1 The deployment status of the Korea national DNS

should be used to analyze the national DNS log. Hadoop enables to analyze the large quantity of data, which has been unthinkable up to now, because there are various analysis tools for big data in its ecosystem.

However, the biggest problem in applying Hadoop is that all logs are needed to be stored in a single HDFS. In Hadoop, several computers are connected with the high-speed network and they construct the distributed file system named HDFS. The distributed DNS logs seem to be combined into a single HDFS because they are dispersed over the network. However, it is impractical to combine the Korea national DNS servers distributed around the world (e.g., U.S., Singapore, Germany, etc.) into a single HDFS because the network connection speed of HDFS is assumed to be much higher than at least 1 Gbps. For this reason, it is difficult to apply Hadoop to analyze the national DNS log even though Hadoop is expected to be the only feasible solution.

71.3 Proposed Method and Experiment Environment

71.3.1 Proposed Method

In order to solve the technical issues described above, we propose "Localized Big Data Analysis (LBDA)". The idea of the LBDA is from the philosophy of Hadoop, "Code moves near data for computation" [10]. In the method, the Hadoop system is installed in each individual national DNS server. After installation, analyses are

locally performed in each individual system and then the results are gathered to the central server. The LBDA has an advantage of saving the delay time because individual DNS servers locally analyze DNS logs instead of sending raw DNS logs to the central server. In addition, the LBDA is quite appropriate for the situation in which everyday's data pattern should be reported within a limited time like the DNS log because the localized analysis on individual DNS servers will significantly reduce the analysis time.

However, the proposed method is applied to only the data mining techniques which are performed in the distributed manner such as sum, average, frequency, max, and min. The statistical value of the techniques is obtained by summing the results of distributed processing. Equation (71.1) represents this concept mathematically and F in the Eq. (71.1) can be replaced with the function of sum, average, frequency, max, or min.

$$F(total) = \sum F(local) \tag{71.1}$$

Since the frequency analysis for the domain access is the most fundamental one in the DNS log analysis, it is important to check whether the proposed method can be applied to the frequency analysis or not. If the frequency analysis with various conditions (e.g., location, time, domain, etc.) is performed with the method, the result will be very useful to understand the DNS access pattern. Expression (71.2) represents the frequency analysis for the domain access using the method.

$$Freq(Total_by_conditions) = \sum Freq(Local_by_conditions) \tag{71.2}$$

71.3.2 Experiment Environment

To verify the validity of the LBDA, we need to install Hadoop on each national DNS server. However, it is impractical to install the Hadoop on all the national DNS servers at this moment due to economic reasons and the national policy. It would be also a wrong way round because this study is designed to verify whether the installation of the big data analysis system on the national DNS servers is effective or not.

We therefore selected the one-day data (August 16, 2013) from the national DNS logs and analyzed it, as if the data was locally analyzed by the individual DNS servers. That is, the one-day data was divided into 15 datasets according to the location of the national DNS servers. For example, we grouped the data collected from the national DNS located in Washington in the U.S. as "g1". Having analyzed these grouped data individually, we compared the result with the one of entire data in order to test the validity of the LBDA. Table 71.1 shows the information on the divided dataset. We excluded two DNS servers out of 15 servers from the test because the two servers had not recorded logs. The data set contains all pairs of query request and response as the record. There are 2.4 billion

Table 71.1 The dataset for the experiment

No	Dataset name	Location	Data size (B)	A number of records
1	b0	Seoul	55,723,153,757	472,854,758
2	c0	Anyang, Gyeonggido	Not collected	–
3	d0	Seoul	1,379,684,939	11,407,520
4	d1	San Francisco, USA	2,901,614,875	25,879,603
5	d2	Seoul	Not collected	–
6	d3	Singapore	11,121,175,162	100,567,085
7	d4	Bundang, Gyeonggido	41,193,871,357	329,419,892
8	e0	Daejeon	2,026,839,286	16,080,929
9	e1	Beijing, China	2,838,299,834	29,680,295
10	e2	São Paulo, Brazil	6,503,301,738	23,679,355
11	e3	Seoul	84,466,201,069	720,866,264
12	f0	Seoul	59,817,095,771	524,415,216
13	g0	Seoul	5,973,103	58,535
14	g1	Washington DC, USA	12,081,819,937	96,143,500
15	g2	Frankfurt, Germany	38,964,469,838	343,219,585

records, which is two times more than the query requests dated on August 16 (1.2 billion requests) as a result. To configure a Hadoop environment, we clustered four PCs with i5 quadcore CPU, 16 GB RAM, and 6 TB HDD.

71.4 Experiment and Evaluation

71.4.1 The Frequency Analysis of a Particular Domain Access

The Pig code shown in Fig. 71.2 was used to analyze the access frequency to a particular domain. We analyzed the frequency of accessing the "ns.hardware.co.kr" domain which ranked the fifth in the access frequency from China dated August 16. To calculate the frequency, we counted the number of query requests to the "ns.hardware.co.kr" at servers. For the experiment, we selected the "ns.hardware.co.kr" domain, which is the game company "Com2Us", because the domain had the relatively higher access frequency regardless of countries.

The results were obtained from 12 datasets through analyses as shown in Table 71.2. After summing up each result, the total number of access frequency was 3,123,206. It was the same value which we got after analyzing the access frequency of the entire DNS logs. It implies that the frequency analysis is processed using the proposed method. The result showed that the big data analysis was successfully performed for the distributed data such as the DNS log without collecting all data to the center when the data mining techniques support the distributed processing.

```
dns = load '0816-e3' using PigStorage(',') as
  (server_name:chararray, day:int, hour:int, min:int,
  sec:int, msec:int, src_ip:chararray, length:int,
  dst_port:int, flag:int, question:int, answer_rr:int,
  auth_rr:int, add_rr:int, domain:chararray, d_type:int,
  d_class:int, cc:chararray, cont:chararray,
  tzone:chararray, region:chararray, isp:chararray,
  city:chararray, lat:float, lon:float);
domain_dns = foreach dns generate d_type, domain,
   server_name;
filtered_dns = filter domain_dns by d_type == 1 and
  domain matches 'ns.hardware.co.kr';
grpd_server = group filtered_dns by server_name;
cnt_server = foreach grpd_server generate group,
    COUNT(filtered_dns)as server_count;
order_server = order cnt_server by server_count desc;
store order_server into '0816-out/hardware-server-e3'
```

Fig. 71.2 A pig code to analyze the access frequency of a particular domain

Table 71.2 The result values of the experiment

e3	b0	f0	d4	g1	g2
815,111	566,400	480,977	436,396	406,425	125,323
d3	e1	e2	e0	d1	d0
111,380	60,423	37,655	34,422	31,516	18,178

71.4.2 Evaluation

The experiment verified that the LBDA, which is proposed in this paper, is the appropriate method for the DNS log analysis. The existing data mining of the DNS log have been limited, but the LBDA shows a new possibility of analyzing the DNS log. Since the LBDA performed the experiment to get a simple statistical quantity of the access frequency, it seems to be similar to a method that records the access frequency in the individual servers. However, the method, which records the access frequency of a particular domain in the individual servers, can only identify the fixed characteristics of pre-designated domains among numerous domains. Compared with the method described above, the LBDA can analyze the DNS logs in various aspects on demand, because the DNS logs are diversely processed with the Map/Reduce of the LBDA. We conducted a lot of analysis on the DNS logs with various conditions such as the domain access status by country, time zone, or cities. In the analysis, we found the malicious C&C domain,

Table 71.3 The elapsed time of performing the experiment

e3	b0	f0	d4	g1	g2
14 min 6 s	9 min 28 s	10 min 12 s	7 min 11 s	6 min 46 s	2 min 36 s
d3	e1	e2	e0	d1	d0
2 min 25 s	1 min 32 s	1 min 31 s	1 min 16 s	1 min 29 s	1 min 10 s

"smartfind.co.kr", and popular domains from Japan, China, USA, Brazil and France. This kind of information is not achieved from the conventional monitoring system which calculates only the pre-defined access frequency.

In addition, the analyzing speed is another benefit of the proposed system because individual DNS servers locally perform analyzing its logs. Table 71.3 shows the time taken to perform the experiment. Even though the running time was different depending on the size of the logs on the servers, the each analysis took 4 min and 59 s on average. When the entire logs were processed, it took more than 1 h, which is much longer than the maximum elapsed time of analyzing logs on the individual DNS servers. The reason is that the Map/Reduce [9] takes more time when it processes bigger data. The all results from the experiment are obtained within 15 min using the LBDA. It means that only one quarter of the time is needed, compared with analyzing the entire logs, thereby the LBDA has the benefit in terms of the analyzing speed.

71.5 Conclusion

Since the DNS log contains all access patterns of the Internet use, it has drawn much interest of data analysts. They have expected that the socially and economically useful information can be extracted through the data mining for the DNS log. However, the DNS log analysis has been used in the security area until now because the size of the DNS logs is too huge and the logs are scattered in world-wide distributed servers. These characteristics make the existing data mining techniques difficult to apply. In order to resolve it, we proposed the Localized Big Data Analysis (LBDA). It analyzes the DNS logs locally at the individual DNS servers and provides a way to combine the results. We applied the LBDA to the Korea national DNS and performed the experiment of the frequency analysis of the domain access. The results prove that the LBDA is an effective method to analyze both huge and distributed DNS logs. We will have a plan to research other data mining techniques to be interworked with the LBDA for further research.

References

1. Mockapetris P, Dunlap KJ (1988) Development of the domain name system. ACM SIGCOMM Comput Commun Rev 18(4):123–133
2. Aitchison R (2005) Pro DNS and BIND, APRESS
3. Snyder ME, Sundaram R, Thakur M (2009) Preprocessing DNS log data for effective data mining. IEEE ICC'09
4. Diana R (2009) Is google public DNS a marketing data warehouse? http://regulargeek.com/2009/12/04/is-google-public-dns-a-marketing-data-warehouse/. Dec 2009
5. Bilge L, Kirda E, Kruegel C, Balduzzi M (2011) EXPOSURE: finding malicious domains using passive DNS analysis. NDSS
6. Whyte D, Kranakis E, van Oorschot PC (2005) DNS-based detection of scanning worms in an enterprise network, NDSS
7. Choi H, Lee H, Lee H, Kim H (2007) Botnet Detection by Monitoring Group Activities in DNS Traffic. Comput Inf Technol CIT 2007:715–720
8. Postel J (1994) Domain name system structure and delegation. RFC 1591
9. White T (2012) Hadoop: the definitive guide. O'Reilly, Cambridge
10. Borhakur D (2007) The hadoop distributed file system: architecture and design

Chapter 72
Development of Bio-Signal Based Continuous Intensity Wearable Input Device

Jinuk Kim, Jihye Kwon, Subin Im, Seongyoon Lee and Jaehyo Kim

Abstract In this paper, instead of the input device that is limited to two-dimensional interfaces, we propose input device which is suitable for the virtual display and three-dimensional interfaces. We suggest the wearable device which is capable of 4-DOF continuous intensity input by using the signal extracted through the AHRS and the EMG electrode attached to the hand. By using this device, it is possible not only to represent the movement of the user's arm but also to reflect the intensity of holding the fist using the magnitude of the EMG. This interface is designed as a universal interface applicable to various displays, and in order to be applied to different locations and to be used in the same way, a wearable-type was selected. In this paper, we present the structure, shape and implementation of the interface and we evaluated the applicability of the device through experiments and simple applications based on MATLAB. Input devices proposed in this paper, was confirmed superior in ease of use and the position control performance when it is compared to the existing pointing input device. There is significance in utilizing as an input device, not only the interface of two-dimension, but the interface with ultra-large display such as a three-dimensional interface.

Keywords Wearable device · Human-machine interaction · Hand gesture · Bio-signal based Interface

72.1 Introduction

Display is being developed in various and complex forms. From representing in two-dimensional plane, now it includes three-dimensional plane and the virtual reality which is formless [1]. About the input device for display, its form is

J. Kim · J. Kwon · S. Im · S. Lee · J. Kim (✉)
Department of Mechanical and Control Engineering, Handong Global University, Pohang, South Korea
e-mail: jhkim@handong.edu

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_72, © Springer Science+Business Media Dordrecht 2014

changed variously, however, it mostly remains still in the concept of two-dimensional interface that extract human's position information using a camera or a motion sensor, and that information is on/off selective [2, 3].

Mouse is a typical input device, but cursor's movement is restricted to two-dimensional plane and it is not suitable for three-dimensional movement or virtual display. Especially, in a supersized display or non-planar display, it is not useful to operate in planar mapping [4]. Moreover, since it is not intuitive, it is not effective for learnability. And touch screen method which has become a popular input device with Smart devices can directly control the display and intuitively communicate so it is effective in bidirectional communication. However, it works with a 1:1 mapping with body, so it is available only to the certain accessible extent. Recently, input devices like Kinect that work as a natural interface on the basis of video or audio have been noted. But it is not appropriate for the use because if multiple users use it at the same time, interference occurs and recognition is not complete, and there is no generalized extraction rule of operation [5].

This paper proposes bio-signal based 4-DOF continuous intensity wearable input device using the user-oriented plane's Euler angle through AHRS attached to hand and EMG extracted on hand. This input device can extract angle's flexion–extension, adduction–abduction, and pronation-supination and it reflects the movement of arm or body in same direction. Also, using the EMG's magnitude from electrode attached on hand, it can measure the intensity of strength to put it or hold the fist. This interface is designed as a single input device in consideration of interaction with user-oriented various displays and as a wearable-type to applicable in the same way at the various place [6].

Most of the existing studies of the EMG-based interface are system that recognizes hand gestures from many electrodes that have adhered to the biceps [7, 8]. In addition, there was an attempt to determine flexion–extension and adduction-abduction by using the magnitude of the multi-channel EMG and to use it as the 2 DOF. However there are limits for use as an interface that it is difficult to attach the electrode to correct position of muscles which act independently of each other and when using the learning algorithm to the modeling, it takes longer train time [8]. For an interface we proposed, by selecting abductor digiti minimi muscle which is large muscle and located in Knife hand and producing an interface in glove form, it is possible to easily attach to the electrodes, and easily adjust the intensity in various levels through the strength of gripping fist.

We present the structure, form and implementation of the interface we propose and evaluate the possibility of use with a simple application based on MATLAB. We also plan to evaluate the performance of the interface by devising the experiment based on the Fitt's law and compare with other interfaces [9]. And we described about the technology necessary to commercialize the interface and discussed the form and work of display that can be effectively used with the interface.

72.2 Materials and Methods

To make a wearable input device capable of multi-dimensional continuous intensity input, we implemented by attaching AHRS to the back of the user's hand and measuring the yaw angle and pitch angle to present the movement of the arm and to make it possible to user to control the pointer intuitively. Further, not only to express the coordinates of the pointer, but also to be able to continuous intensity input we used EMG and roll angle. Such input elements can be utilized as functions like selecting or graphic data input [10]. In evaluation of interface performance, to simplify the process, input was used as in the form of on/off by applying threshold to the EMG. And in graphic input it is possible to represent variously by adjusting the thickness of the line in response to muscle activity. The suggested input device was confirmed whether it is suitable to be used as an input device by implementing the experiment of performance evaluation and application in MATLAB plot window. In this paper, first, we tested the input device through the experiment on the display of single computer.

72.2.1 System Summary

Figure 72.1 shows a series of processes for acquiring and processing the signals to implement the interface with movement of the arm and wrist. The position information of the mouse is represented by converting the yaw angle measured using AHRS to the coordinate in the horizontal direction of the display and by converting the pitch angle to the coordinate in the vertical direction [11]. By the previous studies this amplitude is known to be related to the muscle activation level [11–13]. In addition, by normalizing with the maximum strength that can be issued on a per-user basis and by dividing the muscle activation level relatively, users are be able to intuitively click or graphic data input according to the strength intensity [13].

72.2.2 Attitude Heading Reference System (Motion Sensor)

AHRS, the fusion of the 3-axis gyro sensor, acceleration sensor and geomagnetic sensor, is possible to output the position data of high accuracy and is used to obtain accurate position information. In this study, we received the position data with Euler angle (roll angle, pitch angle, yaw angle) and used the E2BOX's EBIMU-9DOFV2 product. The table below shows the specifications of the sensor (Table 72.1).

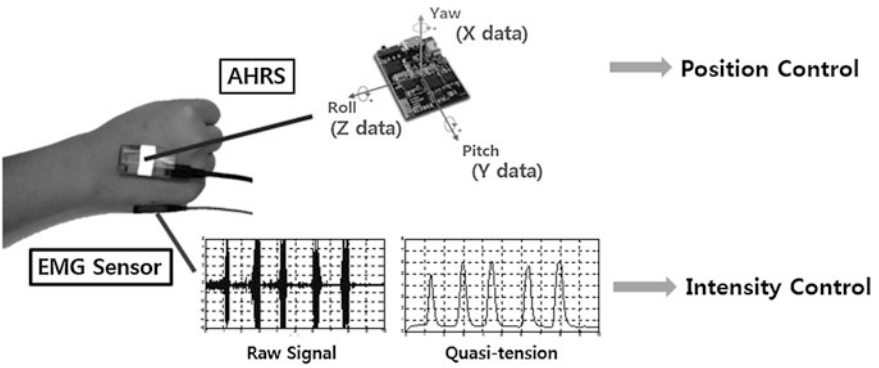


Fig. 72.1 System summary

Table 72.1 Specification of AHRS

Specification	EBIMU24GV2
Type	Wireless AHRS
Embedded sensor	Gyro sensor 3axis
	Acceleration sensor 3axis
	Geomagnetic 3axis
Information	Roll/Pitch/Yaw
Output speed	100 Hz
Wireless transmission	○
Motion capture	○
Output data	Euler angel
Offset	○
Output mode	ASCII
	HEX(binary)

72.2.3 EMG Sensor

EMG is an electrical signal detected at the surface of the skin when person makes movement, and it represents muscle activation level [11]. Quasi-tension is the signal obtained by extracting the amplitude of the EMG, and there is a high correlation between its result and the tension in actual muscle. So it is used to estimate the muscle activation level [11, 13]. For the muscle to measure the EMG, abductor digiti minimi muscle was selected which is located in hand knife. Because its muscle size is larger than that of the other muscles in the hand and its movement is more active, it is easy to measure the clearer signal. Further, when considering AHRS to be attached to the back of the hand to figure out the motion of the wrist, it is appropriate to select abductor digiti minimi muscle to minimize the size of the overall device (Figs. 72.2, 72.3).

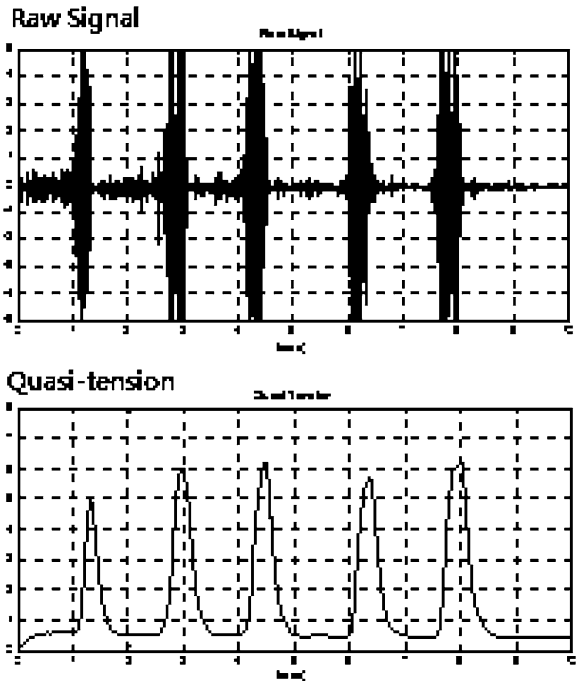


Fig. 72.2 EMG and Quasi-tension

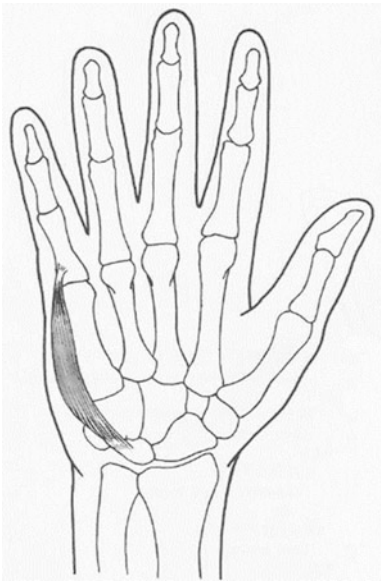


Fig. 72.3 Position of abductor digiti minimi

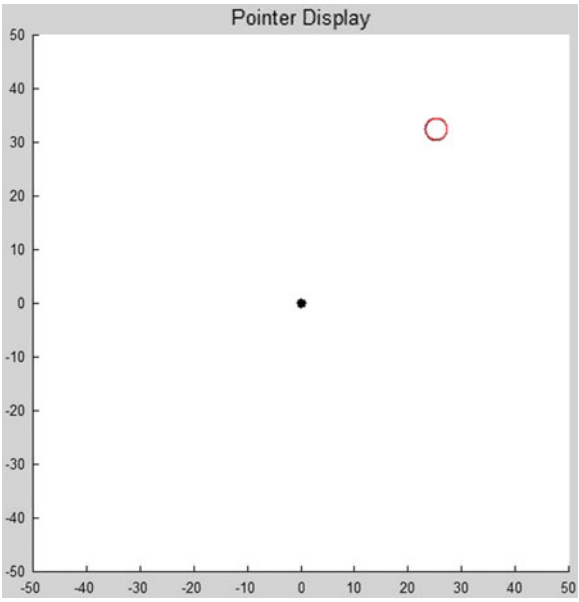


Fig. 72.4 Following the target game

72.2.4 Implementation of Interface

We implemented the game (Fig. 72.4) taking the target on the computer screen using an input device designed in this study and application (Fig. 72.5) able to draw using various line thickness. Interface to take the target is an interface that if the black circle pointer chases the red target circle and catches through click, target circle reappears randomly on the screen and this process is repeated. This can be used to assess to evaluate how can the click combined pointer position manipulation and EMG move the pointer quickly and intuitively. And the drawing application using various line thickness is designed to be drawn a fine line with weak power and a broad line with strong power by utilizing the strength to hold the fist. When use this interface, because the strength intensity and the thickness of the line are set to linearly proportional, it is possible to draw a picture in the sense using a brush.

72.2.5 Interface Evaluation

72.2.5.1 Evaluation Position Manipulation Performance of Interface

We analyzed the position manipulation performance of interface we suggested through the three evaluation criteria. To analyze the accuracy of the interface and the relation to movement velocity, we use Fitt's law, analysis of the movement

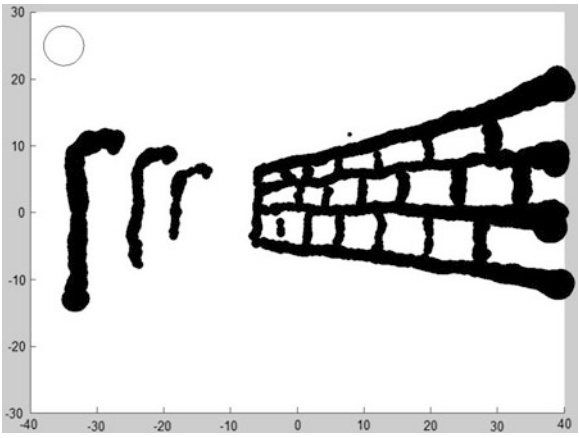


Fig. 72.5 Graphic input using EMG

path and repeated modified model. For evaluation, we present the experimental model in Fig. 72.6.

The experiment was designed as shown in the figure above, for each step, if the target is clicked by moving the cursor, the mission is completed. For a starting point, the target is displayed sequentially in the eight directions. Each mission is formed under the 9 steps with the distance of the object target and cursor (D) in three steps 20, 30, 40 and the width of the target (W) in three steps 2, 3, 4. Each step was measured with 3 subjects at 24.7 years average age repeated two times (Fig 72.7).

$$\mathrm{MT} = \mathrm{a} + \mathrm{b} \times ID, \; where \; ID = \log_2\left(\frac{D}{W} + 1\right) \tag{72.1}$$

First, an experiment was performed using Fitt's law to show that mission's index of difficulty (ID) and movement time (MT) is linearly proportional. Further, the accuracy in accordance with the index of difficulty was evaluated by using the integrated value of the movement path (d). In Crossman and Goodeve's repeated modified model, sub movement is formed about 30 % of the total movement distance [14]. Using it, we analyzed total time and total distance, and evaluated accuracy along the movement trajectory (Figs 72.8, 72.9).

72.2.5.2 Interface Performance Evaluation with EMG Intensity

In this study, experiments were carried out to evaluate whether it can be utilized as an interface by extracting the EMG in accordance with the strength of the force. If the size of the target circle and that of the circle reflecting the magnitude of EMG are same, this process is repeated changing to next goal. We confirmed the validity

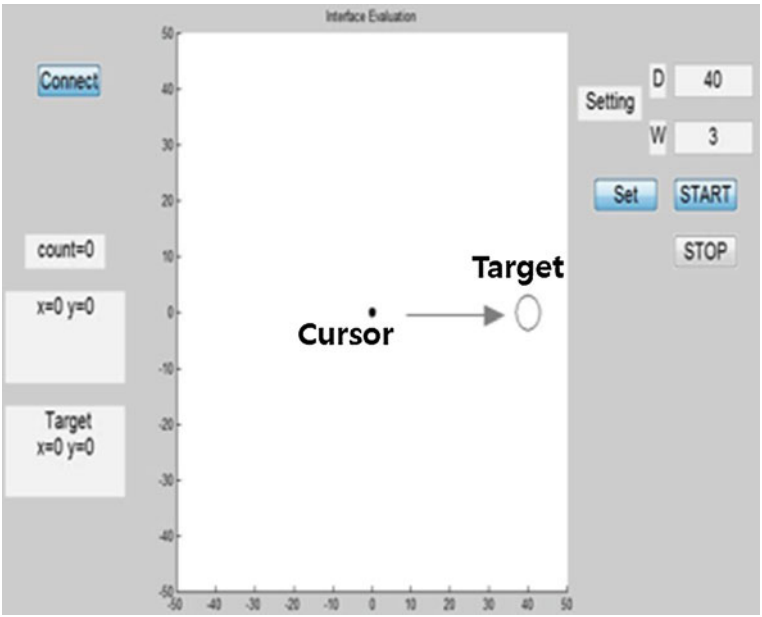


Fig. 72.6 Evaluation of position-control performance model

Fig. 72.7 Fitt's law

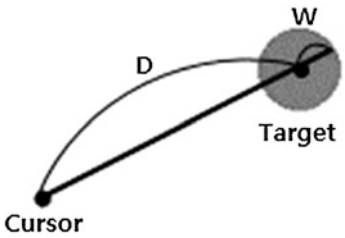


Fig. 72.8 Repeated modified model

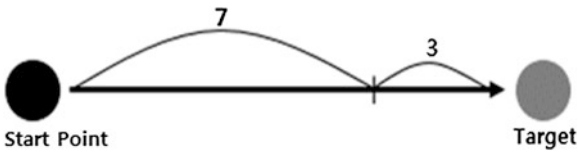
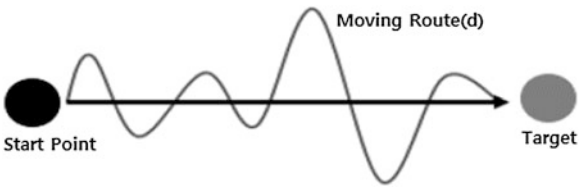


Fig. 72.9 Analyzing movement path



of the input value using the EMG by measuring the performance time between each step (Fig 72.10).

72.2.5.3 Experiment in Comparison with the Existing Pointing Input Device

In order to obtain the indicator for the objective evaluation of the input devices proposed in this paper, we conducted an experiment of comparison with the existing pointing input devices. We use two-dimensional mouse, three-dimensional mouse and wearable three-dimensional mouse as existing pointing input devices (Fig 72.11).

The existing input devices are used only for the position control. So, the experiment was carried out on the performance evaluation model of position control. In addition, the survey was also carried out, asking how comfortable or uncomfortable it was when it was used and how accurate it was running in accordance with the intention.

72.3 Experiment Result

72.3.1 Result of Position-Control Performance Evaluation

Figure 72.12 refers to the total moving distance to complete tasks along the size of the target circle (W) and the direction. We calculated the value obtained by curvilinear as the total moving distances because when the cursor makes the movement it travels from the starting point to the target point in various curves instead of moving in a straight line which is minimum travel path. In general, as the size of the circle becomes larger (namely the difficulty level becomes lower) the total moving distance is relatively reduced because as the size of the circle increases the cursor can be entered into the target circle relatively easily. Also, it is confirmed that the total moving distances to each direction were similar. This fact shows that there are not much difficulties experienced by the user when they move the cursor in all 8 directions. As shown in a graph of arrival time corresponding to the circle size in Fig. 72.13, the fact that the time to complete the tasks becomes shorter as the circle size becomes larger can be confirmed. This fact fulfills the Fitt's law which claims that the task completing time becomes shorter as the difficulty level becomes lower.

To assess the position-control performance of the interface more reliably, we separated the distance between the starting point and the target point in a ratio of 7:3, and we evaluated time and moving distance from the middle point (portion3) to the target point based on the repeat modified model. This experiment was conducted to verify the operability in performing accurate control when the target

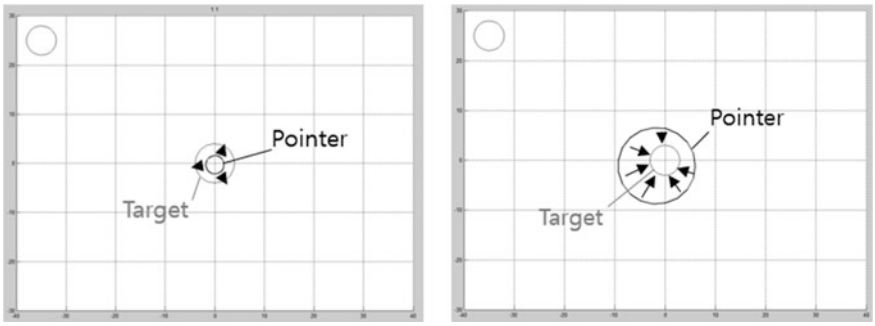


Fig. 72.10 Performance evaluation model with strength intensity

Fig. 72.11 Comparison with existing input device (2D mouse, 3D mouse, Wearable 3D mouse)

is close by. As it can be confirmed in Figs. 72.14 and 72.15, moving distance and time become small as the size of the circle increases. This is because in repeated modified model as the level of the sub modified model becomes low, the number of model is made smaller and it be could allow to position control more intuitively. It was confirmed that the moving distance and the time to complete the mission is reduced as the difficulty level is lowered by using the above results.

72.3.2 Performance Evaluation the Force Intensity

To figure out the characteristics of the EMG emerging in response to muscle activation level, after attaching the electrode on the abductor digiti minimi muscle, we extracted the 3 phases of force separately; weak strength, middle strength and strong strength by holding the fist separately. As Fig. 72.16 shows, the distinguishable magnitudes of the EMG were measured according to the force gripping fist. Because the fundamental magnitudes of EMG are vary from person to person, it should be determined the appropriate threshold for each user when using it as on/off function.

Next experiment was to check whether the arm can create continuous input when strength of the three phases were created and to check whether the arm can

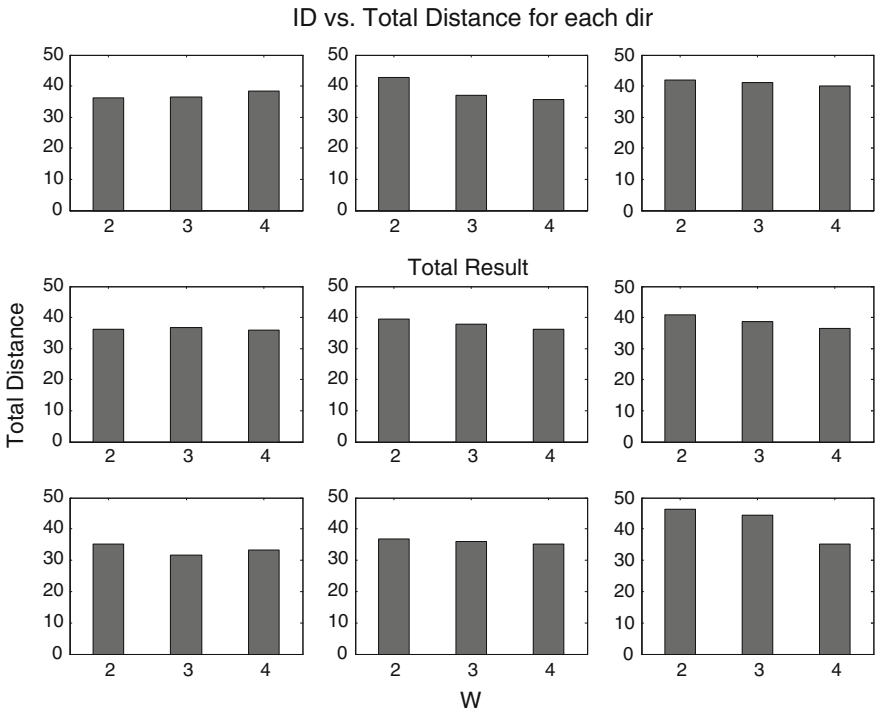
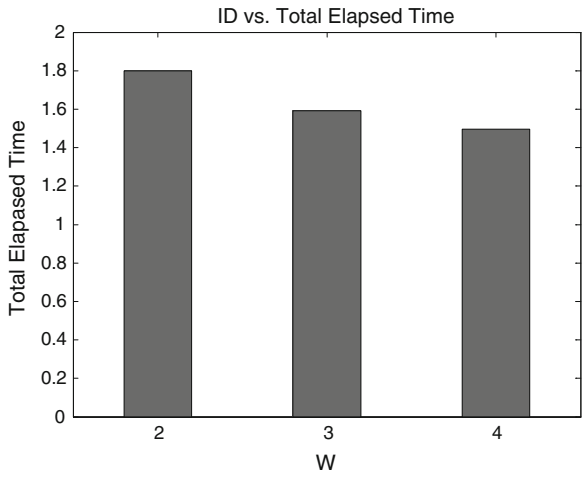


Fig. 72.12 Total moving distance along the directions

Fig. 72.13 Arrival time corresponding to the circle size



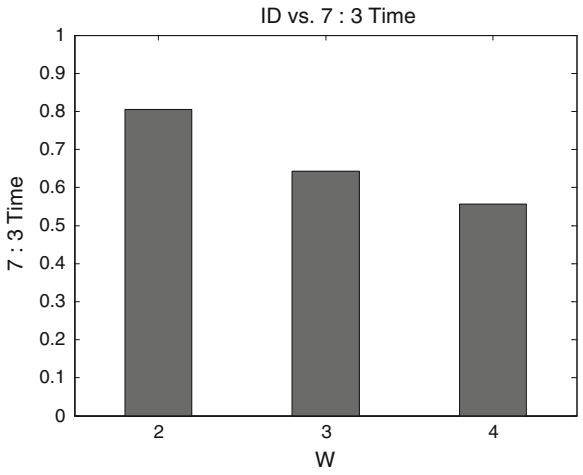


Fig. 72.14 Result of repeated modified model (time)

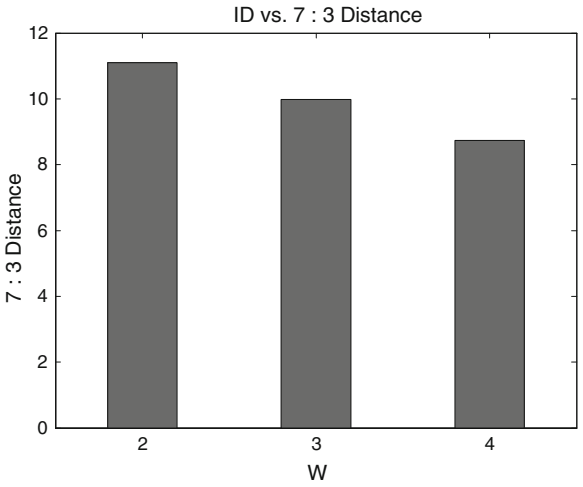


Fig. 72.15 Result of repeated modified model (distance)

put an input regardless of the motion of the arm. Here is the experiment result of EMG at each strength, weak, middle, and strong tension, maintained for 10 s in Fig. 72.17. In general, EMG at each state maintained its magnitude without changing even after 10 s. The magnitude of shaking is increased when it is tried to keep the high tension. According to Fig. 72.18, when three steps of strength was given, it could be confirmed that the EMG is similar in each state when the arm is not exercising comparison to when it is exercising. Therefore it could say that hand is independent of the movement of the arm and its activity continued with time is possible. As the consequence it is possible to get the stability as an interface.

Figures 72.19 and 72.20 is the result of an experiment to evaluate the strength adjustment ability of EMG. When evaluation on the experiment of intensity input

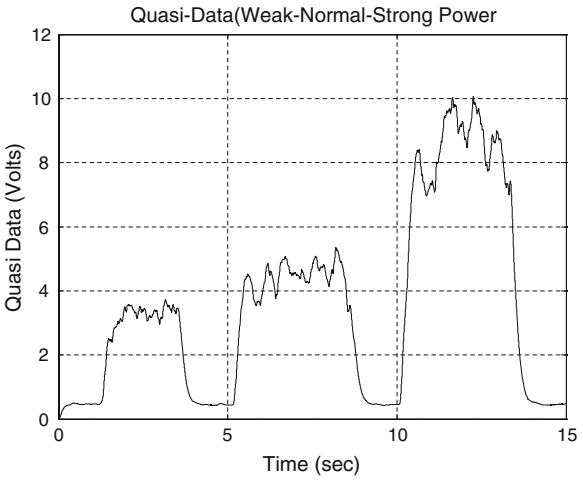


Fig. 72.16 Quasi-tension according to the strength intensity

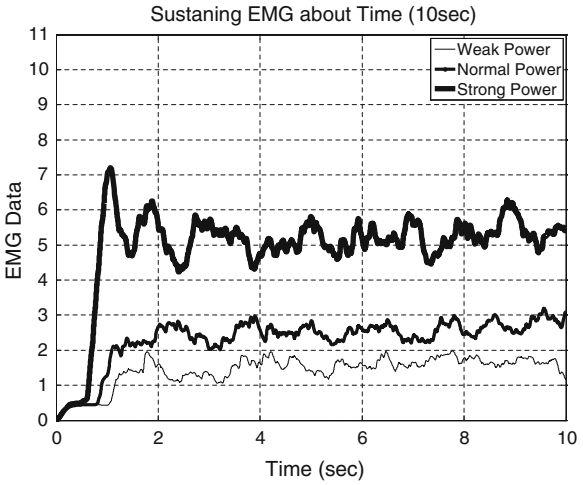


Fig. 72.17 Maintaining strength for 10 s

of EMG is performed, First Match Time means the time it takes to match circle of various sizes that appear in random and circle based on the strength of the EMG. As it can be confirmed in Fig. 72.19, The First Match Time increases as the size of the target circle increases. It tells us the fact that it takes longer time to complete the task when target circle becomes bigger and simultaneously the degree of difficulty increases. For the mission to be successful, after adjusting the target circle the user have to maintain its size between 1.5 s. Holding Time means the time taken by the time the successful to maintain size finally. According to Fig. 72.20, the Holding Time is increasing when the size of target circle increases. The user need to adjust the size of the circle when the size of the target increases, but time for maintaining the signal of constant magnitude continuously is

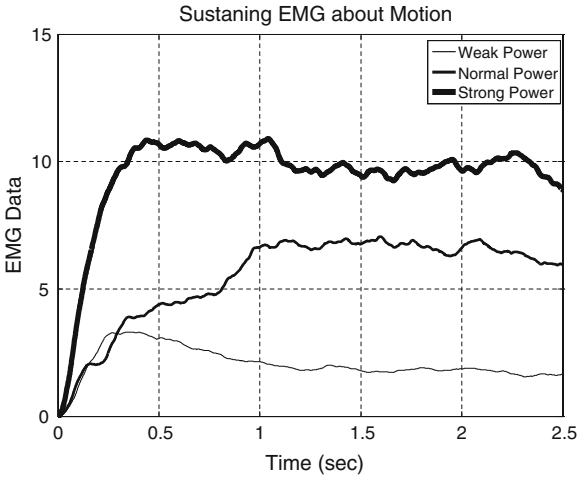


Fig. 72.18 Associated with movement of the arm

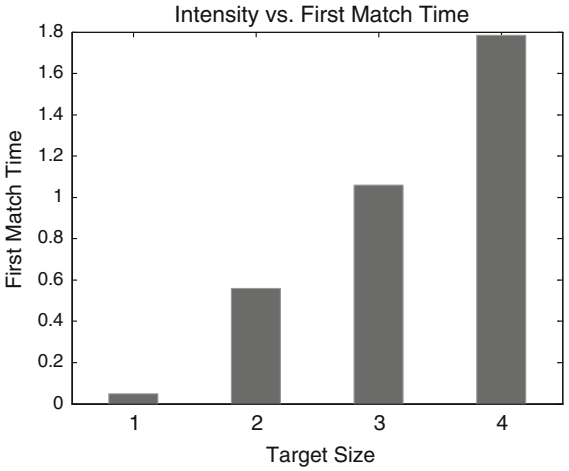


Fig. 72.19 Evaluation according the intensity (arrival time)

increased when giving a stronger strength. However, there is possibility for it to be used as interface because First Match Time only takes about 1.8 s even if the target is at the biggest.

72.3.3 Experiment in Comparison with the Existing Pointing Input Device

Figure 72.21 is the evaluation result of the comparative experiments with existing input devices. Results of the performance evaluation of the position control of the input device, the input device proposed in this study, it can be confirmed that the

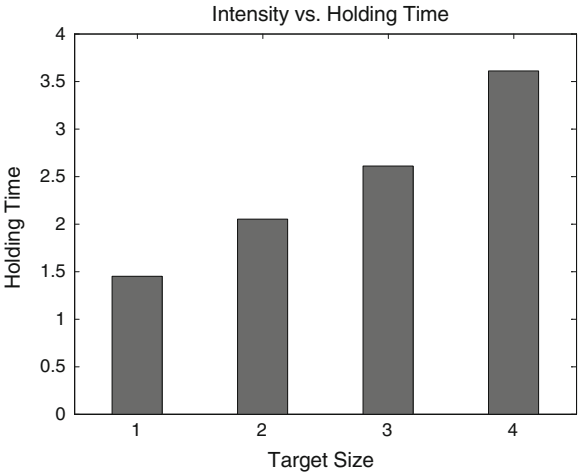


Fig. 72.20 Evaluation according the intensity (maintenance time)

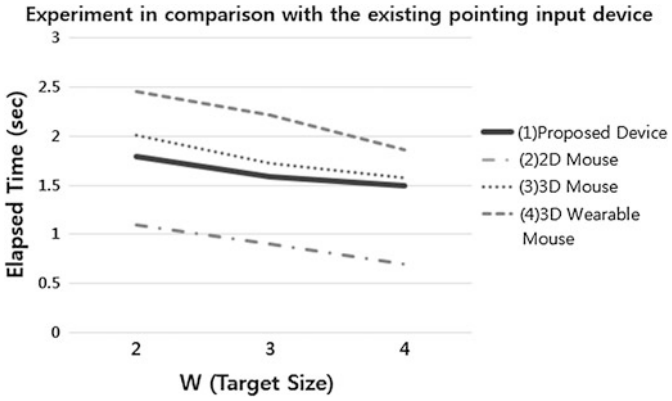


Fig. 72.21 Experimental results compared with the existing input device

time it takes to perform a performance evaluation model of position control is shorter than wearable 3D mouse and 3D mouse. Therefore it is possible to know the fact that the position control performance that is much better (Fig 72.22).

After finishing position control performance, experiment subjects were surveyed on the availability and usability. Through the survey it was confirmed that the convenience of the proposing input device was better or similar to the existing input devices.

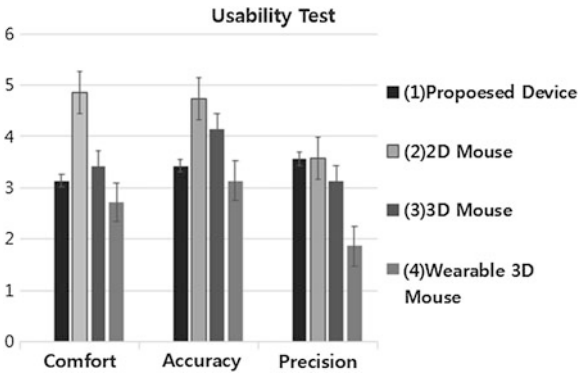


Fig. 72.22 Usability evaluation for pointing input device

72.4 Conclusion

In this paper, we suggested a wearable continuous intensity input device of 4-degrees of freedom utilizing the size of the EMG indicating muscle activation and position control using AHRS. We verified the practicality by devising the application, such as graphic art that is possible to adjust the thickness of the line and the target following game on the display of the computer, and we evaluated the performance in accordance with the intensity of EMG. Unlike previous studies which are utilized primarily discrete gesture based on pattern recognition to date, our study has advantage in terms of applying muscle activation. Further, results of evaluation on usability and position control performance of the existing pointing input device, we can know that the device's performance is better than that existing input devices. There is possibility in utilizing it as unified universal input device, not only for the interface of two-dimension, but also for the interface such as ultra-large display and three-dimensional interface. For it to become true the devised input device needs be compared with existing device for the further evaluation on the efficiency and usability. Also through the stabilization process of the signal of the position control, it is necessary to control position stably in the future. Lastly, design and evaluation for ease of use is required and for mobility and manipulation rules of the uniform by same way is also required when user change the location. Up until now PC-based test has been carried out so far, but for the future commercialization the development of the MCU-based input device is expected to be proceeded.

References

1. Park YW, Joo BK (2011) Overview to display next-generation. Optical science and technology, October 2011
2. Jo C, Park J, Uhm T (2009) Finter gestural interaction with a 3D display: solutions, evaluation and application scenarios. In: Proceedings of the korea multimedia society conference, Nov 2009, pp 551–554

3. Porta M (2002) Vision-based user interfaces: methods and applications. Int J Hum-Comput St 57(1):27–73
4. Suh K-Y, Kim M, Cho H, Park K (2013) Gesture-based user interaction techniques for tiled displays, HCI 2013, Korea
5. Kamata K, Maehara H, Wakimoto K, Usui S (2003) An approach to the design of information interface. In: Proceedings of the 19th symposium on human factors in telecommunications, 2003
6. Yang UY, Kim KH (2013) Personal wearable display technology, 2013 Electronics and telecommunications trends
7. Kim J, Mastnik S, Andre E (2008) EMG-based hand gesture recognition for realtime biosignal interfacing. In: Proceedings of the 13th international conference on intelligent user interfaces, pp 30–39
8. Oskoei MA, Hu H (2007) Myoelectric control systems: a survey. Biomed Signal Process 2:275–294
9. Fitts PM (1954) The information capacity of the human motor system in controlling the amplitude of movement. J Exp Psychol 47:381–391
10. Lee SY, Hong J-Y, Haeng-Suk, Han K-H (2007) Studies of the efficiency of wearable input interface. Sci Emot Sensibil 10(4):583–601
11. Koike Y, Kawato M (1995) Estimation of dynamic joint torques and trajectory formation from surface electromyography signals using a neural network model. Biol Cybern 73:291–300
12. Clancy EA, Bida O, Rancourt D (2006) Influence of advanced electromyogram (EMG) amplitude processors on EMG-to-Torque estimation during constant posture, force-varying contractions. J Biomech 39:2690–2698
13. Burden A, Bartlett R (1999) Normalization of EMG amplitude: evaluation and comparison of old and new methods. Med Eng Phys 21:247–257
14. Crossman ERFW, Goodeve PJ (1983) Feedback control of hand movement and Fitts' law. Q J Exp Psychol 35(2):251–278

Chapter 73
Speaker Adaptive Real-Time Korean Single Vowel Recognition for an Animation Producing

Sun-Min Whang, Bok-Hee Song and Han-Kyung Yun

Abstract Voice Recognition technique has been developed and it has been actively applied to various information devices in Korea such as smart phones and car navigation systems. Since the basic research technique related the speech recognition has been based on research results of other languages such as English and Japanese, it is possible to meet a sort of difficulties or some problems in point of view from the recognition. It should check once at least or a margin for applying the Korean vocal sound system to improve the recognition of Korean speech, 44 since Korean phonemes always have a same phonetic value. However, the scope of this study is the recognition of single vowels for a digital contents producing, particularly lip sync animation, since the lip sync producing generally requires tedious hand work of animators and it seriously affects the animation producing cost and development period to get a high quality of lip animation. In this research, a real time processed automatic lip sync algorithm for virtual characters as the animation key in digital contents is studied by considering Korean vocal sound system. The proposed algorithm contributes to produce a natural condonable lip animation with the lower producing cost and the shorter development period. The system of real time vowel recognition for producing digital contents focusing on formants frequencies is proposed. The recognition process consists of speech signal as the input, filtering, Fast Fourier Transform and identification. The algorithm based on the formant frequency using F1 and F2 was proposed, whose output was applied to the autonomic natural animating of the character' s mouth

S.-M. Whang (✉) · H.-K. Yun
School of Computer Science, Korea University of Technology and Education, Cheonan, South Korea
e-mail: smhwang@kut.ac.kr

H.-K. Yun
e-mail: hkyun@kut.ac.kr

B.-H. Song
Department of Industrial Design Engineering, Korea University of Technology and Education, Cheonan, South Korea
e-mail: bhsong@kut.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_73, © Springer Science+Business Media Dordrecht 2014

shape for small and medium sized animation productions or e-learning contents productions. The result shows the proposed speaker dependent single vowel recognition system is able to distinguish Korean single vowels from dialogue of a dubbing artist with real-time. The average of the recognition ratio was 97.3 % in the laboratory environment. It gives a possibility that the more condonable lip sync produces automatically without any animator involved.

Keywords Speaker dependent · Single vowel recognition · Speech recognition · Formants frequency · Real-time

73.1 Introduction

Movement of the lips and tongue during speech is an important component of facial animation to increase the immersion of users in digital contents. Mouth movement during speech is ongoing and relatively rapid, and the movement encompasses a number of visually distinct positions. Also, the shape of mouth must be synchronized to the dialogue. Since the most of energy and time duration of speech signals are accumulated in vowels, it is required to recognize vowels from the dialogue of dubbing artists to produce animation key for the lip shape of virtual characters. Previous results of researches related speech recognition, especially vowel recognitions, can be applied for synchronizing the mouth movement with a dialogue in digital contents such as animations and e-learning contents [1]. However, the principle techniques were mainly originated form the English-speaking world and the Japanese. Since the vocal sound system of Korean is different from other languages, it should be careful that the existing or previous technique is directly applied for recognizing Korean vowels. Furthermore, most of previous research results related Korean language were derived using the digital signal processing technique which treats Korean-speaking as an input sound signal without considering the characteristic of the Korean vocal sound system. There is possibility to improve the recognition algorithm and rate though the Korean phonemes always have the same phonetic description or value.

Vowel recognition has been widely studied as a main topic in the speech signal processing area in order to find effective their speech recognition. Most of studies used MFCC (Mel-Frequency Cepstral Coefficients) and LPC (Linear Predictive Coding) algorithm. The most popular spectral based parameter used in recognition approach is the MFCC. Due to its advantage of less complexity in implementation of feature extraction algorithm, only sixteen coefficients of MFCC corresponding to the Mel scale frequencies of speech cepstrum are extracted from spoken word samples in database. All extracted MFCC samples are then statistically analyzed for principal components, at least two dimensions minimally required in further recognition performance evaluation [2–4]. A technique for the vowel classification using linear prediction coefficient with combination of statistical approach and ANN (Artificial Neural Network) was proposed in [5]. A speech recognition

system using fuzzy matching method was presented which was implemented on PC. MFCC and LPC with ANN are among the most usable techniques for spoken language identification. MFCC and LPC methods are used for extracting features of a speech signal and ANN is used as the recognition and identification method. However, the most of energy and time duration of speech signals is accumulated in vowels and the shape of mouth generally depends on the vowel of dialogue, which is well known result of previous researches related of speech recognition. Our concern is only for Korean single vowels in Korean dialogue to apply an automatic mouth animation or facial animation for a multimedia contents production, since most of lip sync in digital contents are manually animated by animators according to hearing dialogue. It is one of extremely tedious and cost effective working to produce digital contents.

The proposed solution should meet the real time lip synchronization which is the lip shapes mainly depend on vowels not consonants same as the our previous study. And general pre processing speech analysis techniques are applied to recognize vowels without any specific instruments or tools to be utilized the solution at the small and medium sized animation producing studio or e-learning contents producing studio. To classify vowel, the result of FFT analysis is used to get the characteristic of each vowels. The incoming voice of a dubbing artist is analyzed the frequency component to extract vowels features to recognize vowels.

Since the mouth movement of virtual objects has to be synchronized with a dialogue exactly, lip-sync is one of tedious works for animators and a time consuming work. The mismatch or artificialness between mouth shape of characters and speaking reduces the immersion in the contents. This paper proposes the real time vowel recognition algorithm with the formant analysis which is a new technique to automatically perform lip synching for a computer generated character to match a real speech or a dialogue.

73.2 Related Works

73.2.1 Voice Recognition

The studies related voice recognitions were done by a lot of researchers or institutions [6–8]. They mainly have concerned the recognition words or speakers. Our concern is the real time vowels recognition to apply in a producing 3D animation. That should reduce the load of animators by minimizing simple monotonous hand working to synchronize with dialogues. Hence, the wanted recognition rate is not too high, which gives an affordable rate to be acceptable by producers at the current stage. But the lip sync solution should synchronize the lip animation of character to the dialogues with real time. One of simple application in Korea is the e-learning contents producing. Since the most of e-learning contents in Korea generally are taken recording of the lecturer's performing or the lecture in front of

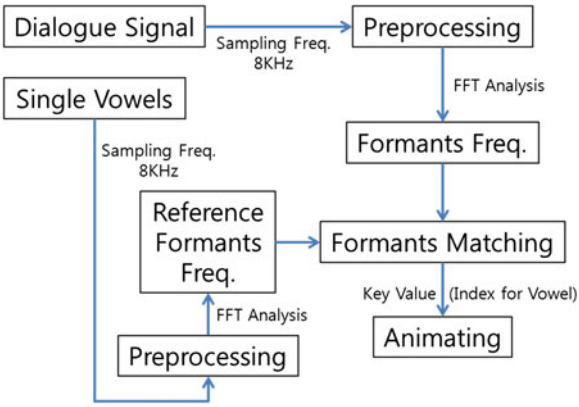


Fig. 73.1 Block diagram for lip sync solution

a blackboard. However, the proposed solution has a big additional advantage which is that the lecturer does not need to realize the camera, since some lectures repeatedly take pictures by their mistaking due to the their performing in the front of camera with trepidation. The proposed method will save the producing time and the cost. For the more, if the real human lecturer is replaced to a virtual lecturer in an e-learning contents, the contents has more freedom to increase the scholastic achievement by adding some animation effects such as resizing objects with real time rendering techniques.

The proposed solution is applied a general preprocessing for speech signals, a FFT analysis to get formants frequencies, comparing with reference data of the speaker to recognize vowels, converting the result of comparing with the reference table, and animating lip shape using the result as the key value for lip in the animation, as shown in Fig. 73.1. The first stage is the feature extracting of vowels for the lecturer before starting the lecture. Lecturer can record his or her voice using the installed microphone in a notebook computer or a desktop computer. The lecturer pronouns 5 times of each single vowels such as ㅏ[a], ㅔ[e], ㅣ[i], ㅗ[o], and ㅜ[u]. Then vowels' features of lecturer are extracted using FFT algorithm as a reference data with real-time. The sampling rate was 8 kHz and 1,024 FFT analysis were applied to get the formant frequencies as the references feature of each vowel since those are enough to show the mouth movement of the speaker if the Korean single vowels triangle is investigated. The next stage is the finding the reference component in the dialogues, which is same as the test of similarity comparison in the image processing. The result is converted to one of indexes of vowels for lip shape animation for the virtual character of lecturer. The next step goes over the subject in this study, which is the animating lip shape using the index as an animation key. This stage should require a real time rendering technique to get more natural facial expression.

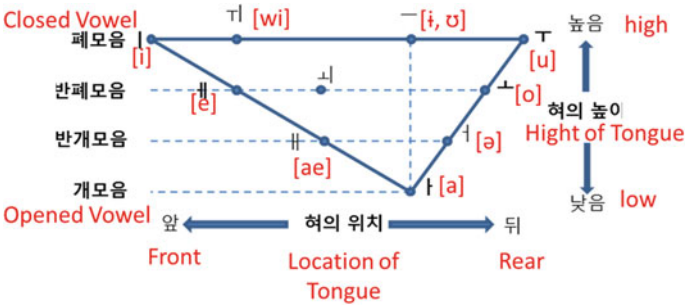


Fig. 73.2 Korean single vowel triangle

Table 73.1 Korean diphthongs

Group	Diphthongs
ㅣ[j]	ㅑ[ja], ㅕ[jə], ㅛ[jo], ㅠ[ju], ㅒ[jɛ], ㅖ[je]
ㅗ/ㅜ[w]	ㅘ[wa], ㅝ[wə], ㅙ[wɛ], ㅞ[we], ㅟ[wi]
-	ㅢ[ɨj or ʊj]

73.2.2 Characteristics of Korean Vocalization

A. *Vowel Vocal Sound System*

The phonetic description of Korean vowels are always invariable, which is a big different from English. The vowels are grouped as single vowels (monophthong) (Fig. 73.2) and diphthongs (Table 73.1) according to the Korean vocal sound system.

Korean single vowels are grouped the rounded vowel and the unrounded vowel by the mouth shapes during pronouncing. The sound of vowels is determined by the lip shape, and tongue's location and height as shown in Table 73.2.

B. *Characteristics of spoken Korean*

The normal reading speed of Korean is 348 SPM (Syllables Per Minute) and the fast reading speed is 426 SPM. The average of speed of spoken Korean is about 256 SPM and it varies from 118 to 409 SPM [9]. The 256 SPM means that 4.4 syllables are pronounced at every second or a syllable is pronounced for about 227 ms. It is equal to about 6.8 frames for a 30 fps (frames per second) animation. Since the maximum spoken speed is about 400 SPM, 6.7 or 6.8 syllables is pronounced per a second. If the sampling frequency is 8 kHz, the number samples is about 1,200 per a syllable. Hence, the real-time 1,024 FFT could be handle the frequency analysis of Spoken Korean.

Table 73.2 Lip shape and tongue's location of Korean single vowel

Max. point of tongue		Front vowel		Back vowel	
Shape of lip		Unrounded vowel	Rounded vowel	Unrounded vowel	Rounded vowel
Height of tongue	High	ㅣ [i]	ㅟ [wi]	ㅡ [ɨ, ʊ]	ㅜ [u]
	Middle	ㅖ [e]	ㅚ [oi]	ㅓ [ə]	ㅗ [o]
	Low	ㅐ [ae]		ㅏ [a]	

73.3 Experiment

The proposed single vowel recognition algorithm is for a natural and condonable lip animation which is automatically produced according to the dialogue of a dubbing artist. And the produced lip animation should minimize dis-matching dialogue with lip animation and human errors of animators.

73.3.1 Experimental Method and Process

Our vowel recognition processing strategy in this paper is using the formants analysis as a result of FFT which is commonly used for speech recognition. FFT is a basic tool for digital signal processing applicable for spectrum analysis. Another transformation is Discrete Cosine Transform. DCT is a discrete transform whose kernel is defined by the cosine function. It is not popular to use in speech recognition, since DCT does not produce clear efficient third formant F3 in speech recognition. It has been known for many years that formant frequencies are important in determining the phonetic content of speech sounds. Several authors have therefore investigated formant frequencies as speech recognition features, using various methods. However, using formants for recognition can sometimes cause problems, and formant frequencies cannot discriminate between speech sounds for which the main differences are unrelated to formants. Thus they are unable to distinguish between speech and silence or between vowels and weak fricative. However, our previous study [1, 11] has shown that the recognition of single vowel is applicable for animation of the mouth shape more natural and faster than manual work for lip sync. The important thing to performing accurate lip synching to the real continuous speech such as dialogue, is the preprocessing for speech signal and is the extraction of feature from speech signal. The first step (left limb in Fig. 73.1) is the feature extracting of vowels for speaker to prepare the reference of single vowels. Speaker phonate his or her voice using the microphone to process by a computer. Speaker pronouns each single vowels (e.g. 'a', 'e', 'e',

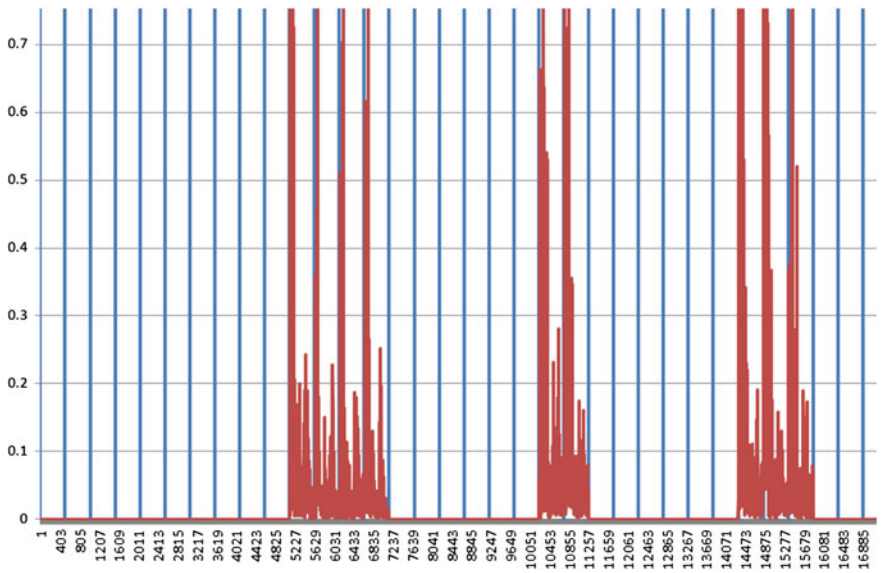


Fig. 73.3 Typical response of the FFT analysis for a sentence

'o', 'u') 5 times, then vowels' features of the speaker are extracted using FFT to get the frequency profile. The sampling rate is 8,000 Hz and 1,024 FFT analysis is applied to get the first and the second formant frequencies which are the reference feature of each vowel. The next step is finding the reference component from the dialogues, which is same as the similarity comparison test of the F1 and F2 for each vowel with the reference table of single vowels. The block diagram of the proposed system is shown in Fig. 73.1. Here, F1 and F2 formant frequencies are defined to calculate as follows:

$$A_k = \sum_{i=0}^{bandwidth-1} a(f_i) \tag{73.1}$$

where A_k is defined as the sum of amplitudes for the k-th band and our experimental band width is 40 Hz, that means the frequency band of FFT analysis is divided 100 for helping the real time processing. Since the range of index k is $1 \leq k \leq 100$, the index k of the first maximum A_k is defined as F1 and the second value of A_k is also defined F2. Then, each index and amplitude is stored for recognizing of vowels in dialogue as the reference which is compared with the incoming voice signal of the speaker (right limb in Fig. 73.1). A typical result of FFT analysis is shown in Fig. 73.3 which is short sentence, 'ga-ja ga a-ga' in Korean, same as 'Let's go kid' in English. The result shows that the proposed recognition algorithm is able to analyze frequency components of the incoming voice with real time. According to the output, the proposed FFT algorithm shows

Table 73.3 Recognition rate of the proposed algorithm

Single vowel	Recognition rate (%)
‘ㅏ’ [a]	98.5
‘ㅖ’[e]	98.6
‘ㅣ’[i]	97.2
‘ㅗ’[o]	96.3
‘ㅜ’[u]	96.1
Average	97.3

about 2 times per a syllable, which means the velocity of speaking is about 0.25 s per syllable. Therefore the speed of speaking should be normal since the average speaking velocity is 4.4 syllables per second or 0.23 s per syllable.

73.3.2 Experimental Result

The proposed algorithm was tested by 3 men subjects who were twenties whether the algorithm was working or not. They attended the experiment 3 times for 2 weeks and each of them made 9 set of experimental data. Each data set is made up reference and tested data since the formant frequency is possible to change according the physiological condition and the environment. Therefore the proposed method is a sort of limited speaker dependent recognition and takes reference data every time before speaking or dialogue to improve the recognition rate. As shown in Table 73.3, ‘ㅗ[o]’ and ‘ㅜ[u]’ are lower rate relatively. The reason is that they are the rounded back vowels and only difference is the height of tongue. If the experimental data is analyzed, ‘ㅗ[o]’ is recognized as ‘ㅜ[u]’ frequently and vice versa.

73.4 Conclusion

In this paper, a limited speaker dependent speech recognition solution for automatic producing lip-sync of objectives in digital contents. The conventional way to produce lip sync is to work manually by animators according to hearing of a dubbing artist’s recording. It is one of tedious works for animator and also requires a plenty of time to get a good quality of lip sync. The experimental result shows that our proposed method is able to recognize single vowels without any problem in real-time. Furthermore, our method has an advantage since they have the same lip shape in spite of the misrecognition between same group of vowels. Though the proposed solution is tested in the laboratory environment, the recognition of single

vowels in the recording studio does not induces any trouble to analyze the voice in real time since the recording studio has better equipment and facilities and the proposed system is isolated from the dialogue recording equipment to keep the sound quality of digital contents. The next experiment is a field test and the proposed system is applied to a real dialogue recording in a recording studio. It has to be confirmed whether the system processes in real time at real situation, not in the laboratory environment.

Korean language, Hangul, has also several diphthongs which have the mouth shape varying characteristic during phonating. However lip sync is relatively faster than other event in digital contents and it is only about 4 or 5 frames per syllable. Therefore the approximate mouth shape is sufficient to implement lip sync of objectives in animation. Similarly, the proposed system could be easily applied for any other language.

References

1. Hwang SM, Yun HK, Song BH (2013) Automatic lip sync solution for virtual characters in 3D animations. ICCT2013 2(1):432–433
2. Umesh S, Cohen L, Nelson D (1997) Frequency warping and speaker-normalization. In: ICASSP-97, IEEE international conference on acoustics, speech, and signal processing, vol 2, pp 983–986, April 1997
3. Umesh S, Cohen L, Nelson D (2002) Frequency warping and the mel scale. Sig Process Lett 9(3):104–107
4. Umesh S, Kumar SB, Vinay MK, Sharma R, Sinha R (2002) A simple approach to non-uniform vowel normalization. In: ICASSP 2012, IEEE international conference on acoustics, speech, and signal processing, vol 1, pp 517–520, May 2002
5. Paul AK, Das D, Kamal MM (2009) Bangla speech recognition system using LPC and ANN. In: Advanced in pattern recognition, ICAPR'09, pp 171–174
6. Kodandaramaiah GN, Giriprasad MN, Mukunda Rao M (2010) Independent speaker recognition for native english vowels. Int J Electron Eng Res Res India Publ 2(4):377–381
7. Kocharov DA (2004) Automatic vowel recognition in fluent speech (on the Materrial of the Russian Language). In: 9th conference Speech and Computer, SPECOM'2004 http://www.isca-speech.org/archive
8. Murakami T, Maruyama K, Minematsu N, Hirose K (2005) Japanese vowel recognition based on structural representation of speech. In: INTERSPEECH, pp 1261–1264
9. Shin Moonja, Han Sook-ja (2006) A study of rate and fluency in normal speaker. Speech Sci 10(2):159–168
10. Chung Hyun-yeol, Makino Shozo, Kido Keniti (1991) Analysis, perception and recognition of Korean vowels. ICEIC 91 2:195–198
11. Hwang SM, Yun HK (2013) Extraction for lip shape using real-time vowel recognition. In: Spring conference, KIIECT 2013, vol 6, no 1, pp 39–42
12. Moon Cho Sung (2003) An acoustic study of Korean vowel system. Korean Lang Cult 20:427–441

Chapter 74
Social Network Service Real Time Data Analysis Process Research

Yu-Jong Jang and Jin Kwak

Abstract Social network service contains various information of users. Because social network service is served as a web service type which displays data from the server on a web browser, only log record data of the web browser is left in the device. The demerit of such data is that it can be deleted easily by users; therefore, can't be used as an evidence. However, as long as the system is not shut downed, the data saved in the memory as a real time data remain so the trace of use of social network service exists as data. This thesis researches about how to collect and analyze data efficiently.

Keywords Social network service · Digital forensic · Data analysis

74.1 Introduction

Recent Social Network Service has increased its utilization rate drastically with the increase of distribution rate of Smart Phone and now is positioning itself as a way of communication with other and expression of oneself. SNS contains diverse information such as content of communication between users, location of users, networking, and the mentality of users. Through digital forensic, such information can be used as useful evidence in an investigation of case. However, unlike computer service saving data in a device to provide its service, SNS provides its

Y.-J. Jang
ISAA Lab Department of Information Security Engineering, Soonchunhyang University, Asan, South Korea
e-mail: yjjang@sch.ac.kr

J. Kwak (✉)
Department of Information Security Engineering, Soonchunhyang University, Asan, South Korea
e-mail: jkwak@sch.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_74, © Springer Science+Business Media Dordrecht 2014

service with real time synchronization of data with a server. The computer device used for SNS saves the SNS log record of the web browser. But such information can be deleted by the users and the information doesn't contain specific details of use of SNS. Besides, an investigator can't avoid difficulties in processing the investigation with deleted data.

Real time data contain the record of the web browsers which used SNS before the shutdown of computer device. This kind of data cannot be deleted freely by users and the real time data still remain in a memory unless it shuts down; therefore, the data can be used in the investigation through an analysis. This thesis studied about the process of collecting and analyzing such data.

74.2 The Need of Digital Forensic of SNS Real Time Data

SNS contains diverse information of users. Such information consists of personal data like schedule of user, conversation with friends and current location. In an investigation of case, personal data of suspect saved in SNS can help through many uses like an identification of living pattern, geographical location, and comprehension of idea and mental state.

However, the difficulty follows in data collection merely by collecting data of personal information section (Personal Page) of users of SNS in order to gather data useful for a case investigation in SNS.

- Data closed setting: Closed data of suspect can't be accessed by open account
- Data editable by many: Number of users can edit the data of suspect
- Inefficient web browser record to be used as a digital evidence.

74.3 SNS Investigation Procedure by Utilizing Real Time Memory Analysis

To collect the real time data from the device using SNS, one must be able to collect live system information of the device using SNS. Real time data can't be recovered once the system is shut down. That is why the device must be confiscated and it should collect evidence data in the activated system [6]. Figure 74.1 is the procedure of real time memory analysis in the investigation process which this thesis suggests. This procedure only indicates the investigation preparation stage, response stage, digital evidence securement and collection stage. The other procedure, evidence transportation and confirmation procedure will be handled in same process of existing forensic model.

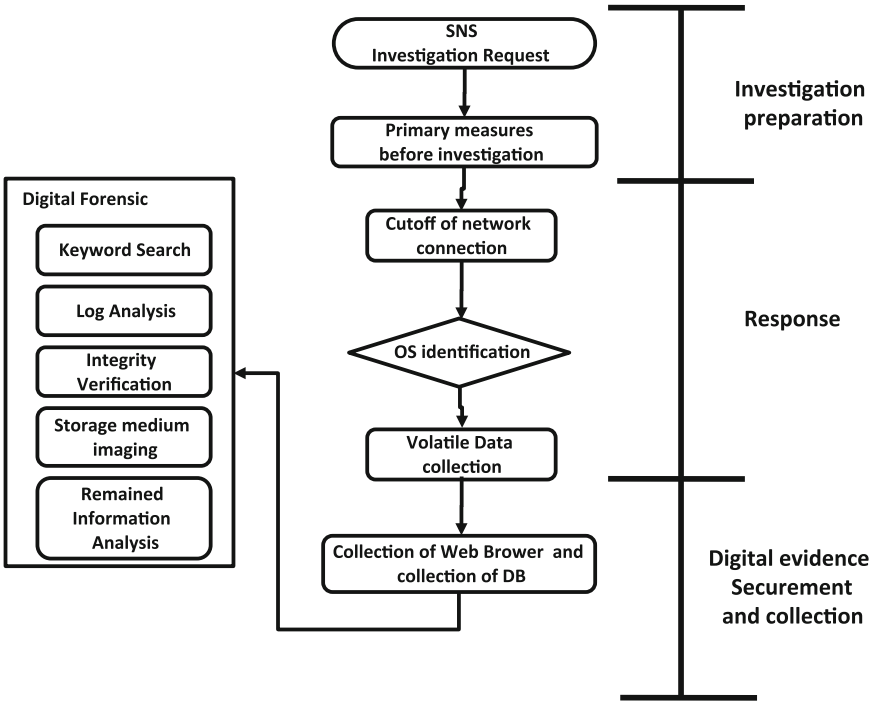


Fig. 74.1 SNS digital forensic investigation procedure

74.3.1 Investigation Preparation

There is a need of preparation before undergoing digital forensic stage about digital device. Such stage will be defined as the digital device and digital forensic.

- Identification of SNS in use

When request of investigation about SNS is assigned to an investigator, the investigator needs to identify SNS used by the suspect. Through this, cooperation for the confiscation of data of the suspect can be demanded from the service company providing the SNS.

When the disk stored the data amid the procedure of general digital forensic is confiscated, it should block the access to data storage of SNS as if using writing protection device so no modification can be done.

- Acquisition of suspect's account

Among the primary measures of SNS investigation, the acquisition process of suspect's account is a process of acquiring SNS account of the suspect through an interrogation. Undergoing this process allows a quick access to the personal information of the suspect and the content created in SNS.

- Information collection using SNS

After the identification of SNS being used by the suspect and the acquisition of suspect's account, the investigator starts to collect information about the type of SNS the suspect uses and his account. Such process collects information of the device before a digital forensic like SNS use timeline, networking, location of the suspect to be compared and analyzed with digital information gathered in the process of a digital forensic.

74.3.2 Response

In case investigation, the initially taken action by an investigator in the field is called Response. This is a process being conducted when the investigator arrives at the site after finishing preparation for the investigation in the procedure of digital forensic about SNS.

- Cutoff of network connection

The devices collected by the investigator must be disconnected from network not to allow any modification of stored information about SNS. This is because it is a real time network service being provided with a service in real time through stored data in SNS server via a network, not a storage type service to provide a service through stored information inside of existing digital divice.

74.3.3 Digital Evidence Securement and Collection

Evidence-analysis-purposed imaging work must be processed for the conservation of original and further analysis when digital evidences are collected. The imaging work of storage medium, as a process to prevent a damage to the original data, is an important initial stage of digital forensic performance procedure.

The equipment or software being used in the storage medium image process print a hash value as it creates a copy from original evidence or image, it must be recorded with the process of evidence collection to secure the integrity and reliability.

74.3.4 Delivering Evidences and Confirmation

Transportation of digital evidence is significant for digital evidences can be damaged or modified easily. Until the moment of a submission of the original as an evidence to analysis office or as the evidence, one must exert an effort more carefully so the copy will be transported safely.

74.3.5 Investigation and Analysis

The procedure of securement and collection of digital evidence is same with the standard of existing investigation. It applies the existing digital forensic process, and it constitutes according to the characteristic of SNS, not the conventional method and algorithm in each procedure.

74.4 Real Time Memory Analysis Technique

74.4.1 Real Time Data Investigation Procedure

The memory information to be used an evidence in this thesis is a volatile data of live system. To use the volatile data which is prone to be damaged as not damaged evidence, the analysis must be performed with the memory as it was when taken as an evidence. To perform such analysis, the device must be confiscated, and the memory analysis must be conducted as imagificating the live system information through memory dump process which creates single image file out of system memory.

Figure 74.2 shows imagificating the physical memory of live system to memory dump.mdd of 3,319.36 MB

74.4.2 Real Time Data Collection Procedure

74.4.2.1 Process List Extraction

To analyze the imagified memory in the live system about SNS data, one must look for a trace of use of SNS in memory. Computer device, unlike a mobile device using application, use a web browser to use SNS. That is why, to analyze the trace of use of SNS, one must find a memory address that uses the web browser. In this thesis, Volatitlity, a memory analysis tool, is used to extract a process list being used from dumped memory image. Figure 74.3 shows a memory offset value and operation hour of process ID (PID) and parent process ID (PPID). This live system shows a feature of using Chrome and Explore.

Figure 74.3 shows the operation hours of memory offset value, process ID (PID) and a parent process ID (PPID) which web browsers use in the dumped live system money use list. This live system shows the use of Chrome and Explorer web browser.

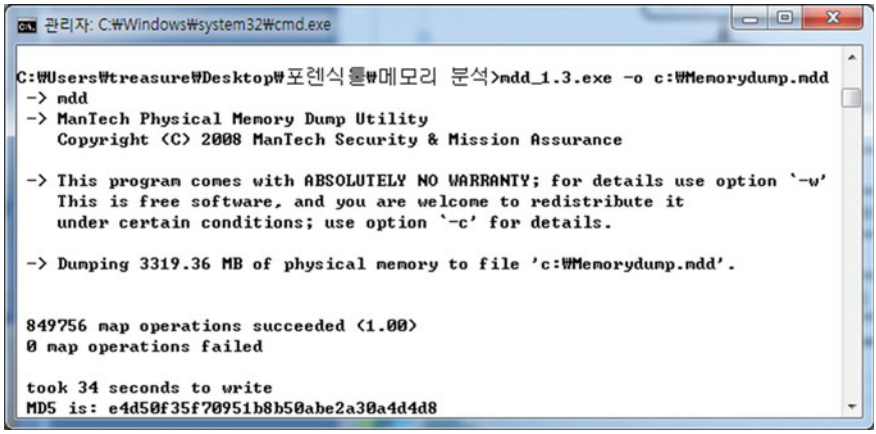


Fig. 74.2 Memory dump process

Offset(V)	Name	PID	PPID	Thds	Hnds	Sess	Wow64	Start
0x87a39030	iexplore.exe	5560	2360	13	573	1	0	2013-09-10 05:35:03
0x85bldaf8	iexplore.exe	1164	5560	29	1003	1	0	2013-09-10 05:35:04
0x85cc7030	chrome.exe	6140	2360	40	1179	1	0	2013-09-10 05:35:28
0x85d68030	chrome.exe	6060	6140	13	315	1	0	2013-09-10 05:35:32
0x85d74458	chrome.exe	2764	6140	13	238	1	0	2013-09-10 05:35:37
0x8868dd40	jucheck.exe	4624	2640	6	405	1	0	2013-09-10 05:35:38
0x85b42540	chrome.exe	2444	6140	16	282	1	0	2013-09-10 05:35:45
0x85bc4708	chrome.exe	1544	6140	24	295	1	0	2013-09-10 05:36:04

Fig. 74.3 Web browser memory use list

74.4.2.2 Memory Imagification for Process

In order to analyze the memory in use by a web browser, there is a need of extraction and analysis process of the previously analyzed assigned memory part of web browser process in the process of process list extraction.

Figure 74.4 shows the process of image dump by assigned process memory such as PID 5560. 1164.6140, and 6060 of Chrome and Explore.

74.4.3 Real Time Data Analysis

In the dump file that has been imagified per process of web browser, there is a used volatile data when operating web browser. Such volatile data is data which can't be modified by the suspect or other in SNS. Moreover, even after the deletion of data shared in SNS by the suspect or others, there is a case of data still remaining in the memory. For such case, it can provide an useful information in the investigation through real time data analysis.

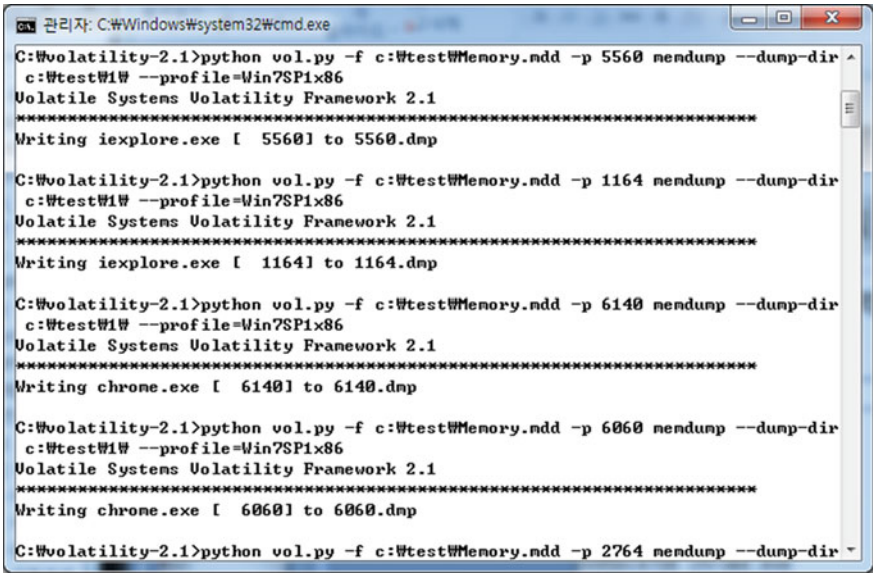


Fig. 74.4 Image of process use in web browser

Real time data analysis can be done through searching related data of SNS from the stored memory data in the image dump file of process memory of web browser.

Figure 74.5 shows the trace of SNS use by a process PID:2764 memory image dump file of Chrome web browser.

- SNS user networking information

Through real time data analysis of SNS, analysis of networking of a suspect that is stored in the volatile memory is possible. When a computer device is being used with a web browser, one can search the profile of SNS users which the suspect visited.

Figure 74.6 shows the profile address of SNS users which one visited by searching a profile statement that presents the profile page of user of Facebook in a process PID:2764 memory image dump file of Chrome web browser.

- Information related to a conversation of SNS users

Through a real time data analysis of SNS, one can analyze the conversation of the suspect with other SNS users with the data stored in volatile memory. In case of a conversation using Web browser with others in SNS, acquisition of any evidence is difficult for users can easily delete it. However, the memory, a real time data, stores the content of conversation, so, as long as the computer device is not shut down, there still remains a conversation record which can be an evidence.

Figure 74.7 shows a collected content of the conversation between profile ID: 10003356748509 with the user of targeted device for analysis at 13792052125 time by searching msg to search the conversation of Facebook in a process

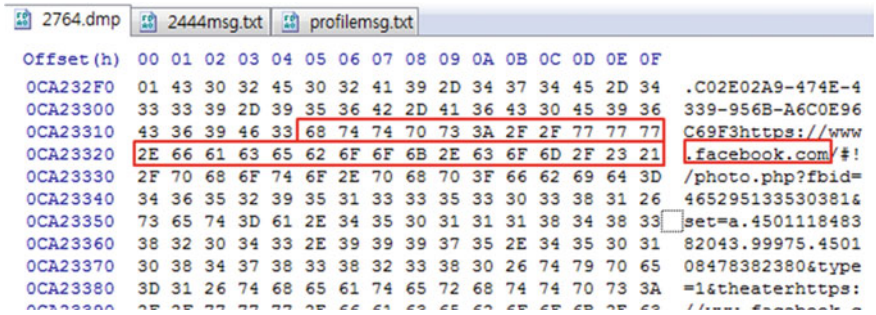


Fig. 74.5 Trace of SNS use in web browser

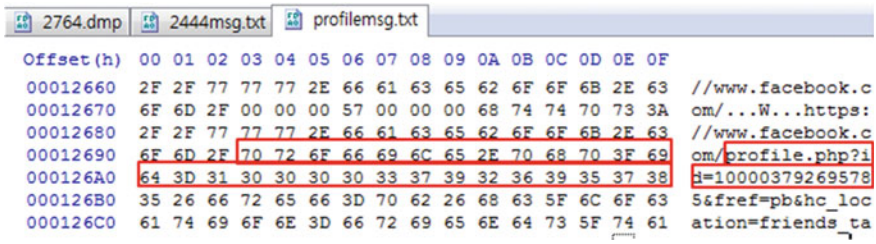


Fig. 74.6 Visited users' profile ID

PID:2764 memory image dump file of Chrome web browser. Since web browser displays a recorded time after 00 h 00 min 00 s of January 1, 1601, it is expressed as 13792052125

- Timeline information of SNS user

Through a real time data analysis of web browser, one can analyze not only networking information of SNS user and related data about SNS conversation, but also timeline of SNS with a web browser. Using such information to compare with cash information and history of web browser enables to investigate a record deleted by the user or others.

74.5 Expectancy Effect

SNS real time data analysis method which is discussed in this thesis enables an effective investigation in SNS environment.

Table 74.1 is a table which arranges a search probability of analyzable digital evidences in accordance of analysis environment. Through this table, the digital forensic by memory analysis is not greatly affected by network environment or data deletion than the information which can be gathered from the existing Web, Mobile devices.

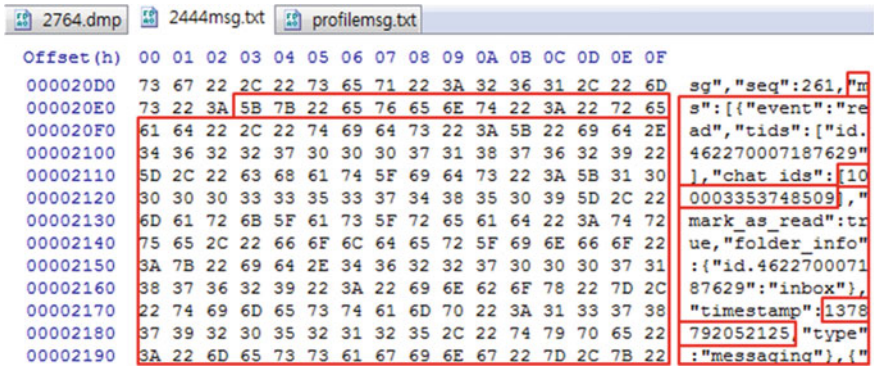


Fig. 74.7 Record of conversation of users

Table 74.1 Possibility of searching SNS digital evidence according to an environment

Classification	Network connection			Network cutoff			Deletion of general data		
	Web	Mobile	Memory	Web	Mobile	Memory	Web	Mobile	Memory
Human networking	O	O	O	X	X	O	X	O	O
Location data	O	O	O	X	X	O	X	X	O
Closed post	X	X	O	X	X	O	X	X	O
Secret information	X	X	O	X	X	O	X	X	O

X searching is impossible, *O* searching is possible

Following effects can be expected through this forensic procedure of social network service

- Preparation for any confusion in an investigation process by systemization of investigation procedure for social network service.
- Assurance of an expedite collection of evidence and integrity with appropriate response in accordance of environment of social network service for a collection of digital evidences.
- Systematic collection of digital evidence through analysis of relevance of various devices, files and information of a web browser by the effective analysis procedure of digital evidence in social network service.

74.6 Conclusion

SNS contains various kinds of information such as personal information of user, location data, and networking. For SNS is served in a form of web service type which displays a data received from a server, only log record data for the use of web browser exists in a device. Such data can be easily deleted by users. In

addition, the demerit of deleted data is being not able to serve as evidence. However, the data of SNS use stored in a memory remains as it is unless the system is shut down. This thesis studied about the way of effective collection and analysis of such data. The SNS digital forensic procedure suggested by this thesis includes general digital forensic procedure and partial addition of a process and through this, it can be applied without any change of overall algorithm. With the suggestion of this thesis, it is expected to use the information of SNS in a case investigation.

Acknowledgments This research was funded by the MSIP (Ministry of Science, ICT and Future Planning), Korea in the ICT R&D program 2013. This work was supported by the Soonchunhyang University Research Fund.

References

1. Jang Y, Kwak J Mobile digital forensic procedure for social network service. J Korea Navig Inst 17(3):325–331
2. Kim I (2013) Crime tail caught in the SNS, etnews. http://www.etnews.com, Apr 2013
3. Lim S (2011) Wife murdering professor complicity with the Mistress. MBC NEWS, May 2011
4. Choi J (2013) Boston terrorist suspects captured? Seoul Newspaper, April 2013
5. Lee JH, Jung H, Lee S (2012) Forensic investigation procedure for real-time synchronization service. J Korea Inst Inf Secur Cryptol 22(6):1363–1374
6. Lee SH, Kim H, Lee S, Lim J (2006) A study of memory information collection and analysis in a view of digital forensic in window system. J Korea Inst Inf Secur Cryptol 11(3):87–96

Chapter 75
A Database Design for Pre-ordering System Based on Prediction of Customer Arrival Time

Aziz Nasridinov, Sangeun Yoon and Young-Ho Park

Abstract Recently, with a busy lifestyle of the people, we can witness that majority of them spend a long waiting time in lines of the restaurants and take-out coffee shops. As modern people should solve many problems in a limited time, systems that takes the pre-orders at the restaurants and take-out coffee shops are necessary. However, currently there is a small number of systems that efficiently support the pre-order services. In this paper, taking into consideration that smartphone services have become an integral part of the modern people's life, we propose a database design for mobile based pre-order system. The proposed system effectively uses the customer's history and efficiently handles the pre-orders. Specifically, in this paper we propose a database design to predict the customer's arrival time based on customer's past experience. By using our system, modern people can reduce the wasting of unnecessary time at the restaurants and take-out coffee shops. Moreover, the proposed database design can serve as a blueprint for other pre-ordering systems in the future.

Keywords Pre-order service · Prediction of arrival time · e-r diagram

A. Nasridinov (✉) · S. Yoon · Y.-H. Park
Department of Multimedia Science, Sookmyung Women's University,
Cheongpa-ro 47-gil 100, Yongsan-Ku, Seoul 140-742, South Korea
e-mail: aziz@sookmyung.ac.kr

S. Yoon
e-mail: kb994346@hanmail.net

Y.-H. Park
e-mail: yhpark@sookmyung.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_75, © Springer Science+Business Media Dordrecht 2014

75.1 Introduction

Recently, people are becoming more busy due to the recession putting pressure on companies and jobs [1]. Such a busy lifestyle leads to less time to be spend on other daily activities, such as dining at the restaurants or spending time in coffee shops. On the other hand, the rapid development of smartphones and tablet PCs has helped people to manage their time more efficiently. However, even with such devices, we can witness that majority of people spend a long waiting time in lines of the restaurants and take-out coffee shops [2, 3]. As modern people should solve many problems in a limited time, systems that takes the pre-orders at the restaurants and take-out coffee shops should be developed. Currently, web or mobile based pre-order services are simple and do not provide an active customer support. In contrast, this paper proposes a database design for a pre-order service that predicts customer's arrival time to the shop, while considering various situations of customer, such as traffic.

The prediction of customer's arrival time is especially important in coffee shops [4]. There have been many smartphone applications that operates in the field of coffee shops, such as With Coffee [5]. With Coffee application proposes various information related to the coffee shop franchise and the menus that they offer. It also proposes a function to search the coffee types. Another widely used pre-order system is pizza ordering service. Domino pizza [6], one of the pizza franchise, is operating a service in which the customers make an order through web or mobile applications and come to the pizza shop to pick up the order. This system requires the customer to enter his/her arrival time and according to it, the system makes a list of orders by their priority. However, existing approaches have many limitations. First, most services are simple and do not provide an active customer support. Second, the existing services do not consider various customers situations. For example, customer may stuck in the traffic while coming to pick the item. Third, for such situations as traffic, there is no active connection between customer and service. In other words, once the order is received, the service prepares the order as it is, unless the customer lets the service know about his/her delay in arrival.

In this paper, taking into consideration that smartphone services have become an integral part of the modern people's life, we propose a database design for mobile based pre-order system implemented at the coffee shop. The proposed system effectively uses the customer's history and efficiently handles the pre-orders. Specifically, in this paper we propose a database design to predict the customer's arrival time based on customer's past experience. By using our system, modern people can reduce the wasting of unnecessary time at the restaurants and take-out coffee shops. Moreover, the proposed database design can serve as a blueprint for other pre-ordering systems in the future.

The rest of the paper is proceeds as follows. Section 75.2 discusses related study. Section 75.3 presents the proposed method. Section 75.4 highlights conclusions.

75.2 Related Study

In this section, we explain the features of the proposed method by comparing it to the related methods.

There are many services that are able to predict customer's arrival time. Such services include car navigation that lets you know the approximate arrival time, and bus, subway and taxi services. However, such systems are single customer oriented, and thus, cannot handle several requests at the same time. In contrast, the proposed method can handle many requests at the same.

The prediction of customer's arrival time is especially important in coffee shops. For example, there are many customers, who want their coffee or tea to be hot when they arrive to the coffee shop. Thus, there have been many smartphone applications that operates in the field of coffee shops, such as With Coffee [4]. With Coffee application proposes various information related to coffee shop franchise and the menus that they offer. It also proposes a function to search the coffee types. Comparing to With Coffee, rather than simply proposing various information related to coffee shop, our method proposes a pre-order system that takes the customer's orders in advance and reduces customer's waiting time.

One of the widely used pre-order system is pizza ordering service. In pizza ordering services, customer usually makes the order through web or mobile applications. Similar to pizza ordering services, most of the pre-order systems prepare the service once the customer makes the order, rather than matching customer's various situations. Comparing to these systems, in our method we predict customer's arrival time and prepare the item without making the customer wait. Recently, Domino pizza, one of the pizza franchise, is operating a service in which customers make an order through web or mobile applications and come to the pizza shop to pick up the order. This system requires the customer to enter his/her arrival time and according to it, the system makes a list of orders by their priority. However, there is no connection between pizza ordering systems and customer after order has been made. In contrast, the proposed method provides an active communication between customer and pre-order system after the order has been made.

However, existing approaches have many limitations. First, most services are simple and do not provide an active customer support. Second, the existing services do not consider various customers situations. For example, customer may stuck in the traffic while coming to pick the item. Third, for such situations as traffic, there is no active connection between customer and service. In other words, once the order is received, it prepares the order unless the customer lets the service know about his/her delay in arrival. In order to overcome such limitations, in this paper, we propose a database design for pre-order system that estimates the customer's arrival time to the coffee shop, matches various customer's situations and offers an active connection between customer and service.

75.3 Proposed Method

In this section, we give the overview to the proposed system, describe it in detail and explain the database scheme used in the proposed system.

75.3.1 Overview of Mobile System

This subsection gives the overall overview to the proposed system. The proposed system is a pre-order system implemented at the coffee shop. It takes an order from the customer and estimates the customer's arrival time while considering various situations of customer. In order to achieve this, we use customer's current location and time information, and customer's past experience. One of the advantages of the proposed method is that it is useful for coffee shop sellers too. Because, it proposes an interactive map, which indicates the current location of the customer and how much time it requires for customer to arrive to the coffee shop. Thus, by using the proposed system, customer can reduce the wasting of unnecessary time at the restaurants and take-out coffee shops, while coffee shop sellers can prepare fresh items prior to customer arrival.

The proposed pre-order system is depicted in Fig. 75.1. It has two main participants, such as clients and server.

- Through socket programming, a mobile client transmits various information to the server, such as customer's information, order information, location and time. On the other hand, a web client connects to the server through the web interface and enters the details of his/her order.
- Server connects and saves information in the database. If the customer is a regular customer at the coffee shop, then it also check the database for customer past experience and predicts his/her arrival time.

75.3.2 Database Design

This subsection explains the database design for our pre-order system. Figure 75.2 shows the entity-relationship (E-R) diagram. In order to modify and manage the customer and order related information, we made 10 tables. Generally, we can divide the proposed database design into four parts, such as customer related information, coffee shop related information, order related information and information that are related to customer past experience.

First, there are a table called MEMBER_USER that stores the customer related data such as customer's name, gender and phone number, and a table called LOCATION that stores the information related to customer's location such as longitude, latitude and time. In order to store the location information by each

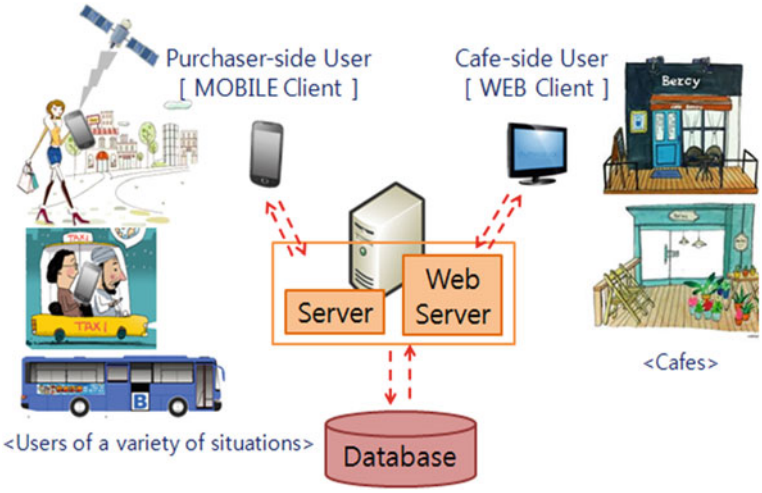


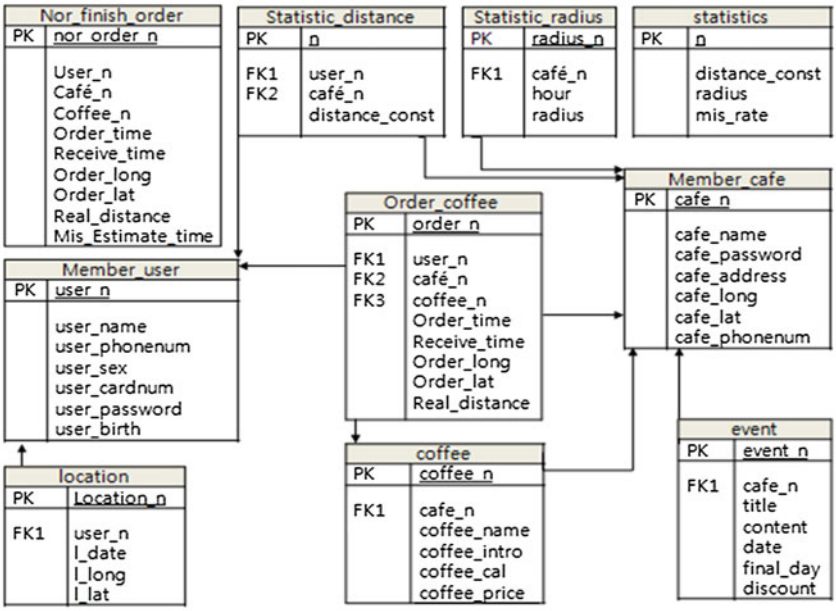
Fig. 75.1 The proposed system

customer, LOCATION table is connected to MEMBER_USER table by customer's number.

Second, there are a MEMBER_CAFE table that stores information related to coffee shop, a table called COFFEE that manages the menus in the coffee shop, and EVENT table that stores various events in the coffee shop, such as discounts on particular item. MEMBER_CAFE table has two main attributes such as *café_long* and *café_lat* that designate the current location of the coffee shop. The coffee shop can have many menus, thus COFFEE table is related to MEMBER_CAFE table.

Third, there are a table called ORDER_COFFEE that has a list of orders in the coffee shop, and table called NOR_FINISH_TABLE that stores a list of orders that coffee shop has already processed. ORDER_COFFEE table uses the customer related information and coffee shop related information, such as menus, thus, it is related MEMBER_USER and MEMBER_CAFE tables. Among main attributes, the ORDER_COFFEE table has *order_time* that stores the time in which the order has been made, *order_long* and *order_lat* that stores the customer's current location, *estimate_time* that stores the customer's arrival time, *distance* that stores the distance between customer's current location and coffee shop, *receive* that confirms if the order is made. NOR_FINISH_ORDER table has following main attributes. It has *real_distance* that stores the real distance of the customer.

Fourth, there are a table called STATISTIC_DISTANCE that stores distance information between coffee shop and customer according to the customer's past experience, a table called STATISTIC_RADIUS that updates the customer's location within 3 min, and a table called STATISTIC that stores the mismatch between the estimated customer's arrival time and real customer's arrival time.



- Tables related purchaser-side user -
MEMBER_USER (user_n, user_name, user_phonenum, user_sex, user_cardnum, user_password, user_birth)
LOCATION (location_n, user_n, l_date, l_long, l_lat)

- Tables related café-side user -
MEMBER_CAFÉ (café_n, café_name, café_password, café_address, café_lon, café_lat, café_phonenum)
COFFEE (coffee_n, café_n, coffee_name, coffee_intro, coffee_cal, coffee_price)
EVENT (event_n, café_n, title, content, date, final_day, discount)

- Tables related order -
ORDER_COFFEE (order_n, user_n, café_n, coffee_n, order_time, receive_time, order_long, order_lat, real_distance)
NOR_FINISH_ORDER (nor_order_n, user_n, café_n, coffee_n, order_time, receive_time, order_long, order_lat, real_distance, mis_estimate_time)

- Tables related past records -
STATISTIC_DISTANCE (n, user_n, café_n, distance_const)
STATISTIC_RADIUS (radius_n, café_n, hour, radius)
STATISTICS (n, distance_const, radius, mis_rate)

Fig. 75.2 E-R diagram and scheme

In order to manage the data according to coffee shop and customer, STATISTIC_DISTANCE table has relationship with MEMBER_USER and MEMBER_CAFE. In order to manage the data according to coffee shops, STATISTIC_RADIUS table has relationship with MEMBER_CAFE.

75.4 Conclusion

As modern people should solve many problems in a limited time, systems that takes the pre-orders at the restaurants and take-out coffee shops are becoming popular. We believe that the need and utilization of such systems will become even more important in the future. Thus, in this paper, we proposed a database design for mobile based pre-order system. The propose database design can serve as a blueprint for other pre-ordering systems in the future.

In the future work, we are planning to present a mobile content that predicts the client's arrival time using the take-out coffee shop example. We are also planning to present an interactive map, which indicates the current location of the customer and how much time it requires for customer to arrive to the coffee shop.

Acknowledgements This work was supported by the IT R&D program of MKE/KEIT (10041854, Development of a smart home service platform with real-time danger prediction and prevention for safety residential environments).

References

1. Menzies T, Hu Y (2007) Data mining for very busy people. Computer 36(11):22–29
2. Chen KWK (2013) Korean restaurant business plan in Taiwan. Thesis in Management Studies, Massachusetts Institute of Technology, Sloan School of Management
3. Castro OS (2013) The impact of technology on customer value in the restaurant industry. Thesis in Business Administration, Saint Peter's University
4. Sheu C, McHaney R, Babbar S (2003) Service process design flexibility and customer waiting time. Int J Oper Prod Manag 23(8):901–917
5. With Coffee application. Available at: http://dawnbeauty.blog.me/90170910450
6. Domino Pizza. Available: www.dominos.co.kr

Chapter 76
Behavior Execution Method for a Robot Using an Online Development Approach

Seungyoub Ssin, Seoungjae Cho, Kyhyun Um and Kyungeun Cho

Abstract This paper introduces a method that applies the behavior execution of a robot to a dynamic environment using an online development approach and demonstrates an approach to improve the convenience of development and overcome the memory limit by moving a key processing unit from the robot to a server. To direct the robot to perform tasks, the server uses image data from 2D cameras. All these tasks are executed in a closed space where the robot can move around. The robot executes tasks through transfer control protocol/internet protocol (TCP/IP) communication with the main server which plays an important role in managing and controlling the motions of the robot and reconstructing the tasks if required. This paper emphasizes on the behavior execution method for a robot by processing images acquired inside a closed space by 2D cameras. In this work the humanoid robot "Nao" was used.

Keywords Online development approach · Task generation · Behavior execution

76.1 Introduction

The robot industry has been focused on the development of robots for replacing humans in manufacturing industries to perform tasks requiring simple repetitive motions. Today, robots are advancing for performing multiple tasks in a variety of fields like health care, defense, and even in routine works.

S. Ssin
Department of Multimedia, Graduate School of Digital Image and Contents, Dongguk University, Dongguk, South Korea

S. Cho · K. Um (✉) · K. Cho
Department of Multimedia Engineering College of Engineering, Dongguk University, Dongguk, South Korea
e-mail: khum@dongguk.edu

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_76, © Springer Science+Business Media Dordrecht 2014

However, robots are not capable of executing tasks under a variety of conditions that may interrupt them from performing predefined motions and movement in a predefined zone. If the development of robots depends only on artificial intelligence (AI) embedded in the robot, the programming may be limited, and a large variety of tasks cannot be constructed during development because of the memory limit of the robot. To solve such issues, an approach must be investigated to include an external system through a server without depending only on the memory and programming in the robot.

The Nao robot from Aldebaran was used in the present study. Nao is a humanoid robot with a total of 25 joints in its body, and it uses an open system with embedded Linux. It has two feet for motion and can be controlled by voice recognition and through transfer control protocol/internet protocol (TCP/IP) and user datagram protocol (UDP) communication using a wireless fidelity (Wi-Fi) network. It was developed using the software Choregraphe and the languages Python and C++. Herein we describe the online development method for generating tasks and exchanging subtask segments using the server and TCP/IP communication. The server system was manufactured using Microsoft Foundation Classes (MFC). The server directs the location movement and executes subtasks by exchanging movement data with Nao in real time.

The server tracks the location of a robot using 2D cameras. The location of Nao is recognized using a color detection symbol (an iconographic symbol) attached to its shoulder. This paper discusses:

1. The color detection of objects by reading the iconographic symbol using the Open Source Computer Vision Library (OpenCV). The detected colors produce two elements: location and direction. After receiving location and direction, Nao moves toward the target designated by a camera using the A* pathfinding algorithm. Then, Nao executes the predefined subtasks in a sequence.
2. The autonomous task execution of a robot through TCP/IP communication with a server in a closed space using 2D cameras to verify the applicability of such implementation in an actual service environment as well as the development method for the server system.

76.2 Related Works

Mixture of Motor Primitives (MoMP) is a method of motion generation that entails an approach to generate tasks for a robot by executing predefined motions and creating new motions by mixing different motions [1]. The concept of mixing motions uses the averages of motion data and changes the motion value depending on the change of speed and angle of an object that enters the detection range of a sensor or a camera. The MoMP algorithm was applied to a table tennis robot. The table tennis robot received several learning sessions and the data was saved in the database of the supervisory system. Next, the robot built up experience by autonomously executing those motions in the database. Motions were determined

by querying the server to identify which learned information was appropriate after measuring the direction and speed of the ball using a camera.

Task reconstructing using a hierarchical task network (HTN) is an approach to reconstructing tasks using the classical planning method by establishing a hierarchical relation with a last node [2, 3]. Tsung-Hsien Yang, basically configured a behavior network using the HTN algorithm and the Web Ontology Language service (OWL-S). There is an initial task at the time of command in the HTN structure, and that task is segmented into subtasks (or sub-services). The subtask list comprises a method, operator, and precondition. The initial task is executed in accordance with the determination of these elements.

The online server system structure for developing and building a tour-guide robot is to develop a service system for the robot guide. A guide robot is assigned to visitors to provide information about a building. The robot recognizes each person in the group and guides them through the building [4]. Anna Hristoskova created an online development approach for a guide robot and proposed a system to manage a robot named Tank for escorting visitors.

Henceforth we propose a structure for a server to identify the status and location of a robot and a procedure for issuing commands in real time using a bidirectional network, not a one-way network such as a Web service, as an online system development approach to controlling the tasks of a robot. The proposed structure enables a robot to execute commands in a dynamic environment. Furthermore, the memory reservoir of the robot is managed in a reasonable way by the proposed approach which the server to carry out unnecessary operations of the robot and obtains results only by assigning functions to divide and reconstruct motions according to feature and not by the combination of simple motions by the server. Moreover, this paper verified the feasibility of controlling multiple service robots in a closed space by assigning an identity to each robot and easily sharing tasks between robots.

76.3 Online Server Development for Robot Control

76.3.1 Generation of Basic Tasks by Demonstration

In the present study, the server approach used to execute the robot's tasks has a structure to control a humanoid robot capable of household chores. The server makes the robot execute tasks by sending indexed motion data. Thus, the robot requires a demonstration [5].

For this demonstration, a teacher holds a robot and moves it in a desired behavioral pattern. Then, the robot imitates the behavior done by the teacher (Fig. 76.1a). The learned motion data (Fig. 76.1b) is saved to a file in a separate motion data server (Fig. 76.1c). The saved files are merged into an indexed file which is sent back to the robot (Fig. 76.1d). This study adopted TCP/IP for sending the indexed files to the robot.

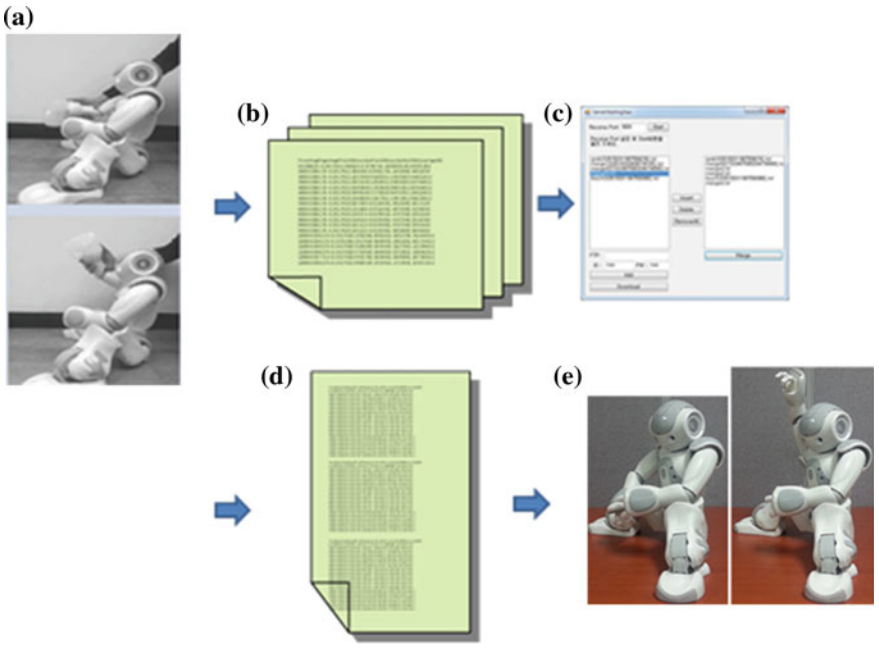


Fig. 76.1 Generation and execution of task after demonstration. **a** robot demonstration by a teacher; **b** separate motion data; **c** separate motion data files in the server; **d** an indexed file of separate motion data; **e** self-execution by the robot

The motion data sent to the robot describes the subtasks to perform the main task. The tasks require the system to generate a table with the indexed motion data as the independent data and save it. For example, a task to pick up an object, move it, and put it in another place can be generated by combining the motions picking up the object, sitting, holding, putting down the object and moving, all of which was acquired through demonstration. The generated task data is indexed as 2, 5, 4, 1, and so on, and saved in the task table. The basic task is executed by playing the recorded data.

76.3.2 Interoperability of Server and 2D Camera

The robot sends task data to the server in the form of network message. The server reads the location and status data of objects in a closed space using a 2D camera. The robot moves toward a target using the object location data from the server and the A* algorithm. Thus, the server and the 2D camera must be connected.

As shown in Fig. 76.2, we identify the objects by attaching iconographic symbols to them and calculate their locations using a 2D camera and OpenCV. The closed space is divided into 9 × 5 cells to assign coordinates to the robot and

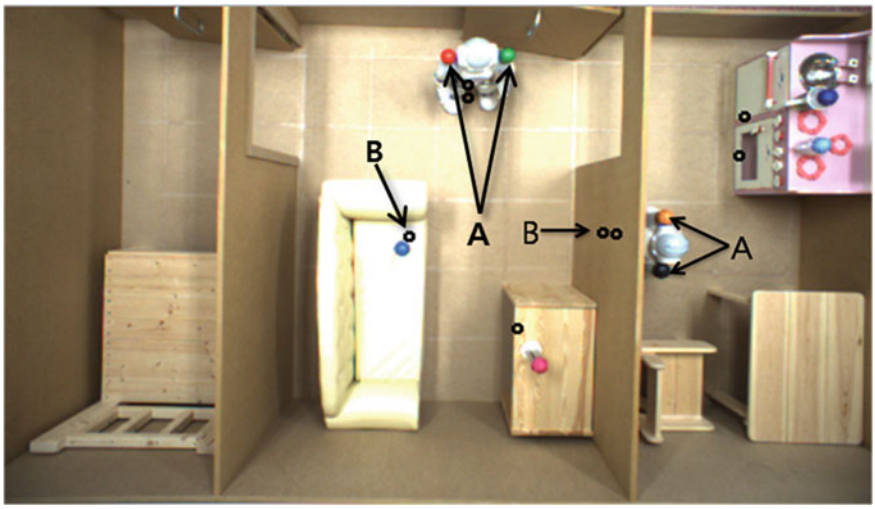


Fig. 76.2 Robot and object image acquired by the 2D camera

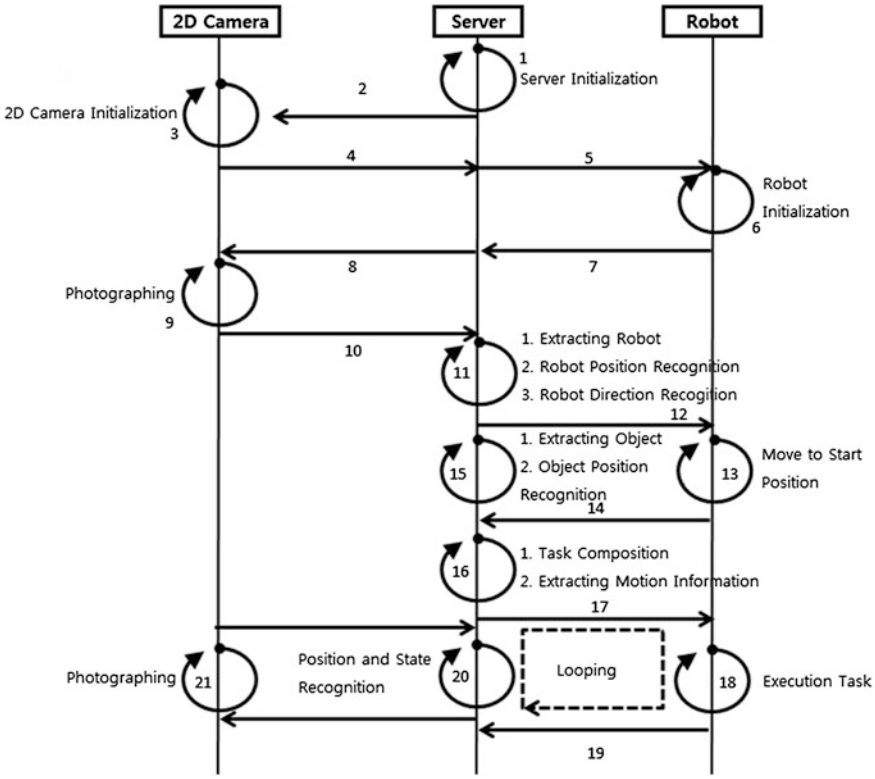


Fig. 76.3 Task execution procedure by the 2D camera, server, and robot

Table 76.1 Processing procedure by camera, server, and robot

Steps	2D Camera	Server	Robot
1	Standby	Initialize	System booting
3	Initialize	Login	Identify joint data
6	Standby	Login	Initialize
9	Taking picture	Login	Standby
11	Standby	1. Identify robot location 2. Estimate robot direction	Standby
13	Standby	Login	Go to the starting point for execution of tasks
15	Standby	1. Identify object 2. Estimate object location	Asynchronous transmission of command completion
16	Standby	1. Configure subtasks to execute using data 2. Generate motion data for each subtask	Standby
18	Standby	Login	Execute tasks using subtask and motion data received

objects. Each object's location is identified in the image using iconographic symbols (A) and an estimate of the object's real location is calculated by projecting the last location vertically onto the floor (B). Then, the location is marked with a circle. Finally, (B) is converted into coordinates suitable for each cell and transmitted to the server program.

With a 2D camera, the server recognizes the robot's location and direction and the object's location. After that, it identifies the target of the robot and the types and locations of the objects. Finally the server executes the predefined tasks corresponding to the current status.

76.3.3 Server Protocol for Task Execution

In the present study, the server processes the data from the image taken by the 2D camera, determines the task data by combining the required motions, sends it to the robot using TCP/IP, and receives a confirmation response from the robot as to whether it executed the commands. An accurately structured protocol system is necessary to execute such complicated procedures in sequence without failures.

Figure 76.3 illustrates the protocol linking the 2D camera, the server, and the robot. The initialization process consists of Steps 1–6. In Step 9, after the initialization process, the camera takes pictures of the closed space (virtual house) shown in Fig. 76.2. The data of the robot location and direction in the space are identified in Step 11. The robot is sent to the starting point, and the server identifies the location of an object using the data acquired by the 2D camera. After receiving

the signal from the robot that it has finished moving to the target, identified in Step 14, the server constructs a task in the Step 16, generates the required motion data, and sends it to the robot. The procedure is repeated in Steps 19–21.

Table 76.1 summarizes the procedures linking the 2D camera, the server, and the robot in sequence.

76.4 Conclusion and Future Research Prospects

This study verified the feasibility of employing a robot to serve humans in a closed space (virtual space) through a demonstration using an online server. Furthermore, herein we proposed an approach to enable the execution of simple tasks by the robot, by linking a 2D camera, the server, and the robot in a dynamic environment that included a variety of conditions. Further studies are suggested to investigate possible approaches for a smooth transition from one camera to another, and the mechanism to enable a server recognize the robot and objects without iconographic symbols.

Acknowledgements This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0011266). And this work was also supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2009148).

References

1. Muelling K, Kober J, Kroemer O, Peters J (2012) Learning to select and generalize striking movements in robot table tennis. Association for the Advancement of Artificial Intelligence
2. Magnenat S, Chappelier J-C, Mondada F (2012) Integration of online learning into HTN planning for robotic tasks. Association for the Advancement of Artificial Intelligence
3. Yang T-H, Lee W-P (2012) A service-oriented framework for the development of home robots. Int J Adv Robot Syst
4. Hristoskova A, Agüero CE, Veloso M, Turck FD (2012) Heterogeneous context-aware robots providing a personalized building tour. Int J Adv Robot Syst
5. Sung Y (2012) Method-action generation method using demonstration-based learning in virtual environments. Dongguk University, Dongguk

Chapter 77
Smart Logistics Service Model Based on Context Information

KyungHun Kwak, NamJin Bae and YongYun Cho

Abstract The Services in the recent logistics environment aims for situation information based logistics service centered around the user by utilizing various IT technologies such as mobile devices, computer systems, and GPS. This thesis proposes a smart logistics service model for providing web-based intelligent logistics service by utilizing smartphones in a smart environment. The proposed service model can use not only the basic logistics delivery information but also the location information and situation information of the delivery vehicle and user as the determining factor information for logistics delivery route and scheduling decisions, and can provide user-based smart logistics delivery service. Through this service model, it is possible to engage in the development applied with various IT logistics convergence services based on situation information between deliverer and user which occurs in real-time.

Keywords Intelligent logistics · Context-aware service · IT logistics convergence

77.1 Introduction

Generally, services in the logistics/delivery environment are mostly labor-intensive and time-consuming due to excessive traffic and frequent changes in user environment. Up to the present, a lot of domestic and foreign researchers have been actively

K. Kwak · N. Bae · Y. Cho (✉)
Department of Information and Communication Engineering, Sunchon National University, Sunchon, South Korea
e-mail: yycho@sunchon.ac.kr

K. Kwak
e-mail: spersdar@hotmail.com

N. Bae
e-mail: bakkepo@nate.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_77, © Springer Science+Business Media Dordrecht 2014

attempting to research and engage in practical development to reduce logistics/delivery time, effort, and provide a faster, safer, and more accurate logistics/delivery service by utilizing various communication devices, satellite technologies, and computing technologies in the logistics service environment [1–3].

Such research efforts are diverse ranging from using position recognition information of logistics vehicles by using wireless telecommunication technology based on RFID and GPS to logistics services through interaction with users. The smart computing environment centered around individual smartphone users, which is recently growing rapidly, is demanding the evolution to human-centered computing services that can reflect the user's opinion and surrounding situation more directly in various service domains. In other words, smart logistics, that can immediately apply this kind of situation information to the logistics/delivery process according to user requirements in the environment surrounding logistics/delivery and frequent changes in environment information, is very important for the enhancement of logistics/delivery productivity and the improvement of service quality.

This thesis proposes a model that can provide smartphone-based smart logistics/delivery services based on various situation information which arises in a ubiquitous computing environment. The proposed service model can use GPS-based position recognition technology as well as various situation information based on sensing information technologies from various sensors in smart space as limiting factors of route and schedule planning of logistics/delivery services. For this, the proposed service model includes GUI-based situation information processing technology using smartphones, and context-aware GUI based logistics service scenario and model technologies which use smartphones.

77.2 Related Work

In recent times, there are active studies underway on logistics automation technologies and IT based logistics smart service technologies in the ubiquitous environment. Especially, practical technology researches are underway on smartphones and RFID/USN technologies for providing context-aware services in various areas from shopping malls to logistics, home delivery, ship and air freight. The studies on logistics environment control technologies using location information and RFID technology has proposed the convergence of IT technologies and logistics regarding various application fields such as control of delivery vehicles using position information based on RFID, method for logistics information and event notification, and management of logistics warehouse [3–5].

Such research has great significance in that it enhances the efficiency of logistics and delivery work by grafting computing models and networking technologies into traditional logistics related work processes. In addition, such studies have developed into studies which attempts to consider the simple data values which is sensed from a sensor for a logistics system and process management and control, along with various situation information such as user and environment information, as the flow

elements of the logistics/delivery process. Through these efforts, various researchers have defined situation information for logistics/delivery environment, and have proposed methods and models for recognizing and processing situation information from logistics/delivery services and systems [6–8].

A study on a system for controlling the logistics transport vehicle using a smartphone proposed a design, which overcomes the limitations of a logistics vehicle control system using only the existing GPS-based location information, and for a smooth logistics vehicle control system in a limited time and location using the 3G network of smartphones [9]. Such research suggests a possibility of networking through the Internet between smartphone and sensors, and the application of various smart services, and system development in a ubiquitous and intelligent space where various RFID/USN-based sensors exist.

In addition, a case study on logistics/delivery services using smartphones presented its usefulness by analyzing the various results which occur when employees in a US logistics/delivery company were actually made to use a logistics/delivery system using smartphones [10]. According to research findings, when Internet based logistics/delivery control system using smartphones was introduced rather than the traditional PDA-based logistics/delivery computer system, it reports that the logistics/delivery productivity was enhanced, and it had a positive effect on the user convenience of people in charge of logistics/delivery using the devices.

Currently, it is predicted that the services using smartphones will have a very wide range and diversity of applicable service domains, target of its execution and application. However, most studies are being concentrated in the tracking of logistics/delivery based on simple position information using GPS and optimization of routes, and there is a lack of studies on the models that can apply not only various types of sensing information communicated from RFID/USN of the logistics/delivery environment but also the delivery requirements communicated from the users into real-time logistics/delivery processes. Therefore, there is a need for the studies on smart logistics/delivery service model that can provide the optimum logistics/delivery service according to user-based situation information arising in real-time in the actual logistics delivery environment.

77.3 Text

77.3.1 Situation Information-Based Intelligent Logistics Service Model

Generally speaking, the actual smart logistics/delivery environment exists in the ubiquitous or intelligent space where various sensors connected to USN carry out mutual computing through networking with wired/wireless communication network. The general structure chart of the service model proposed by this thesis for

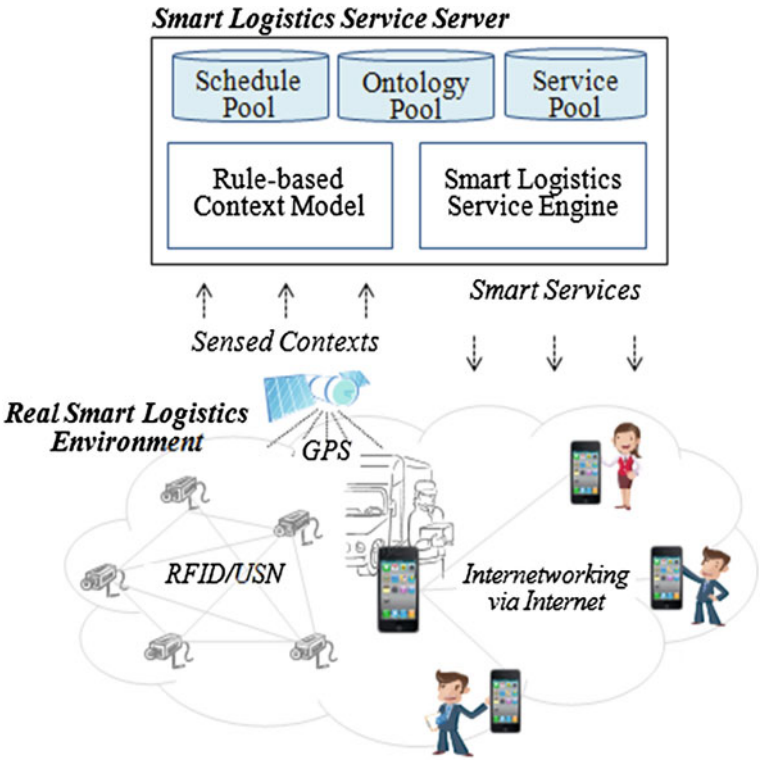


Fig. 77.1 Smart logistics service structure chart

providing a smartphone-based smart logistics/delivery services linked to various sensors in a ubiquitous computing environment is as shown in Fig. 77.1.

The proposed smart logistics/delivery service model consists of the server part for providing the smart logistics/delivery service information based on the actually sensed situation information to the person in charge of logistics/delivery and the user as shown in Fig. 77.1 and the client part for providing the optimum and user location-oriented delivery route information through the smart device of the person in charge of logistics/delivery using the smart logistics/delivery service information provided from the server.

77.3.2 Situation Information Model for Smart Logistics Service

The server receives the input of situation information including the location of the sensed or entered logistics/delivery vehicle, location of the customer, status of the goods, time and method of how the customer received the delivered goods, and

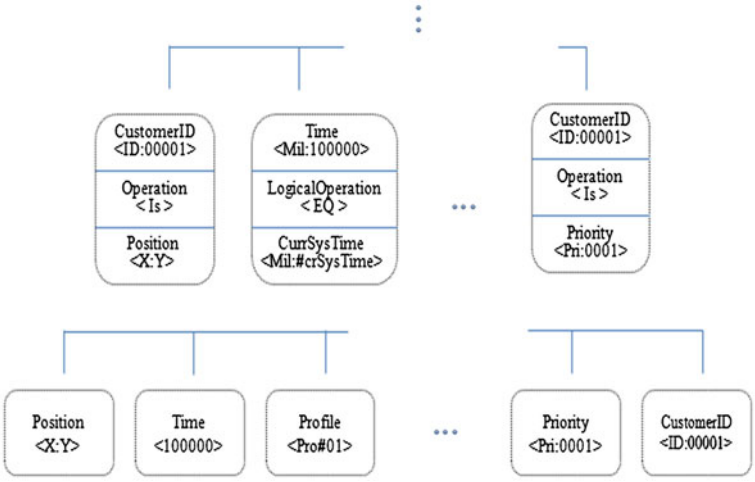


Fig. 77.2 Situation information model for smart logistics service

the real-time route change information of the person in charge of logistics/delivery. In this case, the context model based on the rule of Fig. 77.1 objectifies the low level terminal status information from the client through the RDF-based context model in the form of the status information class form of API level. An objectified low level terminal status information can be processed into a high level status information with a richer meaning through the repetitive rule coupling with another objectified low level terminal status information. Figure 77.2 shows the conceptual diagram of the objectification process by the proposed service model for generating high level status information through coupling that has applied rules through the coupling applied with rules.

The low level data delivered from various sensors in Fig. 77.2 is constituted into types and values according to the ontology knowledge dictionary related to the stored logistics/delivery domain in the ontology storage of Fig. 77.1. A low level data represented with types and values is objectified into a higher level status information through the repetitive coupling process with other low level data. For example, the coordinate information <X:Y> representing the user location transmitted from a position sensor and GPS in Fig. 77.2 is expressed as a pair of position type for showing the actual integer and the semantic information of that value. The generated {Position, <X:Y>} information couples with {CustomerID, <ID:00001>} information for recognizing a specific user, and the user position with ID:00001 can be abstracted into a higher level information called <X:Y>.

The smart logistics/deliver service engine of Fig. 77.1 provides various logistics/delivery related services ranging from the service for providing logistics/delivery route optimization information to the logistics/delivery automatic control service in the form of web service from the service storage using a objectified high level status information. Then, the service engine uses the logistics/delivery plan

information and the high level status information stored in the schedule storage as a limiting factor for executing a specific service. The client sends the sensing information occurring in real-time from various sensors and the RFID and GPS devices installed on the logistics/delivery vehicles to the server. Then, the sensed information can be networked with the user's smart device, and the user and person in charge of logistics/delivery can be communicated through their smart devices. Through this kind of networking, the immediate user requirements can be considered in real time for the new services.

77.4 Service Model Scenario

A simple exercise service scenario for simulating the possibility and usefulness of the proposed smart logistics/delivery service model to be used in actual space is like the following.

(1) Customer A, B, and C is planning to receive their delivered goods at each of their specific location.
(2) The person in charge of the field Y of smart logistics/delivery company X initially receives the delivery route in the order of A, B, and C.
(3) Customer A recognizes that he currently has a personal matter to attend to, and communicates this to the delivery company using his smart device.
(4) Customer B was to personally receive the delivered goods at his home, but B decided to briefly go out nearby.
(5) Customer C was to personally receive the delivered goods at his home, but did not arrive at his house yet.

Now we can obtain the route adjustment and flow of the service that must be newly considered like the following according to the status information by applying the exercise scenario to the proposed smart logistics/delivery service model.

(1) The time information from the private schedule entered by Customer A is objectified into a new status information, and the existing logistics/delivery route is changed using the limitation information for setting a new route. Then, the newly determined route change recommendation information is indicated on the smart device of the person in charge, and the person in charge continues the logistics/delivery service through the newly changed route (B → C → A).
(2) While Customer B is briefly moving near his home where he was initially supposed to receive the goods, he directly sends his own location information to the server using his smart device. Or the location of Customer B is automatically sent to the server through the sensor linked via network to a smart space and the networking with Customer B's smart device.
(3) A new status information is objectified from Customer B's location information, and used as the limiting information for setting a new route. Then, if

Customer B's location should come later than the flow of time context and location context of Customer A or C in the new route which takes into account the time information of Customer A in (1), the new route is decided as C → B → A or C → A → B, and the newly decided route change recommendation information is indicated in real-time on the smart device of the person in charge.

(4) The service engine of the server tracks and monitors in real-time the logistics/delivery service process currently in progress.
(5) If the current location information of Customer B is close to the moving route of the person in charge, the service engine of the server sends a message to Customer B's smart device and provides automatic SMS service so that whether or not the goods were received can be checked midway. At this time, if Customer B reads the SMS text sent to his smart device and changes the reservation to receive the goods midway, the new details are immediately delivered to the person in charge.
(6) Customer C is currently moving to his home, and Customer C's location information is sent to the server in the method explained in (2). At this time, if Customer C's location is before the current location of B or is closer to the route of the person in charge, the new route is decided as C → B → A or C → A → B, and the newly decided route change recommendation information is indicated in real time to the smart device of the person in charge.
(7) The final available route may be B → C → A (If Customer C approaches person in charge Y before he reaches Customer B) or C → B → A (If Customer C can approach person in charge Y after he reaches Customer B).

The location information of each customer can be sent to the server through direct input, and because the proposed smart logistics/delivery service client is already installed and running on Customer A, B, and C's smart devices, the customers' location information can be sent to the server in real-time through various sensors in the smart space and the networking between GPS and customer's smart device. The details examined through the exercise scenario is a consideration of only one instance regarding changes in a very vast and complicated real-time change in status information that may arise in an actual logistics/delivery environment.

77.5 Conclusion

In this thesis, a smart logistics/delivery model using smart devices was proposed. For this, the proposed service model can generate high level status information through a rule-based context model from low level terminal sensing information arising from various sensors existing in smart space and RFID/USN and GPS, and provide status information-based intelligent logistics/delivery service using this. Therefore, the service model proposed in this thesis is anticipated to have high

availability as a diverse mutually interactive logistics/delivery service model based on smart phones, and expected to be of great help in developing related application services and systems in a ubiquitous and intelligent computing environment which is to come in the future.

Acknowledgments This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the CITRC (Convergence Information Technology Research Center) support program (NIPA-2013-H0401-13-2008) supervised by the NIPA (National IT Industry Promotion Agency).

References

1. Lee Y, Jin L, Seo S, Oh S, Lee Y, Tyu K (2004) Design and implementation of the u-logistics alerting system. Korean Inst Inf Sci Eng 31(01):37–39
2. Zhu T, Wang C, Jia G, Huang J (2010) Toward context-aware location based services. In: International conference on electronics and information engineering (ICEIE), pp 409–413
3. Park I, Kim Y, Kim S, Sim C, Shin C (2008) Design and implementation of driver service system for logistics supporting vehicles. Korean Soc Inter Inf 09(02):0197–0200
4. Lee Y, Nam K, Tyu K (2008) Design and implementation of event notification system for location-and RFID-based logistics environment. J Korea Inf Process Soc D 15-D(05): 0599–0608
5. Lee SH, Lee CY, Kim DS, Lee CH, Lee WJ, Min SG, Yoo H (2008) Sensor network deployment for warehouse management system based on RFID. Korea Inf Sci Soc J C 14(01):0022–0030
6. ter Mors AW, Zutt J, Witteveen C (2007) Context-aware logistic routing and scheduling. In: Proceedings of the 17th international conference on automated planning and scheduling, pp 328–335
7. Son VQ, Wenning BL, Timm-Giel A, Görg C (2009) A model of wireless sensor networks using context-awareness in logistic applications. In: Proceedings of the 9th international conference on intelligent transport system telecommunications (ITST 09), pp 2–7
8. Haseloff S (2005) Context awareness in information logistics. Ph.D. Thesis, TU Berlin, Germany
9. Kim M, Oh J, Lee Y-J, Chae J (2010) A design of transportation management system using smartphones. Korea Comput Congr Korean Inst Inf Sci Eng 37(01):212–216
10. Chen JV, Yen DC, Chen K (2009) The acceptance and diffusion of the innovative smart phone use: a case study of a delivery service company in logistics. Inf Manag 46(4):241–248

Chapter 78
Auto-configuration of the Initial Contour Using Level-Set Method for Lung Segmentation on Chest CT Images

Seung-Hoon Chae, Hae-Min Moon, Yongwha Chung and Sung Bum Pan

Abstract Digital medical images assist specialists in improving their diagnostic efficiency and in treating diseases. For example, the chest Computed Tomography (CT) images help in diagnosing the lung disease. The chest CT scan generates multiple images of a patient's lung. A medical image processing technique helps in segmenting these images. It is important to perform this step before processing other medical images. Among the various image segmentation methods available, the method using Level-set is robust to irregular noises. However, the problems faced in using this method include manual input of the initial contour and slow performance speed. Inputting an initial contour to the Level-set that correctly fits the object's form helps in reducing the number of repetitions. This in turn helps in improving the segmentation performance speed. However, it is difficult for a user to input an appropriate initial contour. Therefore, this paper aims at providing a method to auto-configure the initial contour in the Level-set method. Multi-Resolution Analysis (MRA) helps in reducing the pace of the auto-configuration process of the initial contour. In addition, the volume data of a CT image is used to prevent data loss that occurs during the MRA transformation process. Studies have

S.-H. Chae
The Research Institute of IT, Chosun University, Gwangju, South Korea
e-mail: ssuguly@gmail.com

H.-M. Moon
Department of Information and Communication Engineering, Chosun University, Gwangju, South Korea
e-mail: bombilove@gmail.com

Y. Chung
Departmen of Computer and Information Science, Korea University, Seoul, South Korea
e-mail: ychungy@korea.ac.kr

S. B. Pan (✉)
Departmen of Control, Instrumentation, and Robot Engineering, Chosun University, 309, Pilmun-daero, Dong-gu, Gwangju 501−759, South Korea
e-mail: sbpan@chosun.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_78, © Springer Science+Business Media Dordrecht 2014

confirmed that the proposed method facilitates drastic improvements in the performance time and in the segmentation results of chest CT images.

Keywords Medical image · Image processing · Image segmentation · Level-set · Chest CT image · Initial contour

78.1 Introduction

Improvements in the performance of medical imaging equipment and the generation of high-resolution digital images aid computer image analysis in medical diagnosis and treatment of diseases. Several new methods [such as X-Ray, Computerized Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound, and Positron Emission Tomography (PET)] have emerged that aid in extracting or visualizing the information on organ tissues from the diagnostic tomographic images obtained from various medical imaging equipment. This has resulted in active progress of the convergence of medical and IT fields. The accuracy of classifying and discriminating tissues found in the medical images of a complex anatomical structure relies on the analytical skill of the specialists involved in the process. Specialists with identical skill-sets tend to classify or discrimination tissues of a particular image in a different fashion. To avoid data inaccuracy, the need for image analyzes by one specialist has increased significantly. A fast and accurate medical image processing method using a computer helps in solving this issue. Image segmentation is the first and the most important task of this method. Several studies on to this task have been undertaken with utmost interest. Image segmentation is followed by image registration, quantification, visualization and lesion detection. Due to the complexities in the structure of the organs of the human body and similarities in their biological characteristics, it is difficult to conduct the medical image segmentation process [1, 2].

Level-set is preferred over other medical segmentation methods, as it is robust to the impact of irregular characteristics including noise [3]. However, a Level-set requires the user to manually input the initial contour. On the one hand, an inappropriate configuration of the initial contour tends to decline the segmentation performance and the performance speed. On the other hand, an appropriate configuration of the initial contour in the Level-set helps in reducing the number of repetitive computation and in declining the performance time. However, it is impossible for a user to enter the initial contour that would fit the object's form and to configure similar quality of initial contour for all images. Therefore, this paper proposes Multi-Resolution Analysis (MRA) Level-set which auto-configures the initial contour for a Level-set method. This method does not require any additional input by the user and it also helps in improving the performance speed. The MRA Level-set helps to quickly set an initial contour appropriate to the object in the initial contour auto-configuration process. This method helps in reducing the

amount of data using MRA, analyzing medical images, and auto-configuring the initial contour appropriate to the object's form.

This paper also proposes the calibration method of using volume data for the initial contour. This method helps in avoiding data loss during MRA and in improving the quality of initial contour. To evaluate the performance of MRA Level-set method proposed in this paper, lung segmentation of chest CT images were carried out. The experimental results of the proposed method confirmed that this method helped in generating the initial contour appropriate to the form of lung and improving performance time and the segmentation results.

78.2 Proposed Lung Segmentation

78.2.1 Image segmentation Using Level-Set

Level-set is a form of Active Contour Models (ACM). It helps in segmenting images based on the curvature of the initial contour and the images configured over a period of time. It can be formulated by performing mathematical calculations on the contour and the surface. A Level-set does not take the parameters for objects into account. As it is robust to forms and geometric changes, it is suitable to use Level-set in the field of fluid dynamics. Level-set can help in recovering images from irregular characteristics such as noise. In a Level-set, it is also possible to include topological changes independent of locations and forms of the initial contour. Due to these reasons, Level-set is actively applied in the field of image segmentation for extracting, recognizing and identifying a specific object in the medical images [4]. The basic concept of Level-set is that the initial contour keeps moving at a constant rate in a normal direction at a certain point of time. In the Level-set, the speed function "F" of the initial contour moving in a normal direction is calculated as follows:

$$F = F(L, G, I) \tag{78.1}$$

Where L is the curvature speed determined by local geometric information such as the normal direction, and G is a wide range of speed determined by shape and location of the contour. I denotes the speed that can apply force independently regardless of shape of the contour. Level-set considers all three speeds and utilizes them variously in each application fields. The shape progressed at a random time t is expressed as a function $\phi = 0$ and the distance from the random points (x, y) to the curve is denoted by d, and as shown in the following equation, random points can be expressed as Level-set function ϕ.

$$\phi(x, y, t) = \pm d,\ (x, y) \in R^N \tag{78.2}$$

In other words, ϕ value for the points on the curve is 0, and ϕ values for other points in the space are represented as either a negative value or a positive value

depending on whether the point is inside or outside the curve, and the boundary of the curve is represented as 0. In the Level-set, the contour is developed by speed function F. Various speed functions including mean–variance based method and DRSLE-based method are proposed and studied in [5, 6]. Until recently, studies to enhance the performance by improving the speed function of Level-set have been actively underway. The Level-set method is used for lung and brain segmentation as well as for abnormal tissue segmentation in medical images, but the problem faced due to auto-configuration of the initial contour of Level-set that affects its performance has received relatively less attention. Recent studies have shown that segmentation performance can be improved by performing Active Shape Model (ASM) region segmentation or using entropy to set the initial contour of Level-set prior to performing the Level-set method [7].

78.2.2 MRA Level-Set Lung Segmentation Method Using Auto-configuration of the Initial Contour

Level-set is actively applied in the process of image segmentation for extracting, recognizing and expressing specific object from medical images. However, the drawbacks of a Level-set include manual input of initial contour by the user and slow performance speed. This paper proposes the MRA Level-set for an auto-configuration of the initial contour as shown in Fig. 78.1. This MRA Level-set helps in improving user input and in increasing performance speeds. It also complements the existing method that requires a user to input the initial contour.

Auto-configuration of the initial contour consists of the following steps:

- MRA transform
- Initial segmentation
- Reconstruction of the initial segmentation

A MRA transform is a method of separating signals or images into components having high and low frequencies. This is accomplished by changing the resolution of the signal or the image and by analyzing the characteristics of frequencies appearing on these resolutions. MRA can be used for data compression. In this technique, the characteristics of a signal or an image remain constant even when the resolution is altered. Thus, the high frequency and low frequency components of each resolution can be easily separated and viewed. Thus, MRA has been used in various fields for characteristic analysis of signals or images. MRA can used to change the size of an image by reducing its resolution to the desired rate. This implies that multiple blocks of an image can be expressed as one pixel and its resolution can be increased step by step. At the end of this process, one pixel represents one block. In this paper, Wavelet transform is used to apply MRA to images.

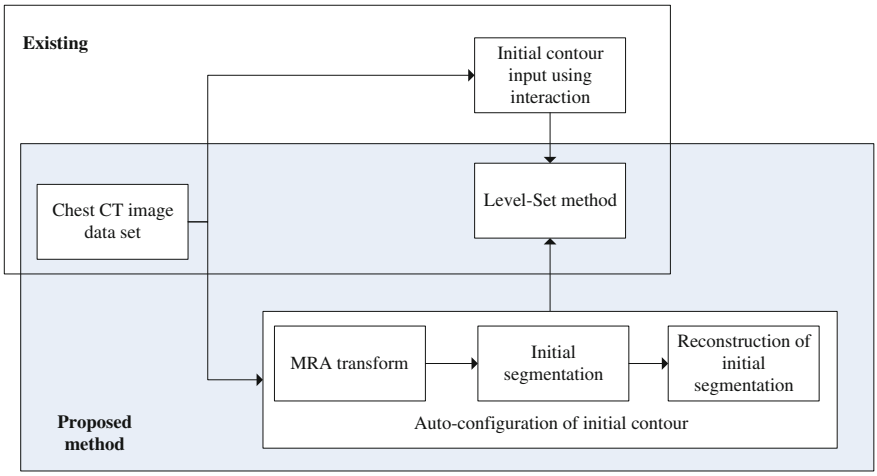


Fig. 78.1 Image segmentation using MRA Level-set

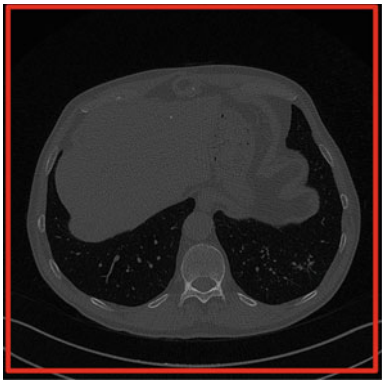


Fig. 78.2 Level-set segmentation for body detection (*red line* initial contour)

The initial segmentation process involves selecting the initial region to enable the auto-configuration of an appropriate initial contour. The initial segmentation process consists of human body detection. This process helps in detecting the body region except the background area and the initial lung area. The body area encompasses the largest area in the chest CT images and is present in the outermost region. After decreasing the amount of data using the level *N*-2 step of the MRA, the body area can be detected by performing Level-set with the initial contour as shown in Fig. 78.2. This Level-set has slow performance speed but it can detect the human body area faster since the amount of image data has been reduced.

After detecting the body area, the next step involves extraction of the candidates for lung region. This process helps in generating the initial contour appropriate to

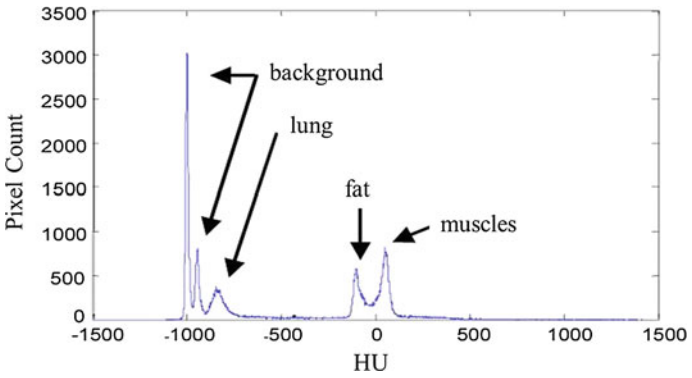


Fig. 78.3 Histogram of the chest CT image

suit the lung form. Due to the existence of other organs in its proximity, the lung's form has complex structure when compared to that of a human body. Therefore, the information on the lung complexity and the organs around it can't be expressed exactly in the *N*-2 level images. Therefore, the candidate of the lung region is extracted at *N*-1 level, which has higher resolution than in the *N*-2 level. The threshold value is used to extract the candidate lung region. The chest CT images have Hounsfield Unit (HU) histogram distribution as shown in Fig. 78.3 [8]. In other words, the density of the lung regions in most images is less than −500 HU. So the lung region can be approximately classified using −500 HU.

The data on body area obtained from the histogram distribution and the body detection process helps in extracting the lung region from the background. However, it is not possible to perform an accurate segmentation process. This is due to the fact that the optimal threshold values in segmentation using histogram vary from one image to the other. In such instances, the segmented region includes the lung region, bronchi and blood vessels. To extract the exact lung region, all regions except the lung tissues should be removed.

The lung region consists of various tissues including lung tissues, air and pulmonary vessels. The bronchi included in the candidate's lung regions are composed only of air as they serve as a passage for air circulation. The interior region of a lung consists of various tissues. And, its exterior region is made of simple tissues or materials. The histogram distribution of the lung's exterior region is thus different from that of its interior region. The graphs shown in Fig. 78.4 represent the distribution of variance and areas of the lung region and other regions respectively. Selection of the final lung region includes:

- Selecting of the lung region using a histogram distribution process.
- Completing the initial segmentation.

The reconstruction process of initial segmentation is the last step in the auto-configuration process. It is performed to remove errors of auto-configured initial contour and to reconstruct the damaged region. As MRA leads to data loss, it is

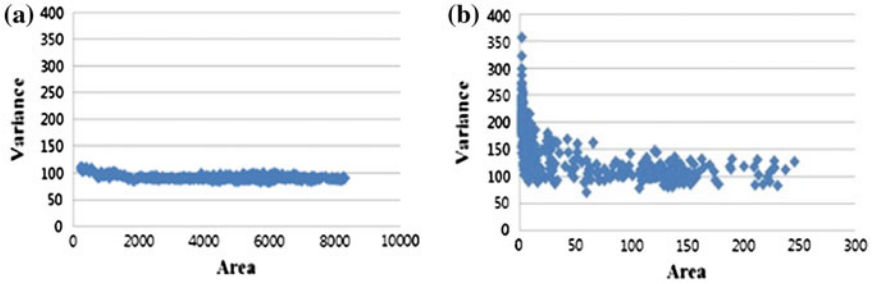


Fig. 78.4 Distribution of variance and area. **a** Lung region, **b** other regions

imperative to undertake the reconstruction process for the region under study. However, there is no loss of data in the lung region present in the beginning and end of the lung as well as in the small region present away from the large region of the specific slice. A small lung region presents an insufficient data making it difficult to segment the lung region using MRA. This initial segmentation is used as an initial contour of Level-set that will be carried out in the original image. Therefore, an improper initial segmentation of the small lung regions leads to problem in the performance of the entire lung segmentation process. The chest image bundles found in the horizontal plane encompass data set of chest CT images. These images help in generating the volume data. Also, forms of the lung between successive slices are similar in nature and are connected naturally. So, the volume data of chest CT images can be used to predict and improve the insufficient data. The connection of coronal plane of the lung has no complex form, so the next slice information can be predicted by applying the linear equation. From the segmentation results, select the n and $n + 1$ as a reference slice. Next, devise the linear equation using coordinates of the boundary points on the two reference segmentation results. With the help of the linear equation using two reference segmentation images, the segmentation boundary coordinates of the $n + 2$ slice can be predicted. By comparing these predicted results with boundary coordinates of the initial segmentation, the segmentation results with error can be reconstructed. In the segmentation reconstruction process, the lung region will be restored using volume data of chest CT images and the linear equation. When the reconstruction process is expressed in formula, it will be as follows:

$$\begin{aligned} P_{n+2} &= L(m_n, m_{n+1}) \\ M_L &= P_{n+2} + m_{n+2} \\ I_{n+2} &= M_L \cap T_{n+2} \end{aligned} \tag{78.3}$$

After defining the linear equation using reference images of m_n and m_{n+1}, the next step is to generate the predicted segmentation results, P_{n+2}. Then, by combining P_{n+2} and m_{n+2}, which are the initial segmentation results, the adjusted result M_L can be calculated. While generating M_L, an error generated due to the absence of the lung region can occur. Generation of I_{n+2} is the next step in this process.

This is done by combining the threshold value segmentation region, T_{n+2}, and by excluding the non-lung region. Finally, the generated I_{n+2} is selected as an initial contour of the Level-set for the lung segmentation.

The initial contour is the segmentation information obtained from initial segmentation of the lung. This information does not provide accurate results for extracting the lung region. A MRA is performed on the initial contour to reduce the performance time of the initial segmentation process. It is thus mandatory to undertake the reconstruction process to restore the size of the initial contour. This implies that after completing the auto-configuration of the initial contour, the initial contour information is reconstructed to its original data, and is then used as the initial contour of Level-set. The initial contour uses the initial segmentation results for auto-configuring purposes. Hence, the lung region is segmented with approximate results. However, owing to its similarity in location and form, the initial contour of the chest CT images can be used on the Level-set to perform the final lung segmentation.

78.3 Experimental Results

To experiment the method proposed in this paper, DB of VESSEL Segmentation in the Lung 2012 (VESSEL12) was used. VESSEL12 DB includes the chest CT image dataset and the segmentation mask dataset, which is the lung region segmentation information. The Level-set used in the experiment uses the DRLES speed function proposed by *Li* and using Dice's Overlap to measure the lung region segmentation performance, accuracy (*S*) and standard deviation (*Std*) were measured [9–11].

Figure 78.5 shows the segmented lung region using the proposed method. And Table 78.1 shows the experiment results of the proposed method. To compare the performance, the Level-set using the initial contour generated through the user input was performed additionally. The initial contour input time by the user was 10–20 s. And, the input time depended on the forms and sizes of the lung.

Both user input Level-set and the proposed method have accuracy rate of more than 0.95. And, the proposed method had higher accuracy than the user input method. The standard deviation of accuracy help to constantly maintain the performance of the user input Level-set in each slice. And, when compared to the proposed method, the image segmentation method of the user input Level-set had constant performance for each slice. Unlike the proposed g auto-configuration method, other segmentation methods intervene in user's input for image segmentation process. This leads to higher performance in the standard deviation. The initial contour set by the user's naked eye was less optimized and slow when compared with the MRA Level-set. This resulted in lower performance of the user input Level-set. The region outside the lung exists in a non-optimized region. It thus showed lower performance than that of a MRA Level-set. In addition, the necessary number of cycles for the user input and proposed method was measured

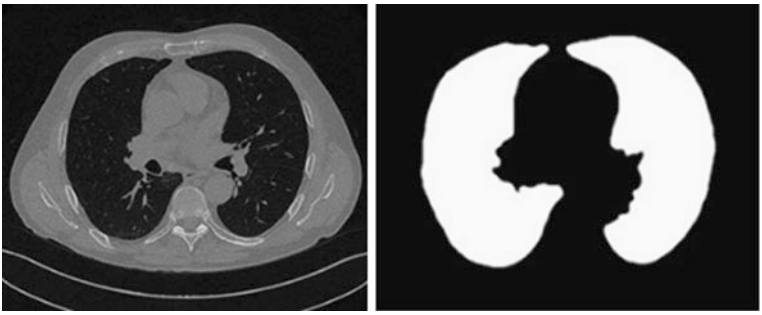


Fig. 78.5 MRA level-set lung segmentation results using auto-configuration of the initial contour

Table 78.1 Lung segmentation experiment results

	User input initial contour	Proposed MRA level-set
S	0.957	0.984
Std	0.128	0.170

as 5.373×10^{11} and 8.569×10^{9}, respectively. This shows that the performance of Level-set was reduced in the proposed method because a more suitable initial contour was entered in this method.

78.4 Conclusion

Owing to the improvements in the performance of medical imaging equipment, medical diagnosis using image analysis with computer has become significant. The medical image segmentation should be performed first before undertaking other steps in of medical image processing. However, the similarities in the biological characteristics of organs cause difficulties in performing region segmentation of medical images. Of all the medical image segmentation methods, Level-set has a slow performance speed as it requires the user to configure an initial contour. However, the performance speed can be improved by entering appropriate initial contour to the form of the object. This paper proposed a MRA Level-set that suggested auto-configuration of the initial contour to as a solution to this problem. In this MRA Level-set, MRA was applied to reduce the computation of auto-configuration process of the initial contour. And, to rectify errors present in the initial contour caused by data loss. To rectify these errors, the volume data characteristics of chest CT images and segmentation reconstruction method using linear equations are proposed in this paper. According to the experimental results, it was confirmed that the proposed method had lower performance and higher accuracy when compared to the Level-set using the initial contour through the user

input, but overall, it was established that the proposed method had higher segmentation accuracy.

In the future, studies on the detailed segmentation for the bronchi, pulmonary vessels, and lobe, which exist within the segmented lung region and on automatic detection of a suspicious region will be carried out.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0023147).

References

1. Costaridou L (2005) Medical image analysis methods. Taylor & Francis, London
2. Rettmann ME, Holmes III DR, Camp JJ, Packer DL, Robb RA (2008) Validation of semi-automatic segmentation of the left atrium. In: Proceedings of SPIE, p 6916
3. Lee M, Cho W, Kim S, Park S, Kim JH (2012) Segmentation of interest region in medical volume images using geometric deformable model. Comput Biol Med 42(5):523–537
4. Malladi R, Sethian JA, Vemuri BC (1995) Shape modeling with front propagation: a level set approach. IEEE Trans Pattern Anal Mach Intell 17(2):158–175
5. Li C, Xu C, Gui C, Fox MD (2005) Level set evolution without re-initialization: a new variational formulation. Comput Vision Pattern Recogn 1:430–436
6. Chen S, Kohlberger T, Kirchberg KJ (2011) Advanced level set segmentation of the right atrium in MR. In: Proceedings of SPIE 2011, vol 7964
7. Lee M, Cho W, Kim S, Park S, Kim JH (2012) Segmentation of interest region in medical volume images using geometric deformable model. Comput Biol Med 42(5):523–537
8. Wei Q, Hu Y, Gelfand G, MacGregor JH (2009) Segmentation of lung lobes in high-resolution isotropic CT images. IEEE Trans Biomed Eng 56(5):1383–1393
9. Li C, Xu C, Gui C, Fox MD (2010) Distance regularized level set evolution and its application to image segmentation. IEEE Trans. Image Process 19(12):3224–3254
10. Dice L (1945) Measures of the amount of ecologic association between species. Ecology 26(3):297–302
11. VESSEL12, http://vessel12.grand-challenge.org/

Chapter 79
Suction Based Wall Climbing Robot for Edge Movement

Jeesoo Lee and Sanghoon Kim

Abstract The intelligent mobile robot with sensors and image processing embedded system combines the suction and aerodynamic attraction to achieve good balance between strong adhesion force and high mobility for the purpose of the warehouse inspection. To be able to move across the edge region, the new robot is designed to connect two robot modules with the waist joint made of strong motor. The height and weight of the robot main body was reduced to improve the suction forces by the physical analysis. It also implements object detection ability using effective color transform and segmentation technique for the exact target detection on the wall using a embedded camera system, communication module and several active sensors. Experimental results showed that the robot can move upward on the wall at the speed of 3.4 m/min and carry 10 kg payload in addition to 2.5 kg self-weight, which record the highest payload capacity among climbing robots of similar size.

Keywords Wall climbing · Edge movement · Embedded image processing · Adhesion force · Object detection

79.1 Introduction

Recently intelligent robot can be used to detect and inspect target objects instead of man in many fields such as construction site and merchandise warehouse. Climbing robots are important to increase the operation efficiency in warehouse environments or difficult-to-access places, and to protect human health and safety in hazardous tasks. Climbing robots with the ability to move on vertical surfaces

J. Lee · S. Kim (✉)
Department of Electrical, Electronic and Control, Hankyong National University,
67 Seokjeong-dong, Anseong, Kyonggi-do, South Korea
e-mail: kimsh@hknu.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_79, © Springer Science+Business Media Dordrecht 2014

are currently being strongly requested by various industries and military authorities in order to perform dangerous operations such as inspection of high-rise buildings, aircraft inspection, surveillance and rescue operations [1, 2], etc. One of the most challenging tasks in climbing robot design is to develop a proper adhesion mechanism to ensure that the robot sticks to wall surfaces reliably without sacrificing mobility [2–4]. So far, four types of adhesion techniques have been investigated: (1) magnetic devices for climbing ferrous surfaces [5]; (2) vacuum suction techniques for smooth and nonporous surfaces [6–8]; (3) attraction force generators based on aerodynamic principles; (4) biomimetic approaches inspired by climbing animals [1]. Magnetic adhesion devices are most promising, but applicable only around on steel structures. Biomimetic adhesion devices are most natural method, but not strong enough for practical use, especially when large payload is desired. In applications for non-ferromagnetic wall surfaces, climbing robots most generally use vacuum suctions to produce the adhesion force. The common defects of the suction-based climbing robots lie in the facts that the suction cup requires perfect sealing and it takes time to generate vacuum and to release the suction for locomotion [7]. Thus they can only operate on smooth and non-porous surfaces with low speed. These constraints greatly limit the application of the robots.

So the robot in this paper combines the suction and aerodynamic attraction to achieve good balance between strong adhesion force and high mobility. Since the robot do not require perfect surface, it can move on virtually various kinds of smooth or rough surfaces. To be able to move across the edge region, the new robot is designed to connect two robot modules with the waist joint made of strong motor. The height and weight of the robot main body was reduced to improve the suction forces by the physical analysis. It also implements object detection ability using effective color transform and segmentation technique for the exact target detection on the wall using a embedded camera system, communication module and several active sensors. Experimental results showed that the robot can move upward on the wall at the speed of 3.4 m/min and carry 10 kg payload in addition to 2.5 kg self-weight, which record the highest payload capacity among climbing robots of similar size. With two 11.1 V lithium-polymer battery, the robot can operate continuously for 40 min.

79.2 The Robot Design and Structure

The robot system consists of body, host PC, Zigbee protocol module, embedded board with S5PV210 microprocessor and 5 M pixel CMOS Camera, remote controller and it is shown in Fig. 79.1 [9]. The robot body includes impeller, chamber, BLDC motors, geared DC motors for driving wheels and rotors, and includes several sensors for detecting range, gas, flame, pressure difference. Especially a embedded board with S5PV210 microprocessor and 5 M pixel CMOS Cameras are adopted for detecting dangerous situation on the wall using color

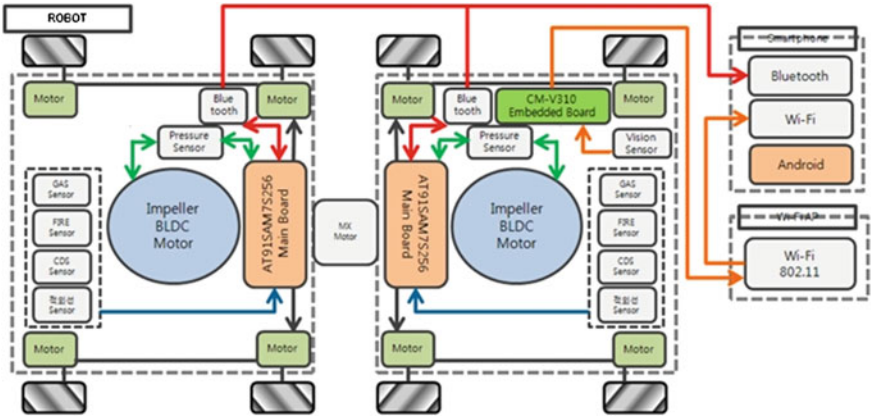
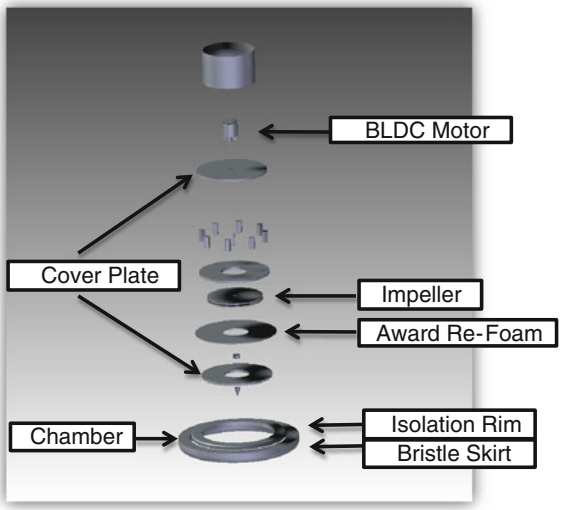


Fig. 79.1 The overall structure of the wall climbing intelligent mobile

Fig. 79.2 Adhesion system and impeller



segmentation technique and sending high quality image processed results to the smart phone or to the neighboring control nodes. AT91SAM7S256 is used to control the motors and sensors.

Figure 79.2 is suction-based adhesion system and it consists of motor cover, impeller, award re-foam, bottom re-foam, chamber. The adhesion device for Arachne is based on the aerodynamic attraction produced by a vacuum rotor package which generates a low pressure zone enclosed by a chamber [2, 4, 7]. The vacuum rotor package consists of a vacuum motor with impeller and exhaust cowling to direct air flow as shown in Fig. 79.2.

Fig. 79.3 Overall exterior view of the robot

In order to generate and maintain attraction force due to the pressure difference, a vacuum chamber is needed to enclose the low pressure zone. Figure 79.2 shows a vacuum rotor package installed on a plate, and a vacuum chamber with flexible bristle skirt seal. When the air is evacuated through the hole on the plate by the vacuum rotor, the larger volume of the chamber, and the smaller gaps between the seal and contact surface, the lower steady state pressure we can obtain, thus increase the attraction force and load capacity. The wheels are outside of the robot frame, making it possible for each module to move from ground to wall easily. Due to efficient placement of the driving system the robot is capable of $\pm 90°$ transitions. We used one motor only for suction ability and 4 more motors for robot driving. The driving motors drive each wheel independently, and via the controllers they drive the front, rear, left and right wheels. To be able to move across the edge region, two robot modules are connected through the waist joint made of strong motor placed in the middle. The two modules are also closely coupled to reduce the torque required to lift up other module. Figure 79.3 shows the overall exterior view of the robot.

Figure 79.4 shows related suction and aerodynamic attraction to achieve good balance between strong adhesion force and high mobility. To be able to move on the wall vertically, the vertical normal force N on the robot should be equal to the friction of the suction (Fvacuum) while the Fvacuum has to be equal to the combined normal forces to each 4 wheel. The Friction of each wheel (FFriction) is the vertical normal force (N) multiplied by the coefficient of friction. Finally, robot can move on the wall safely only when FTraction is greater than the sum of FFriction and FGravity.

And by analyzing the equation regarding the sum of momentums in Fig. 79.4, we can know that reducing the height of the robot chamber and body can improve

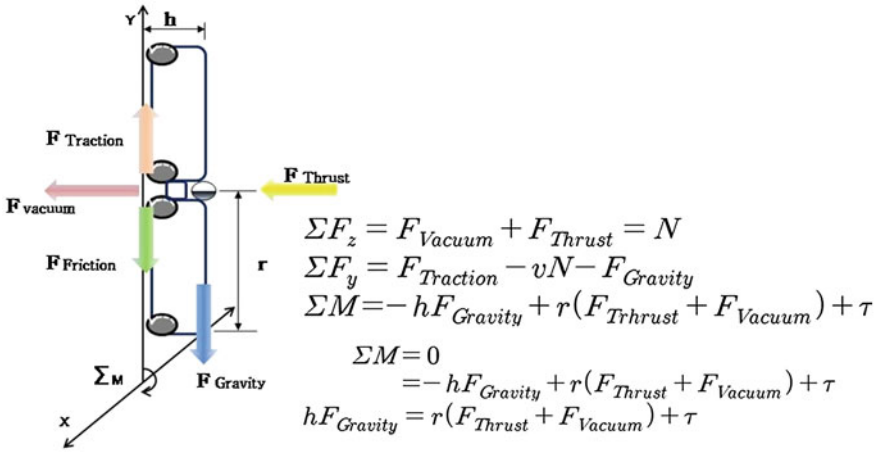


Fig. 79.4 Analysis of existing forces

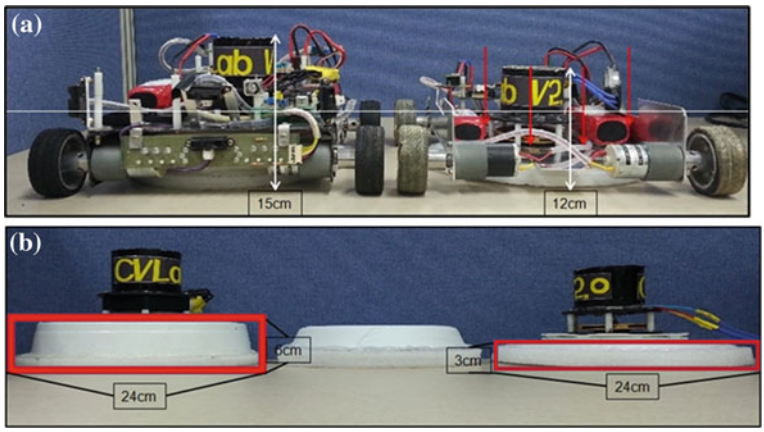


Fig. 79.5 Improvements in the height and weight of robot compared to the version 1 (**a**) reduction of the height of robot (15 > 12 cm) (**b**) reduction of the height of chamber (6 > 3 cm)

the mobility only with the lower suction forces of the robot. The height and weight of the robot main body was redesigned to improve the mobility and keep low suction forces. The change of the size of the robot structure is shown in Fig. 79.5.

79.3 The Embedded Image Processing System

The robot adopted very high performance embedded board with high speed CPU (1 Ghz Samsung ARM Cortex), highly integrated memory and controller, high

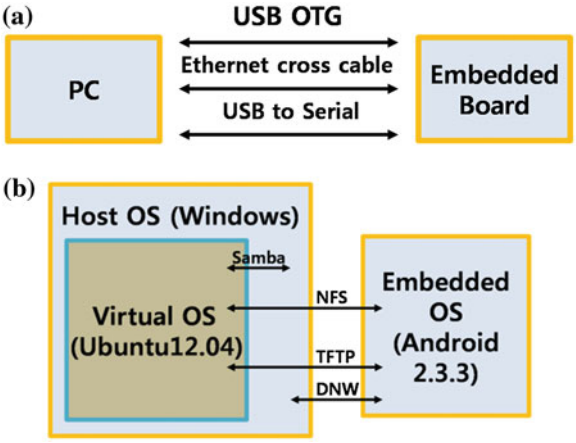


Fig. 79.6 Cross development model for H/W and S/W of robot embedded system. **a** Cross development model for hardware, **b** cross development model for software

speed CMOS Camera and interfaces and LCD interface for the evaluation of the intelligent image processing performance in mobile robot. Figure 79.6 shows Cross Development Model for H/W and S/W of robot embedded system, representing interfaces and relations between PC and embedded board. Because the embedded board do not have compiler internally for the maintaining of simplicity of the embedded environments, the Cross Development Environment (CDE) should be done. The hardware model for cross development between PC and embedded board has interface method of USB, Ethernet and USB to serial connection, while the software model for cross development has virtual OS installation space (Ubuntu12.04) in the host OS (Windows) of PC to communicate with embedded OS of the embedded board safely [9].

And the overall image processing algorithm on the embedded board includes RGB normalization, color transform, segmentation and final object detection [9–11] as shown in Fig. 79.7.

79.4 Experimental Results

The Geared DC motor used for driving robot has the gear ratio of 1:210 and 30.9 RPM. The perimeter of the wheel is 6 cm and the maximum speed on the surface is about 4.8 m/min.

The average speed of the robot in various direction (upward, downward and horizontal direction) are shown in Table 79.1 and the upward move on the wall at the speed of 3.4 m/min is the highest speed among climbing robots of similar size. The robot is able to move on various wall surfaces with the module weight of 2.5 kg and it can generate enough adhesion force to carry additional 6 kg payload (at least). Table 79.2 shows that the payload capacity varies according to the position of the robot.

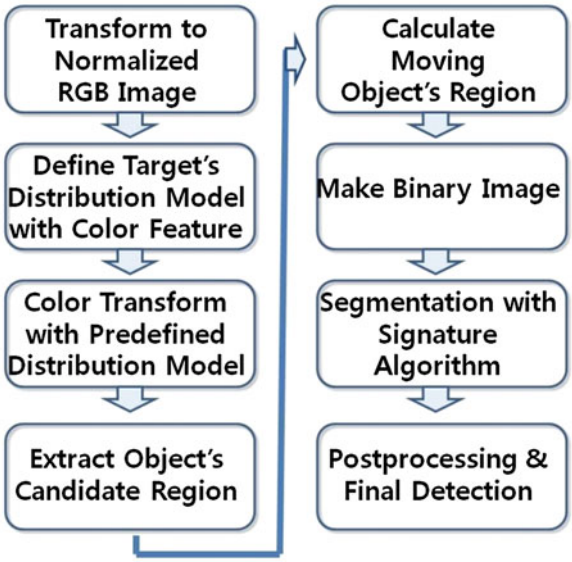


Fig. 79.7 Image processing procedure for object detection

Table 79.1 The average speed of the robot in various direction

Types	Speed (m/min) (ver.1)	Speed (m/min) (ver.2)
Ground move	4.8	5.7
Upward move	2.9	3.4
Downward move	3.6	4.3
Horizontal move on the wall	3	3.8

Table 79.2 Payload Capacity according to the robot position

Robot position	Vertical load (V.1) (kg)	Vertical load (V.2) (kg)	Horizontal load (v.1) (kg)	Horizontal load (v.2) (kg)
Ground	8	12	6	9
Wall	7	10	7	7
Ceiling	5	7	6	6

Table 79.3 Effective battery duration

Endurance for MCU/sensor	Endurance for impeller motor
2 h 30 min	40 min

Table 79.3 shows Effective battery duration of the robot and it can operate continuously for 40 min with two 11.1 V lithium-polymer battery.

The robot's locomotion and attaching ability on various wall environments with payloads is shown in Fig. 79.8. It shows the robot can move well even when on the

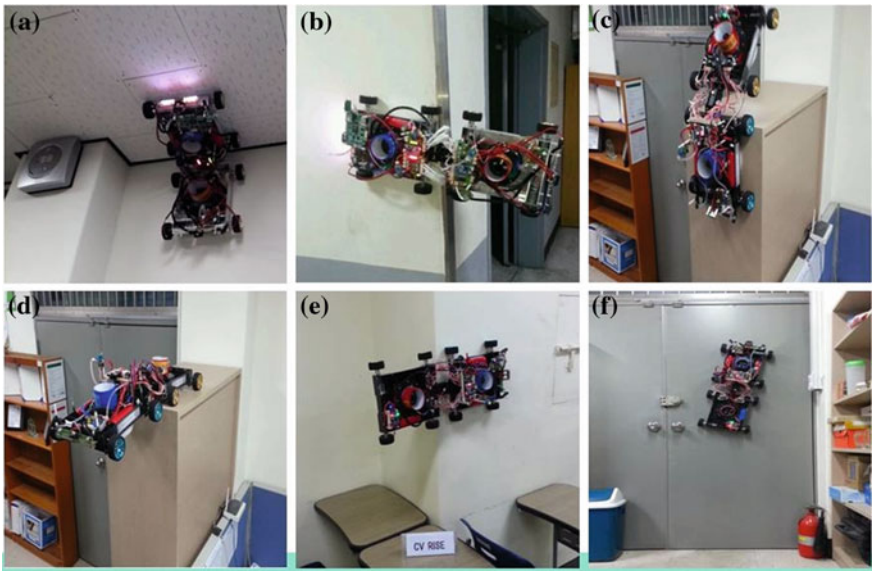


Fig. 79.8 Various robot position with payload (**a**) wall to ceiling (**b**) move the edge region of the wall (**c**) , (**d**) move the edge of the bookshelf (**e**) move the edge region of the wall (**f**) move on the ferrous wall

Fig. 79.9 Image processing and communication between embedded board and the smart phone and test results for object detection

ferrous wall, on the boundary and edge area between ground and wall, on the ceiling. Figure 79.9 shows the embedded board and Android smart phone shares the object detection results and the smart phone controls the robot remotely with user interface program and the robot finally detects and tracts the target object defined by the manager automatically as shown in upper picture of the Fig. 79.9.

79.5 Conclusions

The intelligent mobile robot with sensors and image processing embedded system combines the suction and aerodynamic attraction to achieve good balance between strong adhesion force and high mobility for the purpose of the warehouse inspection. To be able to move across the edge region, the new robot is designed to connect two robot modules with the waist joint made of strong motor. The height and weight of the robot main body was reduced to improve the suction forces by the physical analysis. It also implements object detection ability using effective color transform and segmentation technique for the exact target detection on the wall using a embedded camera system, communication module and several active sensors. Experimental results showed that the robot can move upward on the wall at the speed of 3.4 m/min and carry 10 kg payload in addition to 2.5 kg self-weight, which record the highest payload capacity among climbing robots of similar size.

Acknowledgments This work was supported by a research grant from Gyunggi-do (GRRC) in 2013–2014 (GRRC Hankyong 2012-B02), Development of Automatic Picking System with Traveling Dispensers and Automatic Classification and Defect Detection System with Machine Vision Technology.

References

1. Clark J, Goldman D, Lin P, Lynch G, Chen T, Komsuoglu H, Full R, Koditschek D (2007) Design of a bio-inspired dynamical vertical climbing robot. In: Proceedings of robotics: science and systems, Atlanta, Georgia, USA, June 2007
2. Kang MJ, Moon HP, Choi HR (2010) A study on the design and control of wall climbing robot using impeller. Master's thesis, SeongkyoonKwan University
3. Yamaguchi T, Sorioka Y, Park S, Hashimoto S (2009) SIEN: telescopic-arm climbing-support robot. Department of Applied Physics, Waseda University
4. Silva MF, Machado JT (2006)A survey of technologies and applications for climbing robots locomotion and adhesion. Instituto Superior de Engenharia do Porto Portugal
5. Hirose S, Tsutsumitake H (1992) Disk rover: a wall-climbing robot using permanent magnet disks. In: Proceedings of the 1992 IEEE/RSJ international conference on intelligent robots and systems, Raleigh, NC, pp 2074–2079
6. Li J, Gao X, Fan N, Li K, Jiang Z, Jiang Z (2010) Adsorption performance of sliding wall climbing robot. Chin J Mech Eng
7. Xiao J, Sadegh A (2008) City-climber: a new generation wall-climbing robots. The City College, City University of New York, USA

8. Kim S-H, Jeong Y-S (2013) Mobile image sensors for object detection using color segmentation, Cluster Comput. doi 10.1007/s10586-013-0287-2), Springer, 2013.06,14 (online publish)
9. Kim S, Kwon H (2013) Object detection image processing based on the mobile embedded environment. Korean Soc Comput Game 26(1)
10. Kim S, Lee S, Kim S, Lee J (2009) Object Tracking of Mobile Robot using Moving Color and Shape Information for the aged walking. IJAST 3:59–68
11. Kim S (2006) Object tracking using moving color and active contour information. In: ITC-CSSS2006, Thailand

Chapter 80
Improving Data Hiding Capacity Based on Hamming Code

Cheonshik Kim and Ching-Nung Yang

Abstract Matrix encoding proposed by Crandall can be used in steganography data hiding methods. Hamming codes are kinds of cover codes. "Hamming + 1" proposed by Zhang et al. is an improved version of Matrix encoding steganography. The embedding efficiency of "Hamming + 1" is equal to $(k + 1)2^{k+1} - 1/(2^{k+1} - 1)$, and embedding rate is $(k + 1)/2^k$. Our proposed "Hamming + 3" scheme has a slightly reduced embedding efficiency, but improve highly embedding rate. We therefore propose verifying the embedding rate during the embedding and extracting phase. Experimental results show that the reconstructed secret messages are the same as the original secret messages, and that the proposed scheme exhibits a good embedding rate compared to that of previous schemes.

Keywords Watermark · Steganography · Matrix encoding · Hamming codes

80.1 Introduction

The purpose of data hiding [1, 2] is to facilitate covert communication in the form of concealed messages in a cover media to modify the media. In the case of a single carrier for an application, all secret information such as images, videos, and MP3 files is stored in the carrier. The goal of data hiding is to ensure that embedded data remain inviolate and recoverable. There are two issues with data hiding. One is to provide proof of the copyright, and the other is to provide

C. Kim (✉)
Department of Digital Media Engineering, Anyang University, Anyang, Republic of Korea
e-mail: mipsan@paran.com

C.-N. Yang
Department of Computer Science and Information Engineering, National Dong Hwa University, Hualien, Taiwan
e-mail: cnyang@mail.ndhu.edu.tw

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_80, © Springer Science+Business Media Dordrecht 2014

assurance of content integrity. Therefore, the data should stay hidden in a host signal, even if that signal is subjected to manipulation or degrading such as filtering, re-sampling, cropping, or lossy data compression. However, data hiding generally shows weakness to such manipulation. There are trade-offs between the quantity of embedded data and the degree of immunity to host signal modification. As one increases, the other must decrease. Although this can be shown mathematically for some data-hiding systems such as a spread spectrum, it seems to hold true for all data-hiding systems. The goal of steganalysis is to detect (and possibly prevent) such communication. Generally, steganalysis tools can easily detect a stego image when the error rates are over about 10 % to conceal a message.

Crandall [3] proposed a new data hiding scheme called matrix encoding. The F5 algorithm [4] is based on matrix encoding and implemented by the Westfeld. We can the definition of the cover coding [5–7] in [4]. Matrix encoding was also used in large payload applications [8]. BCH codes were applied to achieve a tradeoff between embedding complexity and efficiency [9]. Westfeld showed matrix encoding using Hamming codes. The CPT method [10] shows the embedding efficiency by hiding messages based on the weighted value of a block. Matrix encoding and CPT can be applicable to LSB steganography. Zhang and Wang [11] showed the ternary Hamming codes using the concept of efficiency by exploiting the modification direction (EMD). The performance of "±steganography" was introduced by the [12]. Mielikainen [13] presented a method based on a pair of two consecutive secret bits. Chang et al. [14] proposed (7, 4) Hamming code for data hiding, which improves on the "Hamming + 1" scheme.

In this paper, we propose novel improving data hiding methods by extending the Hamming codes. Our proposed method can significantly improve the embedding rate of "Hamming + 1" scheme, and perform equally well, or even outperform.

The rest of this paper is organized as follows. In Sect. 80.2, we review current and related work. In Sect. 80.3, we introduce our proposed "Hamming + 3" for grayscale images. In Sect. 80.4, we explain the experimental results. Section 80.5 presents our conclusions.

80.2 Related Works

In Sect. 80.2.1, we will describe the concept of Hamming code and show how to apply Hamming codes to data hiding. In Sect. 80.2.2, the basic theory and efficiency of "Hamming + 1" is presented.

80.2.1 Hamming Codes

Linear codes with length n and dimension k will be described as $[n, k]$ codes. Hamming codes are linear codes and will be described as a $[n, k]$ q-ary Hamming

code, where q is the size of the base field, F_q. A generator matrix G for an $[n, k]$ linear code c (over any field F_q) is a k-by-n matrix for which the row space is the given code. In other words $c = \{xG | x \in F_q^k\}$. Matrix encoding conceals messages with the parity check matrix of linear codes. If $\boldsymbol{c}$ is an $[n, k]$ linear code, the dual to it is an $[n, n - k]$ linear code. If $\mathbf{H}$ is the checker matrix for $\boldsymbol{c}$, H is an $(n - k) \times k$ matrix the rows of which are orthogonal to $\boldsymbol{c}$ and $\{x \mid Hx^{\mathrm{T}} = 0\} = \boldsymbol{c}$.

$$(m_1, \ldots, m_k)^T = H \cdot (LSB(x_1), \ldots, LSB(x_n))^T \tag{80.1}$$

The Hamming codes function is to embed k bits $(m_1, \ldots, m_k) \in F_k^2$ in the LSBs of n pixel gray values $(x_1, \ldots, x_n)$ by at most R changes in the following manner. Note that the covering radius R is the largest number of possible changes and the purpose of Hamming codes is to minimize the average number of embedding changes R_a. In other words, the goal is to maximize the embedding efficiency k/R_a depending on the embedding rate k/n. We note that to correct one error, the position of the erroneous bit must be determined. For an n-bit code, $\log_2 n$ bits are therefore required. Equation (80.2) shows the parity check matrix for a (7, 4) Hamming code:

$$H = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{80.2}$$

For c to be a codeword, it must be in the null space of this matrix, i.e., $Hc = 0$. Let us assume there is a sequence of bits that have an error in the first bit position, e.g., 1101010_b. We calculate the syndrome S with Eq. (80.1). c is a 7-bit binary number and T denote the transpose of a codeword c, that is, the syndrome is $([001])^{\mathrm{T}}$. A syndrome value that is not zero denotes the position of the erroneous bit. If one flips the bit at this position in the codeword, every bit of the codeword will be correct. Binary Hamming codes are $[2^r - 1, 2^r - 1 - \mathrm{r}]$ linear codes with a parity check matrix H of dimensions $r \times (2^r - 1)$ and whose columns are binary expansions of the numbers 1,…, $2^{\mathrm{r}} - 1$. For example, Eq. (80.3) shows the parity check matrix H for $r = 4$. Let us assume that the cover object is an image consisting of $P \times Q$ pixels.

$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{80.3}$$

Example 1 We assume that the codeword $\boldsymbol{c}$ is [1101001]. It is easy to calculate the syndrome using Eq. (80.1) with the parity check matrix H and the codeword: $S = H \times (c)^T = ([000])^T$. If the computed syndrome vector S is 0, as in this case, there is no error in the codeword. Otherwise, there is an error in the bit at position S in c.

80.2.2 "Hamming + 1" Scheme

The parity check matrix of a Hamming code yields a covering function $COV(1, 2^k - 1, k)$, $k \geq 1$, i.e., embed k bits $(m_1, \ldots, m_k)$ into the LSBs of $2^k - 1$ pixel gray values $(x_1, \ldots, x_{2^k-1})$ using at most one change. This covering function is defined by $(m_1, \ldots, m_k)^T = H \cdot (LSB(x_1), \ldots, LSB(x_{2^k-1}))^T$, where H is the parity check matrix of $[2^k - 1, 2^k - 1 - k]$ Hamming code. Hamming covering function combines with one pixel to form "Hamming + 1" scheme, which embeds $k + 1$ bits into 2^k pixels gray values using at most one change:

$$(m_1, \ldots, m_k)^T = H \cdot (LSB(x_1), \ldots, LSB(x_{2^k-1}))^T \tag{80.4}$$

$$m_{k+1} = (\lfloor x_1/2 \rfloor + \cdots + \lfloor x_{2^k-1}/2 \rfloor + x_{2^k}) \bmod 2 \tag{80.5}$$

The first k bits are embedded into LSBs of the first $2^k - 1$ pixel values using the $COV(1, 2^k - 1, k)$ Hamming covering function, and the $(k + 1)$-*th* bit is a function of all 2^k pixels including the appended one. Note that by adding or subtracting one to/from a pixel value x, its $LSB(x)$ always becomes the same binary value $LSB(x) \oplus 1$, however, $\lfloor x/2 \rfloor \bmod 2$, which is the second least significant bit of x, can either be '0' or '1'. Therefore, when Eq. (80.4) does not hold, one pixel value, say, x_i, $1 \leq i \leq 2^k - 1$, has to be changed. By choosing $x_i + 1$ or $x_i - 1$, both Eqs. (80.4) and (80.5) can hold simultaneously without changing x_{2^k}.

On the other hand, when Eq. (80.4) holds but Eq. (80.5) does not, the first $2^k - 1$ pixels need not to be changed, and "Hamming + 1" scheme can modify x_{2^k-1} by randomly increasing or decreasing one to satisfy Eq. (80.5). This means that "Hamming + 1" scheme can embed $k + 1$ bits of message in 2^k pixels with at most one change. This method shows that the embedding efficiency is equal to $(k + 1)2^{k+1}/(2^{k+1} - 1)$, and the embedding rate is $(k + 1)/2^k$.

80.3 Proposed Method

This section proposes new data hiding method, which is called "Hamming + 3". Our proposed "Hamming + 3" improves the "Hamming + 1" scheme, which is a steganographic data hiding method, i.e., "Hamming + 1" scheme embeds $k + 1$ bits into 2^k pixels gray values using at most one change. Our proposed scheme improves the embedding rate compared to "Hamming + 1" scheme, i.e., "Hamming + 3" embeds $k + 3$ bits into $2^k - 1$ pixels gray values using at most 2 change.

80.3.1 "Hamming + 3" Scheme

We propose the following "Hamming + 3" scheme by appending three pixel after the block of Hamming covering function. It embeds $k+3$ bits $(m_1, \ldots, m_k, m_{k+1}, m_{k+2}, m_{k+3})$ into $2^k - 1$ pixel gray values $(x_1, \ldots, x_{2^k})$ using at most two change:

$$(m_1, \ldots, m_k)^T = H \cdot (LSB(x_1), \ldots, LSB(x_{2^k}))^T \tag{80.6}$$

$$(m_{k+1}, \ldots, m_{k+3})^T = H \cdot (\lfloor x_1/2 \rfloor \bmod 2, \ldots, + \cdots + \lfloor x_{2^k}/2 \rfloor \bmod 2) \tag{80.7}$$

The first k bits are embedded into LSBs of the first $2^k - 1$ pixel values using the $COV(1, 2^k - 1, k)$ Hamming covering function [see Eq. (80.6)], and $k+3$ bits are embedded into second least significant bits using the $COV(1, 2^k - 1, k)$ Hamming cover function [see Eq. (80.7)]. Therefore, when Eq. (80.6) does not hold, one pixel value, say, x_i, $1 \le i \le 2^k - 1$, has to be changed. By choosing $x_i + 1$ or $x_i - 1$, both Eqs. (80.6) and (80.7) can hold simultaneously without changing x_{2^k}. On the other hand, when Eq. (80.6) holds but Eq. (80.6) does not, the first $2^k - 1$ pixels need not to be changed, and "Hamming + 1" scheme can modify x_{2^k} by randomly increasing or decreasing one to satisfy Eq. (80.6). This means that "Hamming + 1" scheme can embed $k+1$ bits of message in 2^k pixels with at most one change. This method shows that the embedding efficiency is equal to $(k+1)2^{k+1}/(2^{k+1} - 1)$, and the embedding rate is $(k+1)/2^k$. $[n, k]$ Hamming codes are now a linear space over a field of order q, prime. These q-ary codes are 1-error correcting, relying on the fact that each codeword is at a distance of at least 3 from any other codeword, which in turn relies on the construction of the matrix. Specifically, the fact that no two columns of the check matrix are linearly dependent means that the minimum distance between any two code-words is at least 3.

Proposition *Hamming Codes are 1-error correcting codes.*

Proof We need to check that

$$|C| \cdot \frac{1}{i=0} \binom{x}{y} (q-1)^i = |F_q|^n.$$

The right hand side of this is q^n, where $n = (q^r - 1)/(q - 1)$. The left hand side is

$$\begin{aligned} q^{n-r}(1 - n(q-1)) &= q^{n-r}\left(1 + \frac{(q^r - 1)}{(q-1)}(q-1)\right) \\ &= q^{n-r}(1 + (q^r - 1)) \\ &= q^{n-r}(q^r) \\ &= q^n. \end{aligned}$$

80.3.2 Embedding Procedure

Our scheme is described below in terms of the embedding procedure for hiding secret data in a grayscale image. A cover image is divided into non-overlapping 7-pixel blocks. We present the embedding procedure step by step:

Input: Cover image I sized $H \times W$, a binary secret message δ of maximum length $H \times W - 1$, and the parity check matrix H
Output: A stego image I' sized H × W

Step 1: Divide original images I into $1 \times 2^k - 1$ blocks, letting $c = (b(x_1), \ldots, b(x_{2^k-1}))$, where b(.) denote LSB of a pixel. Letting $c_2 = (\lfloor x_1/2 \rfloor \bmod 2, \ldots, \lfloor x_2^k/2 \rfloor \bmod 2)$. c and c_2 denote code-words and a set of LSB and second LSB respectively.

Step 2: Read all pixels and secret messages into array variable x and δ respectively. $CNT = \lfloor (H \times W)/7 \rfloor$.

Step 3: Calculate the syndrome S by applying Eq. (80.6) to the parity check matrix H and c, i.e., $S = H \cdot (b(x_i), \ldots, b(x_{2^k-1}))^T$, where $i = 1 \ldots n$. Compute $S_1 = S \oplus \delta_j^k$,, where $\oplus$ is XOR operation and $j = 1 \ldots n$, $j = j + k$. As S_1 is the position for 1-error correction, if S_1 is 0 then no flipping any pixel, else flipping a value of $b(x_{i+S_1})$.

Step 4: Calculate the syndrome S_2 by applying Eq. (80.7) to the parity check matrix H and c_2, i.e., $S_2 = H(\lfloor x_1/2 \rfloor \bmod 2, \ldots \lfloor x_{2^k}/2 \rfloor \bmod 2)^{\mathrm{T}}$, where $i = 1 \ldots n$, $i = i + (2^k - 1)$. Calculate the syndrome value for messages, $S_3 = S_2 \oplus \delta_j^k$, where $\oplus$ is XOR operation and $j = 1 \ldots n$, $j = j + k$. If S_3 is 0, then no flipping any pixel, else flipping a value of $(\lfloor x_{i+S_3}/2 \rfloor \bmod 2)$.

Step 5: Decrease CNT by $2^k - 1$. If CNT is greater than 0, return to step 3 to continue the process until there are no more pixels of I.

Example 2 A detailed explanation of the reasons is included in this example. A linear pixels $c = (67\ 79\ 83\ 88\ 91\ 93\ 95)$ is a $1 \times 2^k - 1$ block, reading from left to right and from top to bottom. The secret stream pixels are $\delta = (1\ 1\ 1\ 1\ 1\ 1)$, which is a $k + 3$ block. Calculate $S = (\mathrm{H} \cdot (\mathrm{b}(67)\ \mathrm{b}(79)\ \mathrm{b}(83)\ \mathrm{b}(88)\ \mathrm{b}(91)\ \mathrm{b}(93)\ \mathrm{b}(95))^T) \bmod 2 = (100)$. S is computation using LSB of a block of pixel. $S_2 = S \oplus \delta_j^k = (011)$. Compute $c_{D(S_2)} - 1 = 82$, where $D(.)$ is a function of binary-to-decimal conversion. Next, we show how to conceal secret bits k into second LSB layers. Calculate $S_3 = H \cdot ((\lfloor 67/2 \rfloor \lfloor 79/2 \rfloor \lfloor 83/2 \rfloor \lfloor 88/2 \rfloor \lfloor 91/2 \rfloor \lfloor 93/2 \rfloor \lfloor 95/2 \rfloor)^{\mathrm{T}} \bmod 2) = H \cdot (1\,1\,1\,0\,1\,0\,1) = (0\,1\,0)$. $\mathrm{S}_4 = \mathrm{S}_3 \oplus \delta_j^k = (1\,0\,1)$. Compute $c_{D(S_4)} - 2 = 89$.

80.3.3 Decoding Procedure

Our scheme is described below in terms of the extracting procedure of secret message bits from the stego image. A stego image is divided into non-overlapping 7-pixel blocks. We present the extracting procedure step by step:

Input: Stego image I' sized $H \times W$ and the parity check matrix H
Output: A secret messages δ

Step 1: Divide stego images I' into $1 \times 2^k - 1$ blocks, letting $c = (\mathrm{b}(x_1), \ldots, \mathrm{b}(x_{2^k-1}))$, where $b(.)$ denote LSB of a pixel. Letting $c_2 = (\lfloor x_1/2 \rfloor \bmod 2, \ldots, \lfloor x_{2^k}/2 \rfloor \bmod 2)$. c and c_2 denote codewords and a set of LSB and second LSB, respectively.

Step 2: Read all pixels and secret messages into array variable x and δ respectively. $CNT = \lfloor (H \times W)/2 \rfloor$.

Step 3: Calculate the syndrome S by applying Eq. (80.6) to the parity check matrix H and c, i.e., $S = H \cdot (b(x_i), \ldots, b(x_{2^k-1}))^T$, where $i = 1 \ldots n$. Concatenate δ and S, i.e., $\delta = \delta \| S$. A S denote extracted k bits.

Step 4: Calculate the syndrome S_1 by applying Eq. (80.7) to the parity check matrix H and c_2, i.e., $S_1 = \mathrm{H} \cdot (\lfloor x_1/2 \rfloor \bmod 2, \ldots, \lfloor x_{2^k-1}/2 \rfloor \bmod 2)^T$, where $i = 1 \ldots n$, $i = i + (2^k - 1)$. Concatenate δ and S_1, i.e., $\delta = \delta \| S_1$. A S_1 denote extracted k bits. $j = 1 \ldots n$, $j = j + k$.

Step 5: Decrease CNT by $2^k - 1$. If CNT is greater than 0, return to Step 3 to continue the process until there are no more pixels of I.

80.4 Experimental Results

We proposed a "Hamming + 3" method for data hiding. To prove our proposed scheme is correct, we performed an experiment to verify that it ensures the hidden image can be restored. In addition, the quality of stego image is very important for resisting detection from attackers. Therefore our method is feasible for making good quality stego images from the original grayscale image. To carry out our experiment, 512 × 512 grayscale images were used as cover images. Figure 80.1 is a cover image for experiment to verify our proposed scheme. In our experiments, the qualities of the stego images are measured by the peak-signal-to-noise ratio (PSNR) [14].

The PSNR is the most popular criterion for measuring distortion between the original image and shadow images. It is defined as follows:

$$PSNR = 10 \times \log_{10}(255^2/MSE) \tag{80.8}$$

where MSE is the mean square error between the original grayscale image and the shadow image:

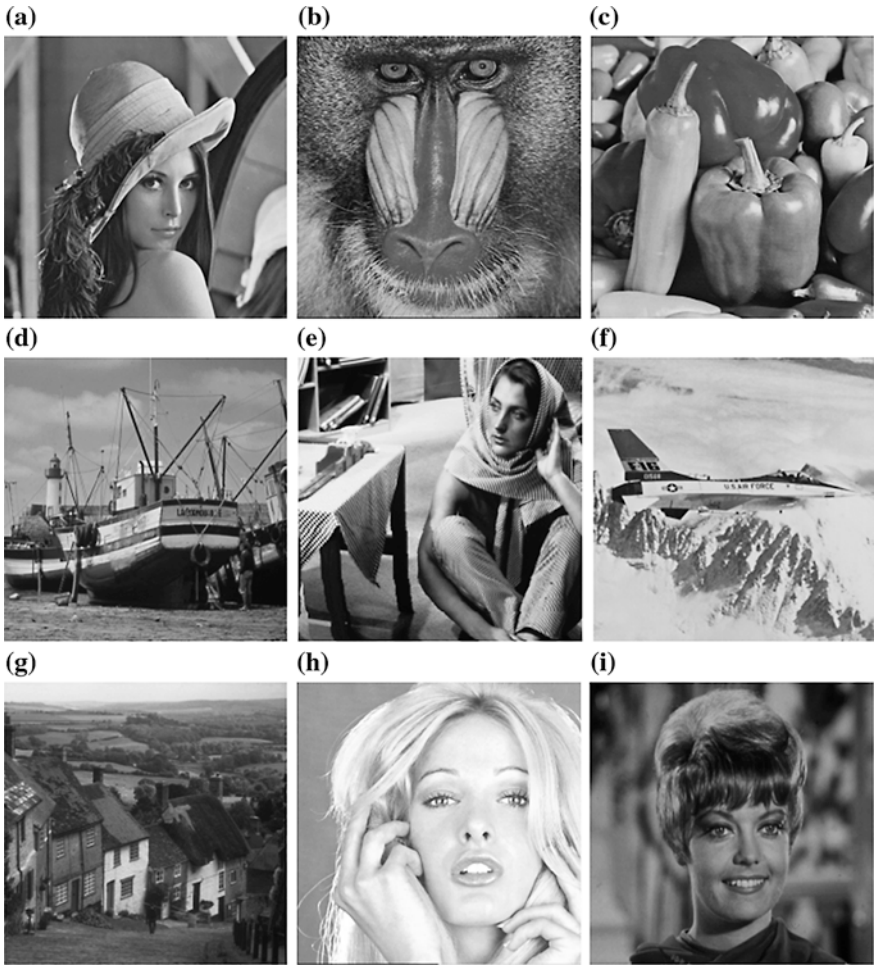


Fig. 80.1 512 × 512 grayscale cover images for data hiding experiment. **a** Lena. **b** Baboon. **c** Pepper. **d** Boat. **e** Barbara. **f** Airplane. **g** Goldhill. **h** Tiffany. **i** Zelda

$$MSE = \frac{1}{m \times n} \sum_{i}^{m} \sum_{j}^{n} [I(i,j) - I'(i,j)]^2 \tag{80.9}$$

The symbols $I(i, j)$ and $I'(i, j)$ represent the pixel values of the original grayscale image and the stego image at position (i, j), respectively; m and n are the width and height of the original image, respectively.

$$p = \frac{|\delta|}{m \times n} (bpp) \tag{80.10}$$

Table 80.1 The comparison result of the matrix encoding, Hamming + 1 scheme and proposed scheme

Images	Method					
	Matrix coding		Hamming + 1		Hamming + 3	
	PSNR	Payload	PSNR	Payload	PSNR	Payload
Baboon	56.44	0.43	53.71	0.499	48.18	0.86
Barbara	54.65	0.43	48.60	0.499	48.22	0.86
Boats	54.75	0.43	49.37	0.499	48.20	0.86
Goldhill	57.02	0.43	53.73	0.499	48.21	0.86
Airplane	55.84	0.43	51.61	0.499	48.20	0.86
Lena	56.05	0.43	52.43	0.499	48.22	0.86
Pepper	54.01	0.43	47.26	0.499	48.22	0.86
Tiffany	53.40	0.43	47.46	0.499	48.20	0.86
Zelda	56.40	0.43	54.04	0.499	48.21	0.86
Average	56.44	0.43	50.91	0.499	48.20	0.86

In Eq. (80.10), p denotes bits-per-pixel (*bpp*), which is an embedding payload. Our experiment compares how many secret bits can be carried by a cover pixel. $|\delta|$ is the number of bits of a secret message δ. There is a tradeoff between a payload and quality of an image. To increase the embedding rate, it is too obvious to require a sacrifice of image quality.

However, if it is possible to keep the balance between payload and quality of an image, we then accomplish our purpose from an aspect of steganography. Table 80.1 shows the visual quality of the stego images created by the matrix encoding, "Hamming + 1", and "Hamming + 3". Our proposed "Hamming + 3" method shows 0.86 bpp with a good visual quality (i.e., the *PSNR* value is higher than 48 dB). From Table 80.1, for the visual quality factor, the matrix coding scheme shows a higher visual quality outcome.

For embedding payload comparison, the proposed "Hamming + 3" show a high embedding payload outcome. Although the visual quality of stego images generated by the "Hamming + 1" scheme is better than the proposed scheme, some images' quality were slightly lower than those of "Hamming + 3". In this experiment, we verified that "Hamming + 3" is worth the steganography method, because our scheme shows reasonable embedding rate and quality as a data hiding scheme. As the *PSNR* of our scheme is over 48 dB, it is not easily detectable by attackers. Therefore, our scheme is highly suitable for various fields of steganography.

80.5 Conclusion

In this paper, we proposed a "Hamming + 3" method that uses both layers, i.e., LSB and second LSB, using cover codes [n, k]. "Hamming + 1" can embed *COV* $(1, 2^k - 1, k)$ at the cost of $1/2^k + 1$ changes. The embedding efficiency is

$(k + 1)2^{k+1} - 1/(2^{k+1} - 1)$. Our proposed scheme shows 0.86 bpp, so "Hamming + 3" has better performance than "Hamming + 1". Moreover, stego images of "Hamming + 3" are over 48 dB, and it denotes that our scheme is a reasonably acceptable steganography method. Thus, we can conclude that the "Hamming + 3" is suitable for steganographic applications.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) by the Ministry of Education, Science and Technology (20120192).

References

1. Kim HJ, Kim C, Choi Y, Wang S, Zhang X (2010) Improved modification direction methods. Comput Math Appl 60(2):319–325
2. Yang CN, Ye G-C, Kim C (2011) Data hiding in halftone images by XOR block-wise operation with difference minimization. KSII Trans Internet Inf Syst 5(2):457–476
3. Crandall R (1998) Some notes on steganography. Posted on steganography mailing list, http://os.inf.tu-dresden.de/westfeld/crandall.pdf
4. Westfeld A (2001) F5: a steganographic algorithm. In: Proceedings of the 4th international workshop information hiding 2001. Lecture Notes in Computer Science, vol 2137, no 1, pp 289–302
5. Bierbrauer J (2005) Introduction to coding theory, Sect. 14.2. Chapman and Hall, CRC Press
6. Galand F, Kabatiansky G (2004) Information hiding by coverings. In: Proceedings of the IEEE information theory workshop 2004, pp 151–154
7. Bierbrauer J, Fridrich J (2006) Constructing good covering codes for applications in steganography. Available: http://www.math.mtu.edu/jbierbra/
8. Fridrich J, Soukal D (2006) Matrix embedding for large payloads. IEEE Trans Inf Secur Forensics 1(3):390–394
9. Schonfeld D, Winkler A (2006) Embedding with syndrome coding based on BCH codes. In: Proceedings of the 8th ACM workshop on multimedia and security, pp 214–223
10. Tseng Y-C, Chen Y–Y, Pan H-K (2002) A secure data hiding scheme for binary images. IEEE Trans Commun 50(8):1227–1231
11. Zhang X, Wang S (2006) Efficient steganographic embedding by exploiting modification direction. IEEE Commun Lett 10(11):781–783
12. Willems F, Dijk M (2005) Capacity and codes for embedding information in gray-scale signals. IEEE Trans Inf Theory 51(3):1209–1214
13. Mielikainen J (2006) LSB matching revisited. IEEE Signal Process Lett 13(5):285–287
14. Chang CC, Kieu TD, Chou YC (2008) A high payload steganographic scheme based on (7, 4) Hamming Code for digital images. In: International symposium on electronic commerce and security, pp 16–21

Chapter 81
Analysis of the Dynamic Characteristics of a Torque-Assist System Using the Model-Based Design Method

Kim Gyoung Eun and Kim Byeung Woo

Abstract In this paper, we present a model-based design method for designing and verifying a control algorithm for a Torque-Assist System in a soft hybrid electric vehicle. The design process can reduce development time and enable improved performance. From the simulation results, it was found that torque variation can be effectively reduced by the proposed control algorithm. By simulating a torque assist system, an appropriate method for controlling a hybrid system with an internal combustion engine and an electric motor can be established.

Keywords Torque assist system (TAS) · Model-based design (MBD) · Automotive simulation models (ASM) · Software-in-the-loop simulation (SILS) · Hardware-in-the-loop simulation (HILS) · Hybrid electric vehicle (HEV) · 42-volt system · Surface permanent magnet synchronous motor (SPMSM) · Interior permanent magnet synchronous motor (IPMSM)

81.1 Introduction

In recent years, with the depletion of fossil fuels and adoption of strict environmental regulations, the automotive industry is greatly increasing investment and support for the production of highly efficient eco-friendly vehicles. In particular,

K. G. Eun (✉)
Graduate School of Electric Engineering, University of Ulsan, Deahakro 93, Nam Gu, Ulsan 680749, South Korea
e-mail: gyg509@naver.com

K. B. Woo
School of Electric Engineering, University of Ulsan, Deahakro 93, Nam Gu, Ulsan 680749, South Korea
e-mail: bwokim@ulsan.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_81, © Springer Science+Business Media Dordrecht 2014

studies are actively conducted on hybrid electric vehicles (HEVs) equipped with a dual electric motors to reduce the loads on an internal combustion engine. Electrical usage in HEVs is gradually increasing as the role of the electric motor is becoming more important in the vehicles. As such, as electricity consumption is increasing, an electrical system having a higher voltage than that of a conventional system is required [1]. When a high-voltage system is installed in a vehicle, it allows for changing hydraulic or mechanical components into electrical components, thus improving the efficiency and performance of the vehicle [2, 3].

Until now, an HEV with a 48V high-voltage system has a hybrid system that directly connects an integrated starter–alternator (ISA) with an idle, stop, and start (ISS) function to an internal combustion engine [4]. This system only drives a vehicle for a few seconds until initial engine startup. The system has a torque-assist function, but the application of such systems to conventional vehicles that use internal combustion engines involves considerable expenses. Therefore, a new motor–starter–generator (MSG) technology with a torque assist function is being developed to achieve high efficiency with low costs [4].

Conventional studies on torque assist systems (TASs) have focused on individual components. Studies from the perspective of the overall vehicle system, however, are very few. Studies have also been conducted based on experiment rather than on the analysis of key component characteristics [1]. Therefore, this study proposes a method for modeling the electric motor (hereinafter referred to as the "motor") and the internal combustion engine (hereinafter referred to as the "engine"), which are the core components of the HEV, using a model-based design (MBD) method that considerably decreases design costs and time when it is applied in conventional vehicles with internal combustion engines. A motor controller was also designed for torque assist to the engine for driving up steep grades. Furthermore, the influence of core components of the TAS on the entire vehicle system including the dynamic characteristics of the vehicle was analyzed rather than the characteristics of the key components.

81.2 Torque Assist System

As shown in Fig. 81.1, the TAS includes a battery, motor, power train, and a dynamic vehicle model. The key components of the TAS were modeled by the Automotive Simulation Models (ASM) program library developed by dSPACE. The ASM program is a MATLAB/Simulink-based third-party program for real-time simulation.

The model configured with the TAS using the MBD method is divided into electrical parts and mechanical parts, i.e., the engine and power train. As shown in Fig. 81.2, the inertial force of the engine is applied to the motor to drive the motor at the same rotation speed as the engine. The motor is modeled such that the torque of the engine is applied to a power train for implementing the assist function.

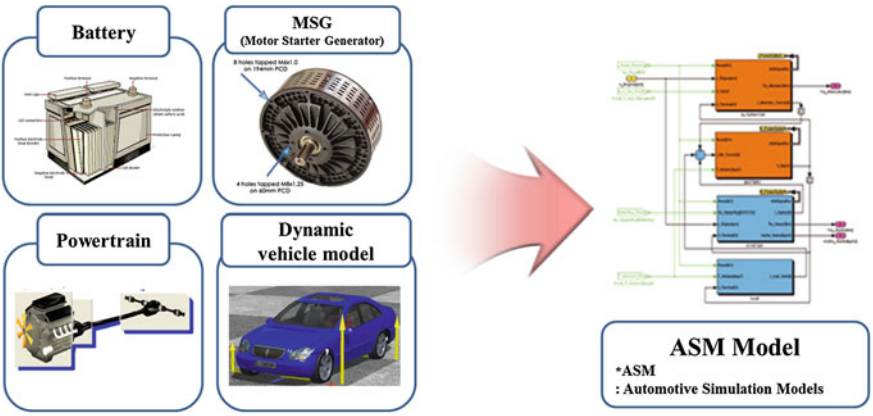


Fig. 81.1 Configuration of an HEV for a torque assist system

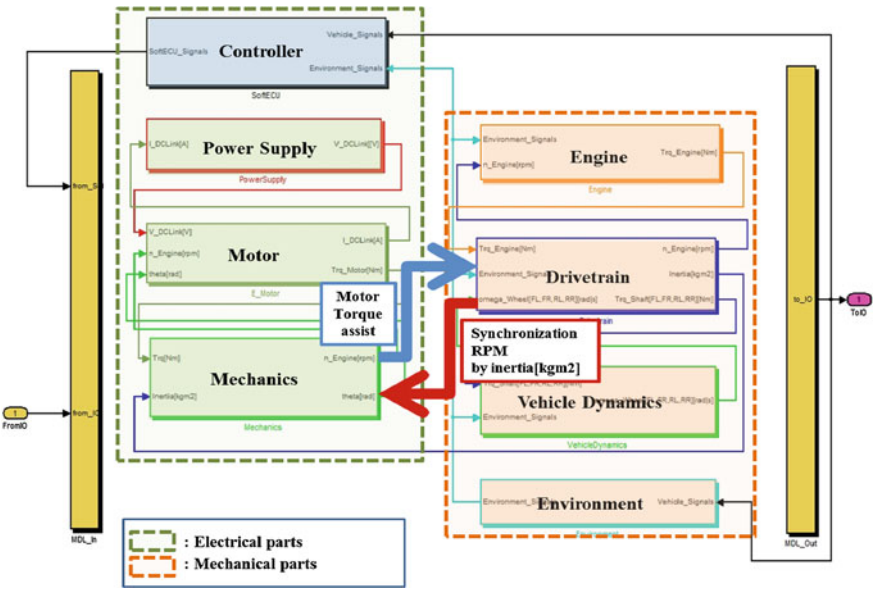


Fig. 81.2 Complete ASM block layout for the torque assist system

81.2.1 Electrical Parts

The electrical parts include a motor (MSG), which is the key component of the electric power train, a battery for supplying power to the motor, and a motor controller (IPMSM controller). The motor is an embedded-type interior permanent magnet synchronous motor (IPMSM) with high output efficiency, and a same volume as that of a conventional alternator.

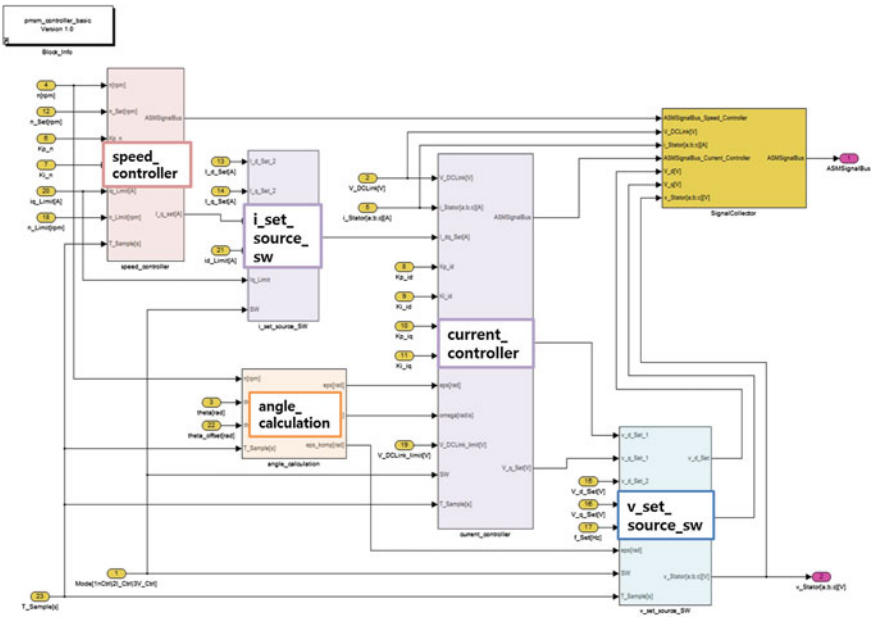


Fig. 81.3 IPMSM controller

The battery was modeled by calculating the state of charge (SOC) by using the following mathematical formula. The initial SOC was set to 80 %.

$$SOC = SOC_0 + i_d \frac{1}{K_N} \int (i_{Battery} - i_{Loss}) dt, \tag{81.1}$$

where SOC_0 is the initial state of charge, K_N is the nominal capacity of the battery, $i_{Battery}$ is the battery current, and i_{Loss} is the loss current.

The quantity $i_{Battery} - i_{Loss}$ represents the state of the battery: if $i_{Battery} - i_{Loss}$ is positive, the battery is charging, and if $i_{Battery} - i_{Loss}$ is negative, the battery is discharging.

As shown in Fig. 81.3, the IPMSM controller model includes proportional integral (PI) controllers for voltage, current, and speed.

81.2.2 Engine

The engine, which is the key component for a mechanical power train, was not modeled by mathematizing the combustion process. Rather, as shown in Fig. 81.4, map data from the Electronic Stability Program (ESP), accelerator pedal position, and practically measured engine speed were used.

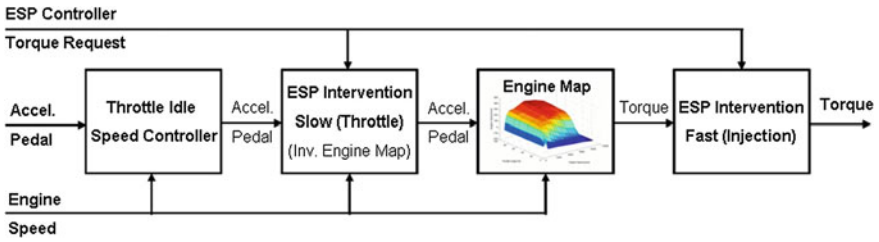


Fig. 81.4 Schematic diagram of the main components in the engine subsystem

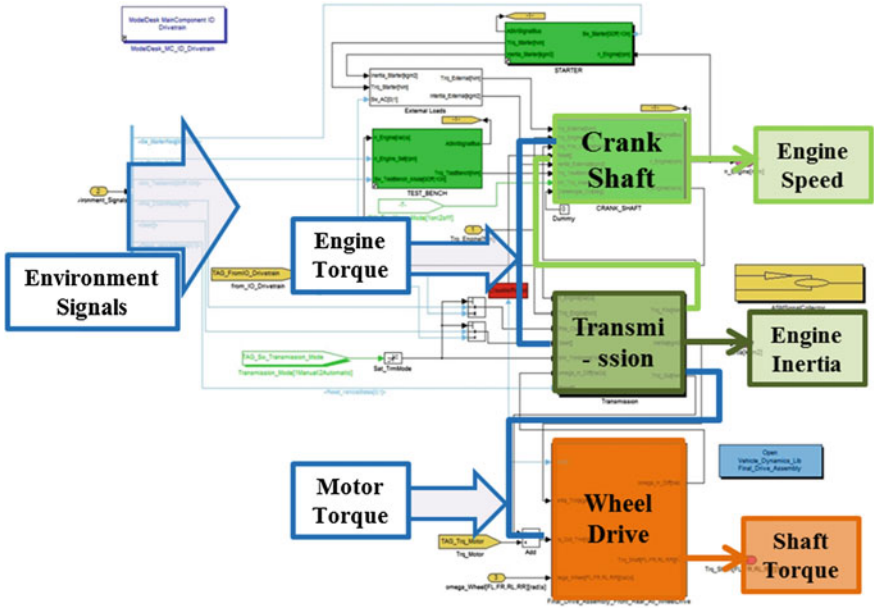


Fig. 81.5 Schematic diagram of a drivetrain

Power generated by the engine of a typical internal combustion vehicles is finally transmitted to the wheels through the power train. The power train transmits power to the wheels through a clutch, gear box, differential and axle drive shaft. In the case of TAS, as shown in Fig. 81.5, the torque of the motor is combined with torque of the engine in the power train and is transmitted to the wheels.

81.2.3 Integrated Controller

As shown in Fig. 81.6, an integrated controller was designed to provide a boosting function that supplements the power of the engine and a power generation function

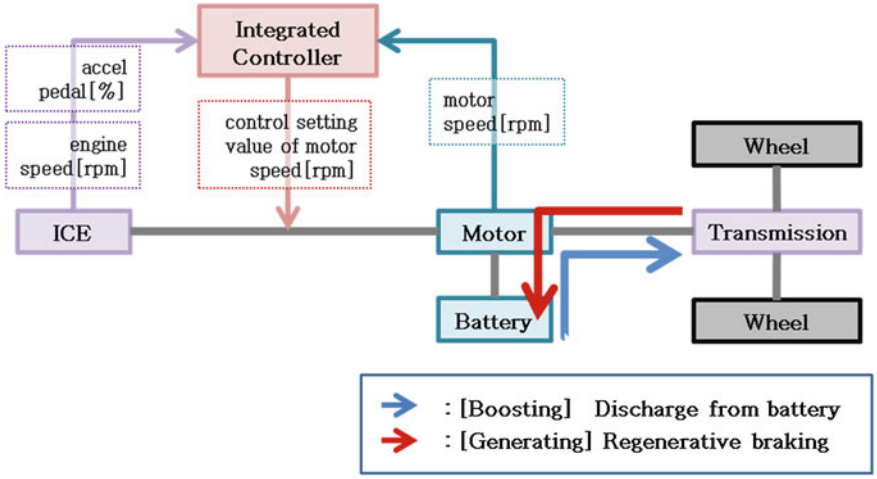


Fig. 81.6 Schematic of the integrated controller

that reclaims energy during deceleration. The support level of the motor torque that is favorable for driving conditions was determined based on the information of the accelerator pedal position, the engine speed, and the motor speed.

For power analysis of a TAS model designed using MBD, a gradient was applied. As shown in Fig. 81.7, the motor controller for the torque assist of the engine was designed considering the power characteristics of an engine and motor. Figure 81.7a shows the torque characteristics map data considering the speed of a motor and engine. It can be seen through the characteristics map data of the engine that the pedal should be depressed more to obtain higher torque output. At the same speed, a higher torque output is required when the vehicle ascends a slope than when it operates on a flat road. Figure 81.7b shows the map characteristics data based on the speed of a motor. It can be seen that the maximum torque output of 40 Nm can be obtained until $\sim$2,400 rpm.

The torque produced by an engine at a specific speed is controlled by the pedal position chosen by the driver. Based on this characteristic, the design considered the maximum efficient torque of a motor in a specific situation, and the characteristic efficiency of an engine to determine when the motor should provide torque assist.

81.3 Simulation and Results

Driving performance was evaluated under various driving conditions considering the effects of the key components of a TAS on a soft hybrid vehicle system. As shown in Table 81.1, according to the structural design criteria for domestic roads,

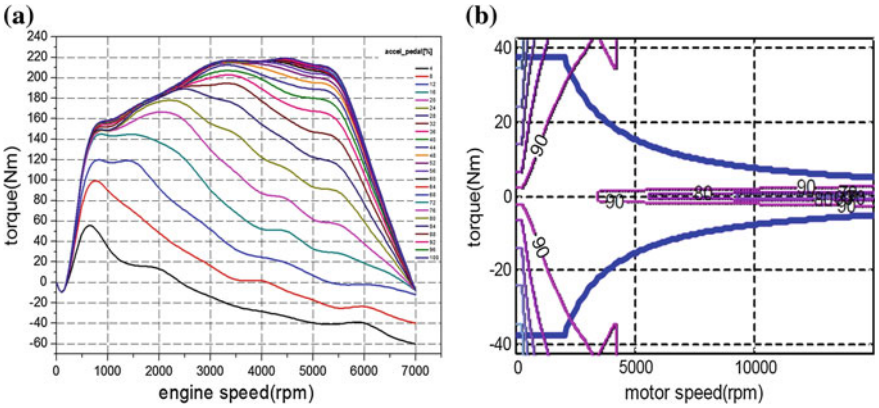


Fig. 81.7 The efficiency map of the engine and motor. **a** Engine map. **b** Motor map

Table 81.1 The maximum of gradient (%)

Gradient of road for compact car				
Design speed (km/h)	Highway		Local road	
	Flatland	Mountain area	Flatland	Mountain area
120	4	5	–	–
110	4	6	–	–
100	4	6	–	–
90	6	7	–	–
80	6	7	–	–
70	–	–	–	–
60	–	–	9	14
50	–	–	9	15
40	–	–	9	16
30	–	–	10	17
20	–	–	10	17

the maximum gradient of a road should be around 4–17 % according to the type of road, topographic conditions, and design speed. Based on this, the road was designed with gradients of 10, 15, and 20 %, which require torque assist. As shown in Fig. 81.8, the gradient was adjusted in straightway 2 by fixing the length at 20 m and changing the height by 2, 3, and 4 m, and driving environment was designed using the information in Table 81.2. According to the driving scenario applied on the designated grade, the vehicle accelerates until it reaches the target speed of 80 km/h from a stationary state, and then it maintains a constant speed.

Figure 81.9a shows accelerator pedal position of a vehicle according to time. During that interval, the accelerator pedal position varies by 10–25 %. Figure 81.9b is a graph the location of the vehicle with respect to time in the simulation on a

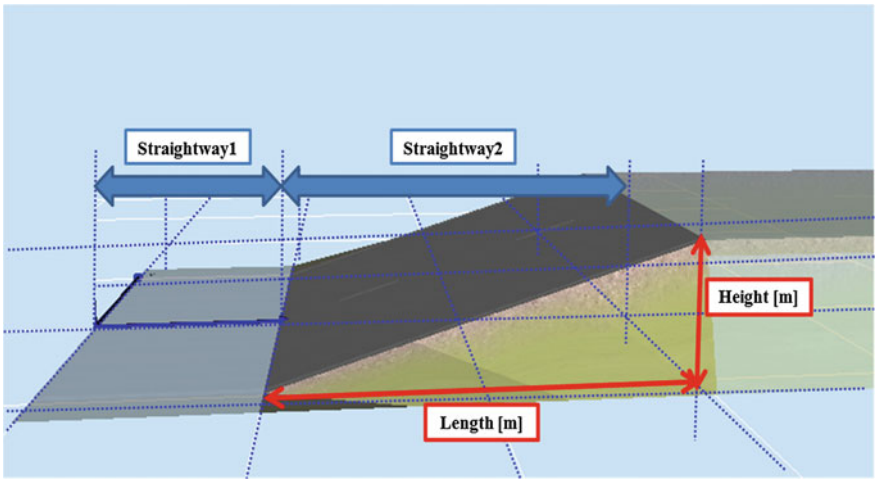


Fig. 81.8 Road condition for the simulation

Table 81.2 Design condition for simulation

Straightway1	Horizontal profile	Length (m)	10
Straightway2	Horizontal profile	Length (m)	20
	Longitudinal profile	Height (m)	2/3/4
		Gradient (%)	10/15/20

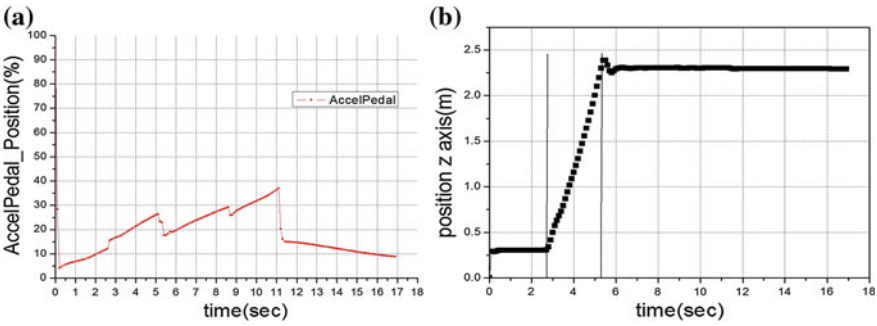


Fig. 81.9 Driving environment for simulation. **a** Accelerator pedal position. **b** Position on the *z* axis (10 % grade)

10 % grade. Figure 81.9b shows the time required to reach the grade from the flat land. Climbing of the grade starts at 3 s, and at 5 s the top of the grade is reached. This means that the torque assist controller had enabled torque assist to climb the vertical grade at the same time irrespective of value of the gradient. Torque

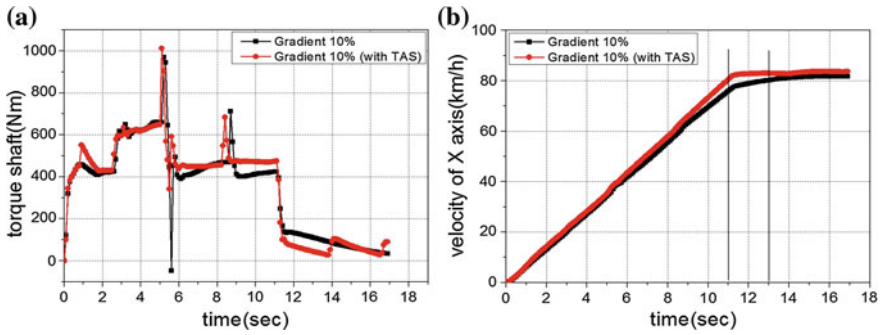


Fig. 81.10 Road conditions for simulation (10 % grade). **a** Torque at shaft (10 % grade). **b** *x*-axis velocity (10 % grade)

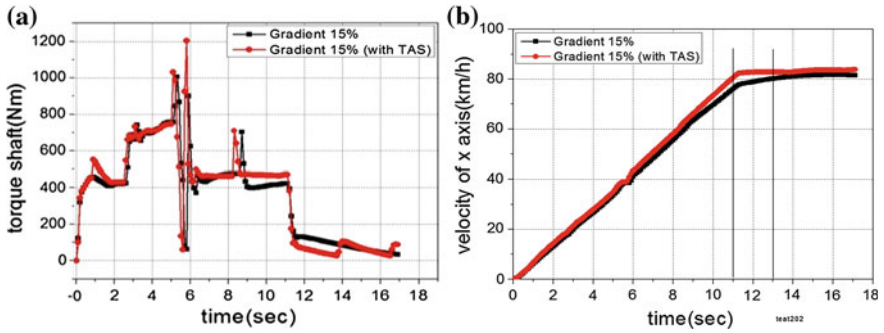


Fig. 81.11 Road conditions for simulation (@gradient 15 %). **a** Torque at shaft (15 % grade). **b** *x*-axis velocity (15 % grade)

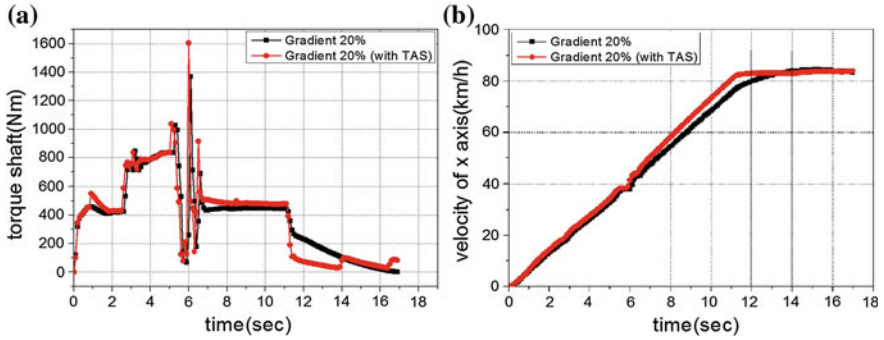


Fig. 81.12 Patterns of road condition for simulation (20 % grade). **a** Torque at shaft (20 % grade). **b** *x*-axis velocity (20 % grade)

variations according to grade can be seen in the analysis by traversing of the same distance in the same time at a target speed of 80 km/h.

In the gradient condition section, cases with and without the integrated controller were compared. The range of torque transmitted to wheels according to the variable position of the pedal becomes wider with the increase in the grade. Figures 81.10a, 81.11a and 81.12a show that the torque assist applied by the integrated controller produces more torque than a conventional drivetrain. Figures 81.10b, 81.11b and 81.12b show that the time required to reach the target speed increases with the increase in gradient; however, in the case with torque assist, the difference in the time required to reach the target speed increases with the increase in the gradient.

81.4 Conclusions

In this study, we proposed a method to model a TAS mathematically by applying the MBD method to a soft hybrid vehicle system. Grade conditions were applied to evaluate the power characteristics of the TAS system. Hence, the performance of the TAS system on the entire vehicle system could be evaluated.

The key components of the TAS were modeled using the MBD method based on an analysis model of the TAS components and a system to evaluate the influence of key electric power components on the overall performance of the TAS system. A motor control method was implemented for evaluation and efficient control of the output generated by the engine and motor of a vehicle equipped with a TAS. The power characteristics of the TAS were evaluated under actual driving conditions and on various grades.

Therefore, the proposed MBD method for the TAS core components and system can be applied to the design of a driving control logic system for a hybrid vehicle and to validate the performance of a hybrid system. In addition, reliability and completeness conforming to international safety standards ISO26262 can be ensured.

Based on this study, future studies will be aimed at a practical validation by comparing experimental data of an actual vehicle equipped with a regenerative braking system in addition to a TAS.

Acknowledgments This research was supported by the Ministry of Knowledge Economy (MKE), Korea, under the Convergence Information Technology Research Center (CITRC) support program (NIPA-2013-H0401-13-1008) supervised by the National IT Industry Promotion Agency (NIPA).

References

1. Ko YG, Cho SH, Park KM, Ro YS (2009) Commercial HEV performance simulation model development. In: KSAE 2009 annual conference, KSAE 09-A0515, pp 3060–3065
2. Kim J, Oh K, Lee J, Kim H (2005) Analysis of fuel economy for a 42-volt ISG vehicle using performance simulator. Trans KSAE 13(3):1–9
3. Kim J, Oh K, Lee J, Kim J, Kim H (2004) Development of 42-volt ISG vehicle drive strategy and simulator. In: KSAE 2004 annual conference, pp 1773–1778
4. Lim WS, Park YI (2005) Development of the performance analysis program of hybrid power transmission system. In: KSAE symposium, KSAE paper number 05-L0023, pp 21–25
5. Lee Y, Kim B (2009) A study on vehicle window regulator control logic using model based design. In: KSAE 2009 annual conference, pp 1936–1940
6. Kwon O, Chai M, Ji J, Kim H (2013) Design of integrated starter generator for parallel CVT hybrid electric vehicle. In: KSAE 2013 annual conference, pp 1834–11837, May 2013
7. Trigui R, Jeanneret B, Malaquin B, Badin F, Plasse C (2007) Hardware in the loop simulation of a diesel parallel mild-hybrid electric vehicle. In: Vehicle power and propulsion conference, 2007. VPPC 2007. IEEE, 9–12 Sept 2007, pp 448–455
8. Lian J, Zhou Y, Ma T, Shen X, Mi M (2009) Design of doubly salient permanent magnet motor for small displacement ISG mild hybrid. In: International conference on mechatronics and automation, 9–12 Aug. 2009, pp 4698–4703
9. Sundstrom O, Guzzella L, Soltic P (2010) Torque-assist hybrid electric powertrain sizing from optimal control towards a sizing law. IEEE Trans Control Syst Technol 837–849
10. Zafeiropoulos S, Di Cairano S (2013) Vehicle yaw dynamics control by torque-based assist systems enforcing driver's steering feel constraints. In: IEEE American control conference, pp 6746–6751

Chapter 82
Improved Branch Prediction for Just-in-Time Decompression of Canonical Huffman Bytecode Streams

Chisoon Jeong, Su Hang and Bernd Burgstaller

Abstract Stack-based virtual machines (VMs) are being employed with embedded systems for their space-efficient instruction format. Bytecode can be compressed to further reduce the memory footprint. A technique introduced by Latendresse et al. (Sci Comput Program 57:295–317, 2005) [13] decodes Huffman-compressed bytecode without a prior decompression step of the program as a whole. Instead, the VM instruction dispatch determines the next opcode from a sequence of table-lookups from the compressed bytecode stream. In this paper we identify indirect branches as a major performance bottleneck of the Latendresse Huffman decoder. We show conclusively that the heuristics of CPU branch predictors are in-effective with Just-in-Time (JIT) Huffman decoding, and we provide a revised decoder where indirect branches have been eliminated. We experimentally evaluate our proposed method as a stand-alone decoder and as part of the instruction dispatch of TinyVM (Hong et al. in Softw Pract Exp 42:1193–1209, 2012) [8]. A representative selection of benchmarks from the MiBench suite (Guthaus et al. IEEE international workshop on workload characterization, WWC-4, pp. 3–14, 2001) [7] showed improvements between 20 and 35 % in overall interpreter performance.

82.1 Introduction

VMs are a viable means to overcome hardware idiosyncrasies with embedded systems platforms. VMs allow for compact program representations: (1) stack-based VMs benefit from space-efficient instruction encodings, because operands

C. Jeong · S. Hang · B. Burgstaller (✉)
Department of Computer Science, Yonsei University, Seoul, South Korea
e-mail: bburg@cs.yonsei.ac.kr

C. Jeong
e-mail: dissu@yonsei.ac.kr

S. Hang
e-mail: suhang@yonsei.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_82, © Springer Science+Business Media Dordrecht 2014

are implicitly addressed by their stack position, and (2) bytecode compression methods can be applied to further shrink the program representation. Bytecode compression is needed with small-scale embedded devices such as sensor nodes, which have severe energy constraints and thus can only provide on the order of a few kilo-bytes of RAM and ROM [16]. Bytecode compression is applicable with VMs such as the Java VM (JVM) and Google Android's Dalvik VM that receive applications over the network, where small program sizes translate to reduced download times.

Decompression of compressed bytecode has been addressed by two main methods: (1) ahead-of-time (AOT) decompression into a memory buffer and executing the inflated bytecode, and (2) just-in-time (JIT) decompression of opcodes and operands during bytecode interpretation. The latter method has a smaller memory footprint because it does not need to inflate the compressed bytecode into a temporary buffer.

JIT-decompression of Huffman-encoded bytecode has been introduced by Latendresse et al. [13]. Latendresse's method is based on an extension of the VM instruction dispatch mechanism such that the next opcode is determined from a sequence of table-lookups from the compressed bytecode stream. Although space-efficient, this JIT decoder technique was shown to incur significant run-time overhead: on recent $\times 86$ CPUs, slowdowns of up to a factor of $4\times$ were experienced with the TinyVM embedded systems VM for C [8]. It should be noted that unlike [13], with the TinyVM experiment the bytecode representation consisted of alternating Huffman codes for opcodes and operands to maximize space-savings; thus, the run-time overheads occurred for decoding of both opcodes and operands.

To improve the efficiency of Huffman decoding, this paper makes the following contributions:

1. We identify indirect branches as a major bottleneck of the Latendresse Huffman decoder on contemporary CPUs with speculative execution. We show conclusively that the heuristics of CPU branch predictors are in-effective with indirect branches and incur a high run-time overhead with Huffman decoding.
2. We provide a JIT decoding method for Huffman-compressed bytecode that does not require indirect branches.
3. We experimentally evaluate our proposed method as a stand-alone decoder and as part of the instruction dispatch of TinyVM. A representative selection of benchmarks from the MiBench embedded benchmark suite [7] showed improvements between 20 and 35 % in overall interpreter performance.

The remainder of this paper is structured as follows: Sect. 82.2 presents background information; Sect. 82.3 presents the related work. We introduce our proposed method on compressed bytecode decoding in Sect. 82.4. We present experimental results in Sect. 82.5 and draw our conclusions in Sect. 82.6.

82.2 Background

TinyVM [3, 8, 11] is an embedded systems VM for the C and nesC programming languages. Its major design goals were execution efficiency and small memory footprints. TinyVM is based on the VMgen [6] interpreter generator and supports the threaded-code and Huffman-compressed bytecode program-representations. Unlike [13], TinyVM's JIT decoder supports alternating Huffman codes for compressed opcodes, numerals and address references.

Huffman coding [9] is a compression algorithm used for lossless data compression. Huffman coding is a part of the JPEG image format and of the MP3 audio format (both formats are lossy, but not on the part of the Huffman code). Huffman code has been applied to compress bytecode of resource-constrained embedded systems VMs [8, 11, 13]. Canonical Huffman code [15] is a Huffman code-variant where the numerical values of the codes of a given length form a consecutive sequence. As observed by Latendresse et al. [13], canonical Huffman code has a compact representation of the bijection between the codes and the encoded objects. This property makes Canonical Huffman code easier and faster to use than "plain" Huffman code.

Pipelined microprocessors, which constitute the main-stream among current HW architectures, show good performance with straight-line code, but less so with branch instructions. Branch targets are resolved late in the processor pipeline and the instruction decoder thus needs to speculate and start execution of the predicted branch target on good faith. If the branch target has been mispredicted, a pipeline stall is incurred. Branches can be classified into two types, direct and indirect branches. Direct branches encode targets directly in the opcode and have a maximum of two branch targets. Indirect branches receive their branch targets from a register value and have thus more than two branch targets, which makes them hard to predict for history-based branch predictors. VM threaded code dispatchers, virtual method calls, switch-case statements and goto statements with "labels as values" of GNU GCC use indirect branches.

82.3 Related Work

To improve VM performance, various approaches to exploit CPU caches and branch predictors have been devised already. Ertl and Gregg [5] reported on the inefficiencies of CPU branch target buffers (BTBs) with the indirect branches of the instruction dispatch mechanisms of threaded-code and switch-based interpreters. To increase the prediction accuracy of BTBs, replication of VM instructions and combining of VM instructions to super-instructions are proposed. Brunthaler's approach from [2] alleviates instruction cache-miss penalties by

re-arranging interpreter instructions such that frequent instruction sequences are co-located in memory. Hot instruction traces are determined by dynamic profiling; a modified Python interpreter showed a maximum speedup of 1.142 on Intel's Nehalem architecture. McCandless and Gregg [14] reordered arms of switch-case statements in the interpreter loop to improve code locality. Various iterative, feedback-directed reordering strategies are proposed and implemented with the Python and Lua VMs. Our proposed method is orthogonal to the before-mentioned work, because it alleviates branch mispredictions in the JIT Huffman decoder and not in the VM instruction dispatch itself. In particular, our approach does not attempt to improve the prediction accuracy of a particular branch, but rather replaces costly indirect branches by direct branches.

Several methods have been devised recently to parallelize Huffman decoding [1, 4, 12]. Because Huffman code is a variable length code, these methods need to overcome the problem of not knowing the start position of codewords in the bit-stream. These approaches do not apply with JIT-decoding of Huffman codes, where the decoded information as a whole is not desired because of memory constraints.

Our proposed method is a variant of the Huffman decoder introduced by Latendresse et al. [13]. Their decoder traverses the Huffman tree through a sequence of table-lookups from the encoded bytecode stream. Their method employs indirect branches to encode transitions between Huffman tables. In contrast, our approach combines lookup tables in one global data structure. We replace each indirect branch by a combination of direct branches and table-lookups. Direct branches have only two outcomes (taken/not taken) and are thus easier to predict than indirect branches. The reduction of mispredicted branches improves decoder performance.

82.4 Proposed Decoding Method

In this section we develop a method that eliminates all indirect branches from the Huffman decoding algorithm of Latendresse et al. [13]. To motivate our approach, we first show that misspredicted branches in Latendresse's algorithm are indeed a huge source of execution-time overhead. By eliminating all indirect branches from the decoder, we increase its performance. The elimination of the majority of misspredicted branches and the associated CPU pipeline stalls results in more consistent performance than with the previous state-of-the-art.

In the following we summarize the decoding scheme for canonical Huffman decoders; for details we refer to Latendresse et al. [13].

Integrating the Huffman decoder into the dispatch mechanism of a VM requires a tree of precomputed lookup-tables. Starting at the root of the binary Huffman tree, those k-bit lookup-tables essentially represent k-bit high sub-trees of the Huffman tree; they are computed by branch-and-bound or greedy methods from [8, 13]. At run-time, the VM's instruction pointer (IP) points to the current position

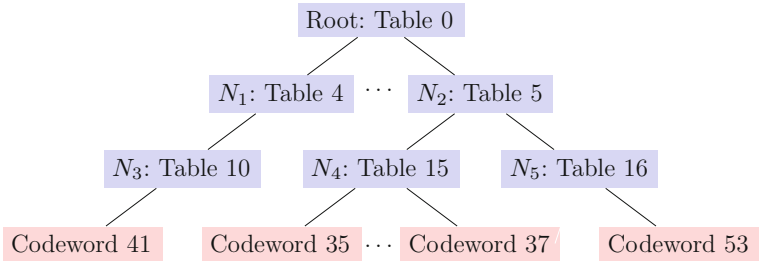


Fig. 82.1 Relevant nodes of the tree of lookup-tables used in the experiment. Table numbers are not consecutive because of nodes that have been omitted for space considerations

in the encoded bytecode stream. The decoder reads the next k bits from the stream and updates the IP accordingly. The k bits from the bitstream are used as an index into the lookup-table at the root. The result of the table lookup represents an address which is used with an indirect branch ("computed goto") to the code-segment that contains the next table look-up.

The decoder mechanism consists of several code fragments with transfer of control being exercised via decoder tables. The following code fragment from the TinyVM decoder illustrates this principle.

```
decode<x>: idx = get_bits(k);
           goto *table<nr>[idx]; ...
code<op>:  shift_left(<k>);
           return op;
```

The decoder starts executing a new codeword in `decode0`. It uses the first k bits of the bytecode stream as an index for the first table-lookup. The computed goto statement transfers control to either a new code segment that will process more bits from the bytecode stream, or to a leave node (i.e., the codeword has been identified). If a codeword was identified, more bits than necessary might have been read and the bitstream pointer is adjusted by `shift_left(<k>)`. Thereafter the opcode value is provided (returned) to the VM instruction dispatcher. As it follows from the above code sample, this decoding method executes a large amount of indirect branch statements. Each goto-statement has up to k branch targets.

A sample tree is depicted in Fig. 82.1. Assuming that the bit-representation for codeword 37 is to be decoded, a Huffman decoder will perform the following steps:

1. a k_{Root}-bit lookup in Table 0, followed by a `goto` to node N_2,
2. a k_{N_2}-bit lookup in Table 5, followed by a `goto` to node N_4, and
3. a k_{N_4}-bit lookup in Table 15, followed by a `goto` to the implementation of codeword 37.

In the above sequence of steps to decode Codeword 37, we identified the goto statements to be detrimental to the performance of the decoder. These goto statements are so-called *indirect branches:* their branch targets are computed

dynamically, depending on the bit-representation of the currently decoded codeword. Indirect branches have a potentially unlimited number of branch targets. E.g., with Node N_4 in Fig. 82.1, any one of the codewords 35–37 are possible branch targets. This is unlike direct branches, which only have two outcomes (i.e., branch taken or branch not taken).

CPU branch predictors commonly attempt to predict the target of a branch from a branch history, e.g., if a goto statement previously resulted in a branch to memory address x, then the branch predictor will predict address x as the outcome for the next execution of that goto statement. However, with the above Huffman decoder, branch targets are independent from the branch history, because they solely depend on the currently decoded codeword. History-based CPU branch-prediction heuristics are thus ill-suited with Latendresse's Huffman decoding mechanism.

To validate the detrimental performance impact of CPU branch predictors on Huffman decoding, we isolated the Huffman decoder from the TinyVM virtual machine. In the following, this decoder is called the stand-alone decoder. With the stand-alone decoder, we decoded a sequence of codewords to measure the decoding time. With each experiment, the sequence of codewords consisted of a prologue of codewords to set up the CPU branch predictor history, and of a single codeword to be decoded after the prologue. Depending on the branch predictor history, different decoding times were expected for a given codeword.

Figure 82.2 shows the codeword sequences that were used to preload the CPU branch predictor for different percentages of correctly predicted branches. For all sequences, the last codeword, i.e., codeword 37, was the one for which we measured the execution time. As follows from the decoding tree in Fig. 82.1, the first sequence leaves the branch predictor in a state where 100 % of all branches are preloaded wrongly for decoding codeword 37. The subsequent sequences preload the branch predictor with 33, 66 and 100 % of correct branch targets. Figure 82.3 depicts the execution times for decoding of codeword 37 for various percentages of correctly predicted branches. As expected, the execution time is the highest for 0 % of correctly predicted branches. Execution times decrease when the number of correctly preloaded branches increase.

Note that the decoding time in Fig. 82.3 varies by a factor of 3× between 0 and 100 % of correctly preloaded branch targets. From this we concluded that the CPU branch predictor has the potential to drastically slow down Huffman decoding on contemporary CPUs.

We have changed Latendresse's canonical Huffman decoding algorithm not to use indirect branches. Our Algorithm is depicted in Fig. 82.1.

	Codeword sequence prologue	Codeword
0%:	Codeword 35 → Codeword 53 → Codeword 41	Codeword 37
33.33%:	Codeword 35 → Codeword 53	
66.66%:	Codeword 35	
100%:	Codeword 37	

Fig. 82.2 Codeword sequences to preload the CPU branch predictor for different percentages of correctly predicted branches. The left-most column denotes the percentage of correctly preloaded branches for decoding codeword 37 from Fig. 82.1. The middle-column denotes the codeword sequence prologue employed to get the CPU branch predictor into the corresponding state. Each codeword from the prologue was decoded twenty times in an attempt to erase the CPU's branch predictor history. After executing the prologue, the execution time for decoding the codeword from the right-most column was measured

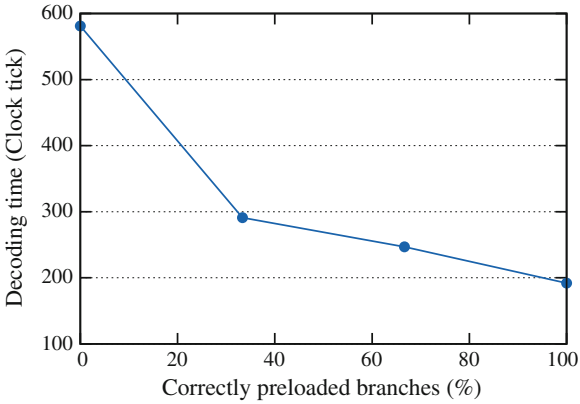


Fig. 82.3 Relation between decoding time and correctly preloaded branches on an ×86 HW architecture

Algorithm 1: Huffman decoder without indirect branches

Input : VM instruction pointer IP into the compressed bytecode stream
Output: Opcode number decoded from the compressed bytecode stream

```
1  rd ← next bytes from the bytecode stream; update IP
2  base ← 0
3  index ← 0
4  crd ← rd >> (w − k_r)
5  val ← (crd >> initial_shift) & initial_mask
6  Entry :
7      index ← base + val
8      base ← opcode_table[ index ]
9      if base < 0 then
10         return −base ;                    // Leafnode (codeword) reached
11     val ← (crd >> shift_table[ index ]) & mask_table[ index ]
12     goto Entry
```

To remove indirect branches, we convert the tree-of-tables data-structure into one global table with subtables. Each subtable starts at a constant offset from the global table. Transfers of control between tables via goto-statements are converted to offset update computations to select the next subtable. Our decoder uses zero value and negative codeword values to distinguish leaf nodes from internal nodes of the decoding tree.

There are two major differences from Latendresse's algorithm to the proposed one. First, we maintain a base index to determine the current subtable within the table of subtables. Bookkeeping information that the Latendresse-algorithm encodes directly in the code segments of sub-tables is kept in dedicated tables (shift-table and mask-table in Algorithm 1). Second, our method uses only one unconditional branch (line 12) and one direct branch (line 9), because referring to the global table of subtables substitutes for jumping to another code-section. Both branches of our method are easy to predict for CPU branch predictors: the unconditional branch has only one target, and the direct branch has two targets.

Our method needs to distinguish op-codes (leaves) from table indizes. As depicted in lines 9 and 10 of Algorithm 1, we use negative values with op-codes to distinguish them from table indizes.

82.5 Experimental Results

To check the validity of our idea, we created a stand-alone decoder from the decoder of our TinyVM virtual machine. This stand-alone decoder was modified according to our proposed idea. After we established the proof of concept with the stand-alone decoder, we implemented our idea with the TinyVM Huffman decoder. We applied the Likiwid Performance tools [16] to access the hardware performance counters of the $\times 86$ CPUs. Likiwid reads the values of the MSR performance counters to check the decoding speed of the stand-alone decoder and the execution speed of TinyVM. This tool was proper to check the relation between improvement of the execution speed and reduction of indirect branch mispredictions. We selected Mibench [7], an embedded benchmark suite, as a representative collection of programs for our experiment. All the benchmark programs we used for the experiment ran on a 1.86 GHz Intel(R) Xeon 5120 CPU, Linux kernel version 2.6.18-308.11.1.el5 and glibc version 2.5. The server has 8 GB of DDR2 FB-DIMM. We compared the decoding speed of the old decoding method and our proposed decoding method on the above machine and the execution speed of TinyVM executables with the old method and our proposed method. Measurements were taken selecting the median value of 10 evaluation results of each benchmark program.

With the stand-alone decoder, we used a sequence of 10 bytecode symbols as an input to both the previous and the proposed stand-alone decoder. The comparison of the decoding times yielded a performance improvement of 24.95 %.

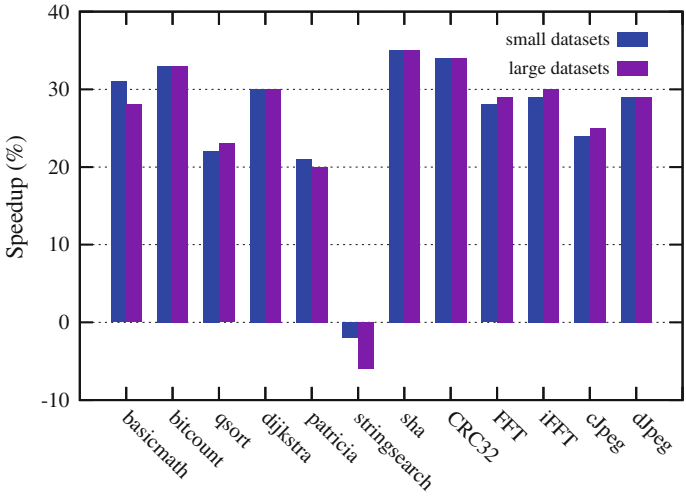


Fig. 82.4 Speedup of TinyVM through opcode decoding with our proposed method and Mibench benchmark suite

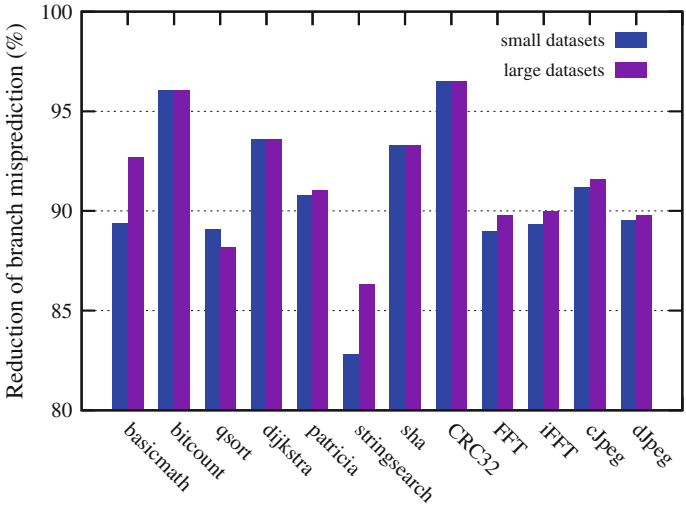


Fig. 82.5 Branch misprediction reduction ratio with our proposed method and Mibench benchmark suite

We implemented our proposed method with TinyVM to check if our proposed method can improve the execution speed of TinyVM as well. We created a Python script that changes the source code of the Huffman compressed bytecode decoder of TinyVM. We used Mibench for the performance evaluation of TinyVM with our proposed method. Figure 82.4 shows the result of the evaluation. Almost all

benchmarks showed more than 20 % of performance improvement. The stringsearch benchmark program did not show speedups, because its main parts consist of machine code. According to Amdahl's law, our improvements with the bytecode of the benchmark program thus cannot improve the overall performance with the stringsearch benchmark program.

We investigated the reduction of mispredicted branches by our proposed method. As depicted in Fig. 82.5, most benchmark programs showed a reduction of about 90 % in the number of mispredicted branches.

82.6 Conclusions

We have proposed a JIT Huffman decoding algorithm by improving the method by Latendresse et al. [13]. We have shown that misprediction of indirect branches constitute a major overhead with [13] on contemporary CPUs. We have provided a revised decoder where indirect branches have been eliminated altogether. We have experimentally evaluated our proposed method as a stand-alone decoder and as part of the instruction dispatch of TinyVM [8], an embedded systems VM for C. A representative selection of benchmarks from the MiBench suite [7] showed improvements between 20 and 35 % in overall interpreter performance.

Acknowledgments Research partially supported by the National Research Foundation of Korea (NRF) funded by the Korean government (MEST) under grant no. 2012K2A1A9054713.

References

1. Biskup MT, Plandowski W (2009) Guaranteed synchronization of huffman codes with known position of decoder. In: *Data compression conference, 2009. DCC '09*, pp 33–42
2. Brunthaler S (2011) Interpreter instruction scheduling. In: Proceedings of the 20th international conference on compiler construction: part of the joint European conferences on theory and practice of software, CC'11/ETAPS'11. Springer, Berlin, pp 164–178
3. Burtscher M, Livshits B, Zorn BG, Sinha G (2010) JSZap: compressing javascript code. In: Proceedings of the 2010 USENIX conference on web application development, WebApps'10, Berkeley, CA, USA, 2010. USENIX Association, p 4
4. Capocelli RM, Gargano L, Vaccaro U (1988) On the characterization of statistically synchronizable variable-length codes. IEEE Trans Inf Theory 34(4):817–825
5. Ertl MA, Gregg D (2003) Optimizing indirect branch prediction accuracy in virtual machine interpreters. In: Proceedings of the ACM SIGPLAN 2003 conference on programming language design and implementation, PLDI '03, New York, NY, USA, 2003. ACM, pp 278–288
6. Ertl MA, Gregg D, Krall A, Paysan B (2002) Vmgen—A generator of efficient virtual machine interpreters. Softw Pract Exp 32(3):265–294
7. Guthaus MR, Ringenberg JS, Ernst D, Austin TM, Mudge T, Brown RB (2001) Mibench: a free, commercially representative embedded benchmark suite. In: 2001 IEEE international workshop on workload characterization, 2001. WWC-4, pp 3–14

8. Hong K, Park J, Kim S, Kim T, Kim H, Burgstaller B, Scholz B (2012) TinyVM: an energy-efficient execution infrastructure for sensor networks. Softw Pract Exp 42(10):1193–1209
9. Huffman DA (1952) A method for the construction of minimum-redundancy codes. Proc IRE 40(9):1098–1101
10. Jang H, Kim C, Lee JW (2013) Practical speculative parallelization of variable-length decompression algorithms. In: Proceedings of the 14th ACM SIGPLAN/SIGBED conference on languages, compilers and tools for embedded systems, LCTES '13, New York, NY, USA. ACM, pp 55–64
11. Kim H, Hong K, Kim S, Kim T, Park J, Ko Y, Burgstaller B, Scholz B, Billy get your guns: fast barrel-shift decoding for in-place execution of Huffman-encoded bytecode streams
12. Klein ST, Wiseman Y (2003) Parallel Huffman decoding with applications to jpeg files. Comput J 46:487–497
13. Latendresse M, Feeley M (2005) Generation of fast interpreters for Huffman compressed bytecode. Sci Comput Program 57(3):295–317
14. McCandless J, Gregg D (2011) Optimizing interpreters by tuning opcode orderings on virtual machines for modern architectures: or: how i learned to stop worrying and love hill climbing. In: Proceedings of the 9th international conference on principles and practice of programming in Java, PPPJ '11, New York, NY, USA, 2011. ACM, pp 161–170
15. Schwartz ES, Kallick B (1964) Generating a canonical prefix encoding. Commun ACM 7(3):166–169
16. Treibig J, Hager G, Wellein G (2010) LIKWID: a lightweight performance-oriented tool suite for $\times 86$ multicore environments. In: 2010 39th international conference on parallel processing workshops (ICPPW), pp 207–216

Chapter 83
Effective Path Modification Using Intuitive Gestures

Hongzhe Liu, Seoungjae Cho, Yulong Xi, Kyungeun Cho and Kyhyun Um

Abstract Human–Computer Interaction (HCI) approaches have been rapidly changing in tandem with advances in computer technologies. The Natural User Interface (NUI) concept is being actively investigated at present because it enables more intuitive and easier access than existing user interfaces. Most of the existing research on path planning aims at automatically generating paths for autonomous driving. However, current automatic path generation approaches have difficulty coping with emergency events during autonomous driving. This paper proposes an interface that overcomes this difficulty by enabling rapid path modification using gestures during autonomous driving. The gesture execution methods, algorithm to recognize gestures, and gesture recognition accuracy are described.

Keywords Human–computer interaction · Natural user interface · Gesture · Path modification

83.1 Introduction

Natural User Interface (NUI) technology is currently being investigated to enable direct interaction with computers using the human body without any intermediary input device. NUI applications utilize recognition of gestures, faces, voice, eye lines, and brainwave. Current research in this area include efforts to implement a virtual fitting room by combining Kinect and augmented reality [1] and control of smart houses using voice [2].

Gesture recognition is one of the representative input methods used by NUI applications. The latest case for successful implementation is video conferencing

H. Liu · S. Cho · Y. Xi · K. Cho (✉) · K. Um
Department of Multimedia Engineering, Dongguk University, Seoul, South Korea
e-mail: cke@dongguk.edu

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_83, © Springer Science+Business Media Dordrecht 2014

based on human gestures [3]. The primary aim of gesture-based video conferencing is to deliver commands to a computer via hand motions. This would enable the conference to proceed as if participants in various places were sitting in the same space using the augmented reality technology.

There are several events that require path planning in daily life. Navigation can be an issue for path search while traveling by car, flight path search, and autonomous path search by robots. Path planning is an essential element in navigation systems as they need to plan paths that allow movement from an initial point to a target point without collision.

Path planning is divided into two kinds depending on the interaction status. One is autonomous drive path planning, in which all the path planning processes are controlled by a computer. The other is path planning with human intervention. Most existing research approaches assume that the autonomous driving takes place in a static environment. Path planning in static environments is defective because it cannot properly cope with unexpected events. However, path planning with human intervention may help to rectify this.

Path planning with human intervention is presently in the research stage. This paper proposes an approach that inspects the collision between a hand and a path in 3D space and dynamically modifies the path in real time in order to facilitate real time path changes in 3D environments.

83.2 Related Work

A gesture is a linguistic system in which the location of a palm and fingers is used as a communication tool. Nielsen proposed an approach in which multiple gestures are defined with independent symbols, each linked to a different function [4]. In the proposed approach, users interact with a computer by combining symbols. However, the disadvantage of this approach is that users need to remember too many combinations for each function and gesture.

Kim proposed a movement method in a virtual space that uses touch gestures [5]. In the proposed method, the coordinates of the point where a finger touches a touch screen are acquired and identified as the navigation coordinates. It facilitates sliding gestures to make an object move to the relevant direction of sliding in virtual space in real time. However, this approach is not easy to use because it has no function to automatically plan paths and so requires too much human intervention.

Conventional path planning is primarily based on autonomous driving conditions and is mainly applied in the robot industry. Zhen-zhong et al. [6] and Mojtahedzadeh [7] adopted fuzzy logic and data convergence technology to autonomously drive a robot. Their approach collects environment data using a Kinect sensor in order to identify the relative path to an object and thereby navigate while avoiding obstacles. However, the approach cannot cope with emergency events because its path planning is based on static environments.

The performance and functions of Leap, a device for collecting external data [8, 9], has been positively analyzed by Weichert, and further used by Vikram to develop a precise handwriting application. These activities indicate that it may be possible to implement a practical gesture recognition system based on Leap.

On the basis of the existing research outlined above, this paper proposes an approach to real time path modification by defining the intuitive interaction approach between a hand and a path in 3D space.

83.3 Gesture Definition for Path Modification

Our event mechanism model is depicted in Fig. 83.1.

The key function of this model is to determine the collision point between a hand and a path in 3D space. The relevant point in the path is subsequently changed in real time by interaction with the collision point in accordance with the user's gesture.

The system first receives hand data from Leap. It then identifies the gesture made by the user from the hand data. When a hand meets a certain point in a path, a collision event occurs and the command corresponding to the user's gesture is executed. After the implementation of the command, the final path is visualized to the user.

Figure 83.2 illustrates the process used to determine whether collision has occurred between a hand and a path.

To determine whether collision has occurred between a hand and a path, a bounding sphere is set at the coordinates corresponding to each finger. This bounding sphere is also set to each valid point on a path. When the bounding sphere corresponding to a finger meets that of each valid point on a path, it is determined that a hand and a path have intersected, and a collision event then occurs.

When a collision event occurs, the commands vary depending on the kinds of gestures. Table 83.1 illustrates the kinds of gestures used.

A point gesture generates a new valid point on a path. When a user touches a specific point on a path using a point gesture, a new valid point is generated at the relevant point.

In addition, a user can generate a new path on a path using a sliding gesture. When a user touches a specific point on a path with the sliding gesture and subsequently touches another point, a path connecting the two points is generated.

The scissors gesture divides the exiting path into two parts. When a user touches a specific point on a path with the scissors gesture, the path is divided into two parts around the specific point touched by the user.

The drag gesture moves the valid points on a path. When a user touches a valid point with the drag gesture, the relevant valid point follows the hand from the point that is touched. Then, the user can move the valid point to any desired point.

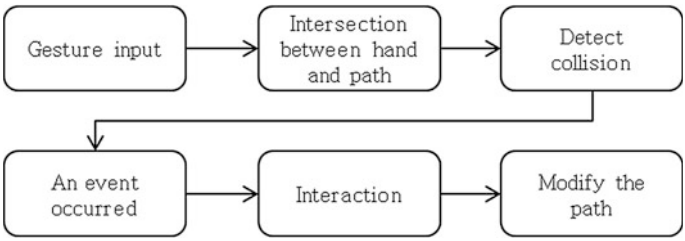


Fig. 83.1 Event mechanism model

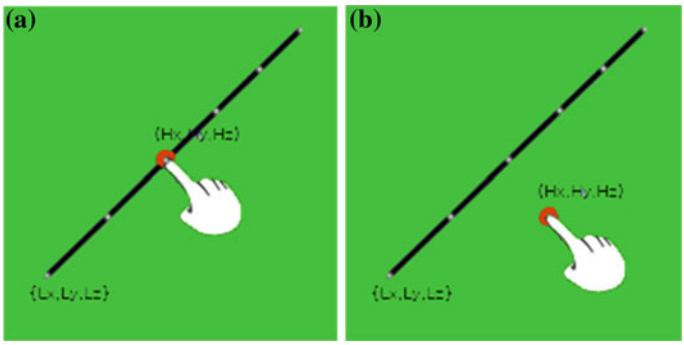


Fig. 83.2 Process used to determine whether collision has occurred between a hand and a path: **a** collision; **b** no collision

Table 83.1 Kinds of gestures

Name	Point	Slide	Scissors	Drag	Stop
Icon					

Fig. 83.3 Pseudo code for movement of valid point using drag gesture

```
TranslateValidPoint () {
  Success ← False
  FOR i ← 0 To ValidPointCount - 1
    IF LinePosTranslate[i] = Yes THEN
      Translate(ValidPointPos[i])
      ValidPointPos[i] ← HandPos(Hx,Hy,Hz)
      Success ← True
    END IF
  END FOR
  RETURN Success

}
```

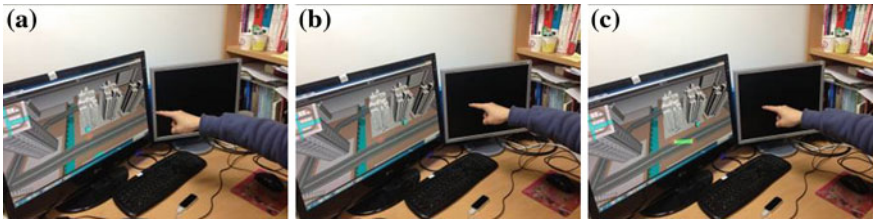


Fig. 83.4 Point gesture experiment scenes. **a** 1st collision. **b** 2nd collision. **c** Path generation between both points

In this paper, we describe a path modification method that uses drag gestures. Figure 83.3 depicts the pseudo code used for moving the valid point by recognizing the drag gesture.

If the hand doing the drag gesture intersects with a specific valid point on a path, the value of LinePosTranslate[i] becomes Yes, then the position of the valid point is replaced with the position of the hand. This procedure is repeated in every frame, allowing the valid point to move along with the hand.

When a user makes the stop gesture, any gesture that is being executed at that moment stops.

83.4 Experiments and Analysis

We implemented our application that enables generation and modification in a 3D virtual space as an experiment in path modification using gestures. The scenes in which each gesture was executed were captured using the application.

Figure 83.4 presents the point gesture experiment scenes taken.

Figure 83.4a shows the generation of additional valid points on a path using the point gesture. Figure 83.4b illustrates the addition of a valid point on a different point. Finally, Fig. 83.4c depicts the automatic generation of a path connecting two valid points.

Figure 83.5 presents the slide gesture experiment scenes.

As shown in Fig. 83.5a, the starting point of the slide was selected using the point gesture. Figure 83.5b shows the hand moving to the ending point. As soon as the hand moves, the path is generated. The final path is determined by the stop gesture.

Figure 83.6 presents the scissors gesture experiment scenes.

As illustrated in Fig. 83.6a, the point is not divided yet. Next, two fingers are outstretched like scissors, as shown in Fig. 83.6b, and the path is divided into two parts with the relevant point as the basis.

Figure 83.7 illustrates the drag gesture experiment scenes.

Figure 83.7a shows a collision with the valid point to move using the drag gesture. Then, as illustrated in Fig. 83.7b, the relevant valid point is moved.

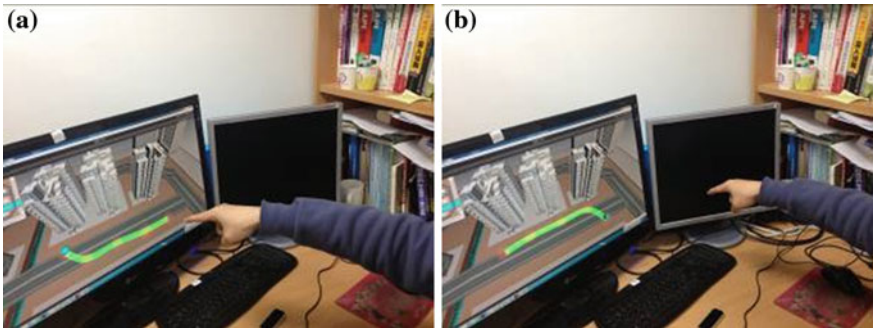


Fig. 83.5 Slide gesture experiment scenes. **a** Selecting the starting point. **b** Path generation after slide gesture

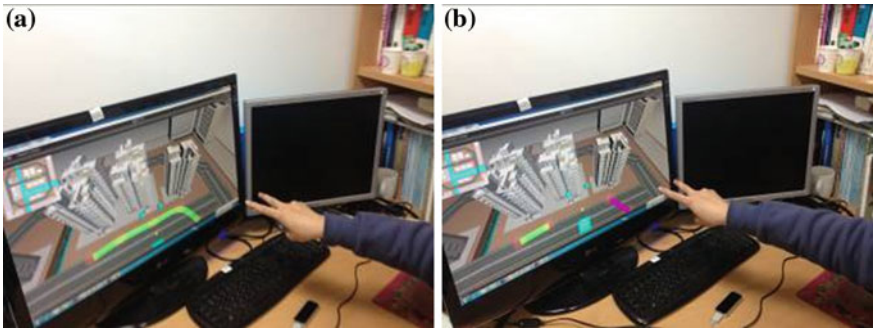


Fig. 83.6 Scissors gesture experiment scenes. **a** Before path separation. **b** After path separation

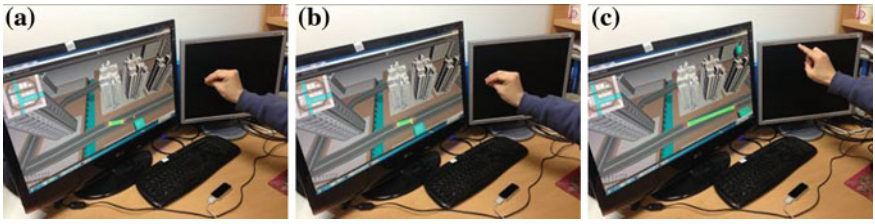


Fig. 83.7 Drag gesture experiment scenes. **a** Before drag. **b** Dragging. **c** After drag

Figure 83.7c presents the result of path modification to make the existing path pass through the valid point moved.

Table 83.2 is the analysis results for the recognition rate after each gesture was executed several times.

From Table 83.2, it can be seen that the lowest recognition rate is 80 %, which is the recognition rate of the slide gesture. This low rate occurred because of the

Table 83.2 Recognition rate analysis by gesture

Gesture	No. of executions	No. of executions recognized	Recognition rate (%)
Point	50	47	94
Slide	50	40	80
Scissor	50	42	84
Drag	50	47	94
Stop	50	50	100
Average	50	45.2	90.4

similarity between simply making a hand moving gesture and the slide gesture. However, the overall average recognition rate for all gestures is 90.4 %. These results confirm the possibility of effective and intuitive path modification using the gestures above.

83.5 Conclusion

This paper proposed several gestures to dynamically modify paths in real time. The results of experiments conducted confirm that paths can be modified using intuitive gestures without any intermediary input devices. Accordingly, users can easily modify paths even after learning for only a short time. The gestures for path modification can be applied to a number of applications. The gestures can be applied to education programs for children who are not accustomed to computers or for simulations of army movement path planning in real time.

The algorithm for recognizing slide gestures has a relatively low recognition rate, and therefore must be improved to recognize the gesture more distinctively. In addition, the gestures tested in this paper need to be developed into a standard gesture template that can be applied in more and varied fields with additional intuitive gestures.

Acknowledgments This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2013-H0301-13-4007) supervised by the NIPA (National IT Industry Promotion Agency).

References

1. Rong L, Kong Z, Xiang X, Yueqiao L, Zan L (2011) Research of interactive 3D virtual fitting room on web environment. In: 2011 fourth international symposium on computational intelligence and design (ISCID), vol 1, pp 32–35
2. Abid MR, Melo LBS, Petriu EM Dynamic sign language and voice recognition for smart home interactive application. In: 2013 IEEE international symposium on medical measurements and applications proceedings (MeMeA), pp 139–144

3. Gelb D, Subramanian A, Tan K (2011) Augmented reality for immersive remote collaboration. To be published and presented at IEEE workshop on person oriented vision (POV) 2011, Kona HI, 7 Jan 2011
4. Nielsen M, Störring M, Moeslund T, Granum E (2004) A procedure for developing intuitive and ergonomic. In: 5th international gesture workshop, pp 409–420
5. Kim J, Gracanin D, Matkovic K, Quek F (2009) iPhone/iPod touch as input devices for navigation in immersive virtual environments. In: IEEE proceedings of virtual reality conference 2009, March 2009
6. Zhen-zhong Y, Wei-cou Z, Xin L, Jing H (2013) Real-time local path planning for mobile robot based on Kinect. Proc Comput Eng 39(4):243–247
7. Mojtahedzadeh R (2011) Robot obstacle avoidance using the Kinect. Master's thesis in computer science (30 ECTS credits) at the systems, control and robotics Master's Program Royal Institute of Technology year 2011
8. Weichert F, Bachmann D, Rudak B, Fisseler D (2013) Analysis of the accuracy and robustness of the leap motion controller. Sensors 6380–6393
9. Vikram S, Li L, Russell S (2013) Handwriting and gestures in the air, recognizing on the fly. In: CHI 2013 extended abstracts, 27 April–May 2, 2013, Paris, France, ACM 978-1-4503-1952-2/13/04

Chapter 84
DSRC Based Self-Adaptive Navigation System: Aiming Spread Out the Vehicles for Alleviating Traffic Congestion

Hohyeon Jeong, Jung-Won Lee, Jaehoon (Paul) Jeong and Eunseok Lee

Abstract In the past decades, traffic problems has become a major research field of both academic and industrial community. In this paper, we propose a Self-Adaptive Navigation System (SeNS) to provide optimal tailored driving paths, reflecting real-time road traffic information. The current navigation systems are using server independently, and only use the static time stamp-based traffic information. For not consider the changing of traffic in near future, they tend to provide the path that sharing some roads. This kind of individual optimization could—create new bottlenecks on some roads, which was light traffic. Our proposed system consists of Traffic Control Center (TCC), Road Side Unit (RSU), mobile terminal (e.g. smartphone). TCC and RSU are for collecting road statistics, merging them and generating overall traffic information, in real-time. Based on this, mobile terminal can provide tailored optimal path, apply the algorithm of maximizing the traffic flow, for alleviating traffic congestion on target area of TCC. Therefore, based on the prompt exchange of road traffic conditions through the interaction among the mobile terminals, RSUs, and TCC, SeNS can provide drivers with a new type of self-adaptive navigation service, based on dynamic traffic conditions, such as temporal traffic jam and car accidents.

H. Jeong (✉) · J.-W. Lee
Department of Electrical and Computer Engineering, Ajou University,
Suwon, Republic of Korea
e-mail: good92489@ajou.ac.kr

J.-W. Lee
e-mail: junwony@ajou.ac.kr

J. (Paul)Jeong · E. Lee
Department of Computer Engineering, Sungkyunkwan University,
Suwon, Republic of Korea
e-mail: pauljeong@skku.edu

E. Lee
e-mail: leees@skku.edu

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_84, © Springer Science+Business Media Dordrecht 2014

Keywords Intelligent transportation system · Adaptive navigation system · Self-adaptive software · Dedicated short range communications · Collective intelligence

84.1 Introduction

In modern society, vehicles are very common for daily transportation. The expansion of vehicle's usage and propagation is getting faster daily, hence, road networks are getting more complex and bigger. Therefore, navigation systems are highly required. The existing navigation systems' feature as follow.

First, existed navigation systems update their traffic information by constructing their own infrastructure [1] or reporting the road statistics report, made by drivers. The former one may be expensive to construct and maintain, but it can be more accurate. And the latter one is fully depend on their number of users and frequency of report. Second, both of mentioned collecting methods use mobile network. It can be expensive if raise up the frequency of update for fresh traffic information [2]. And also, it can be throw down the QoS of default service of mobile network (e.g. calling). Third, these existed systems use the independent system for each other. So, they provide the path that sharing some roads, depend on options (e.g. fastest, shortest, ECO driving [3]). This individual optimization can make new bottleneck on provided path.

These kind of systems are difficult to react to dynamic road traffic network instantly. Therefore, it is hard to aware the traffic situation, spread out the vehicles and use the road network effectively. Thus, these existed systems should provide more satisfaction to get their destination for driver. The satisfaction depends on how efficient we spread out the vehicles and how fast to get to their destination. In this paper, we propose Self-Adaptive Navigation System that has objectives as follows: (1) Providing traffic information to drivers in a fast and accurate way by constructing SeNS, adapts to dynamic traffic situations. (2) Architecting DSRC-based infrastructure in order to collect traffic information efficiently. (3) Providing a routing algorithm to compute a path with minimum travel time by using vehicles' travel paths.

84.2 Related Work

84.2.1 Infrastructure for Generate and Collect the Traffic Information

Recently, there are various infrastructures implemented in road network for providing navigation service. Such infrastructures can be roughly divided into private infrastructures and public infrastructures. In case of private infrastructures, they communicate with terminal by using their own server and mobile network. WAZE

[4], TOMTOM [5], TrOASIS [6], T Map [7] are the navigation system that use mobile network (TOMTOM connect with network via bluetooth tethering). Because of this, communicating traffic information between server and terminal, using quiet large amount of data, can be decline the QoS that default service of mobile network. And the case of public infrastructure, Traffic control center that managing traffic information, collect traffic information by communicating with loop detector, vehicle ID recognition unit, Road side Equipment (RSE, or Road side unit, RSU) [8, 9], CCTV. And these are connecting with optical network [10] and using another bandwidth. Therefore, even if traffic information's size is heavy, they can decrease the workload of mobile network [11].

84.2.2 Generate and Collect the Traffic Information

Basically, the navigation systems has GPS module for aware the information of current location and speed in terminal. Through this terminal, terminal can give the raw data to server. First of all, the information received from server is traffic information that provide which road has been jammed in visually to driver [12]. Next, the information transmit to server is collectable data from vehicle. Such as, requesting find new path, current location and current speed. Than the methods of collecting this kind of information is report the road status directly wrote by user and report the vehicle's status automatically. The former one should be non-periodic way, because this way is fully depend on the user. And the latter one should be periodically report to server. Additionally, the method that wrote by user directly, may have high probability of accident by using device in driving.

84.2.3 Traffic Information Providing Service

Through the navigation system, user can take two big services. First, user can take traffic information in terminal. This traffic information induces which road is congested and fluid to user. Second, user can take the trajectory from current location to destination. The calculation of path can be operate in server and terminal (TrOASIS does not support calculating path). Via this service, user can get the traffic information. Such as shortest, fastest path, CCTV stream, report from other users.

In this paper, we design the navigation system that aware and self-adapt the traffic information by using Road Side Unit, easy to deploy and collect information, and automatic report periodically for concentrate on driving [13]. Additionally, finding the optimal path will be implemented in terminal to reduce the workload of server. When re-routing, existed navigation systems handle the re-route request from user directly. So, it can hardly consider about user or designer's preference. Therefore, we propose the self-adaptive navigation system that reflect

the intention through utility function, which include the experience of user or experts [14], when re-routing. And also, propose the system, which can provide optimal fastest path that effectively spread out the vehicles.

84.3 Self-Adaptive Navigation System

84.3.1 System Requirement

Table 84.1 is the comparison of existed navigation system mention in Sect. 84.2. Most right column shows the requirement of our system. It represents the limitation of workload in mobile network that most of existed navigation system used, the limitation of updating traffic information in social network based navigation system, like WAZE, the possibility of navigation system that based on wireless network, such as DSRC-based RSU, in growing ITS [15]. To overcome the limitation of vigation system, we have goals such as, present the merged traffic information made by modelling the road network, guarantee the fastest path avoid the congested road, aware the dynamic traffic information in real-time for auto-re-routing that reflect the intention of user or designers.

Therefore, we propose the system that use DSRC-based TCC-RSU-Vehicle architecture which is being built at the national and dynamically adapt the traffic information, keep changing. Our proposed system is more fast and accurate, generate the traffic information effectively, by using these, provide the tailored optimal path that fastest and spread out the vehicles.

84.3.2 Design

The components of proposed Self-Adaptive Navigation System has goals, functions and roles as follows. Figure 84.1 shows the interaction among the components in order to achieve goals.

- **Objective 1** Managing and Providing Overall Traffic Information

TCC (Traffic Control Center) has goal of generating overall traffic information of target area. To accomplish this goal, TCC has function of communication with all of RSU, deployed on target area of TCC, and transmission the traffic information, corresponding to request of vehicles. Traffic information, modelled at TCC, has a goal of including the road traffic network of target area. Hence, TCC has a role of merging the overall traffic information of target area and providing overall traffic information to vehicle.

- **Objective 2** Managing each Road Statistics

Table 84.1 Comparison of existed navigation systems

Navigation system Infrastructure		Waze Internet	TOMTOM Public infra	TrOASIS Internet	T map Private infra	Proposed method Public infra
Generating and collecting traffic information	Source	Report from client device	STI, device signal	Report from client device	Report from client device	Report from client device
	Collecting method	User's report	• STI • Anonymous phone signal	• User's report • Automatic gathering[b]	• User's report • STI	• Automatic gathering
	Update period	Non-periodic	2 min	Non-periodic	5 min	less than 1 min
Providing traffic information service	Service range	Traffic information, source-destination route	Traffic information, source-destination route	Traffic information only	Traffic information, source-destination route	Traffic information, source-destination route
	Information type	• Shortest/fastest path	• Shortest/fastest path	• CCTV • STI • user report	• Shortest/fastest path	• Shortest/fastest path/ prediction
Self-adaptiveness	Real-time service	Medium[a]	High	Medium[a]	Low	High
	Re-routing event handler	Manual	Automatic	Manual	Manual	Automatic

[a] If report from user frequently
[b] Drive direction and speed

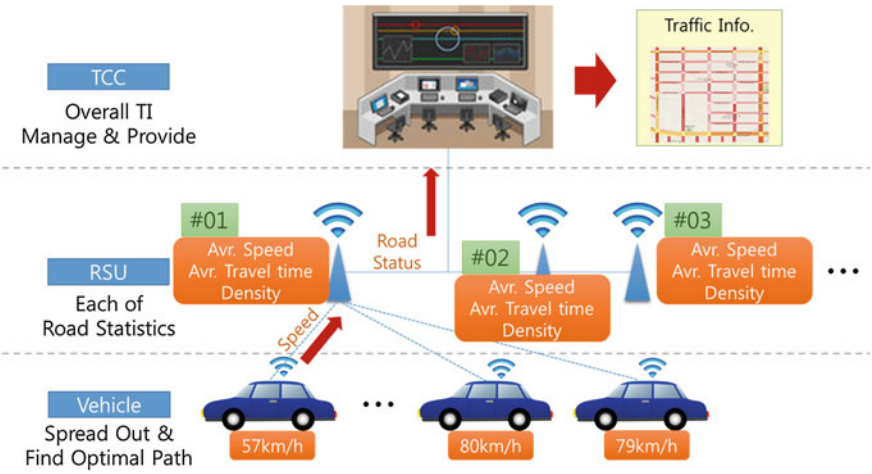


Fig. 84.1 Interactions among components

RSU (Road Side Unit) has a goal of generating road traffic statistics. To accomplish this goal, RSU directly communicates with vehicles via DSRC. And RSU deployed more than one unit in order to fully cover the road segment. Therefore, RSU has function of receiving speed data from vehicle and generating road traffic statistics based on received data and transmitting road traffic statistics to TCC. Hence, RSU has a role of collecting data of road segments and processing collected data and transmitting the processed data to TCC.

- **Objective 3** Spread Out the Vehicles and Find Optimal Path

Mobile Terminal installed on vehicle (smartphone or tablet) has a goal of delivering measured speed to near-by RSU and provide optimal path to user based on overall traffic information from TCC. To accomplish this goal, mobile terminal has 802.11 p module for DSRC. And mobile terminal has function of measuring speed using GPS module and providing optimized fastest path based on traffic information. Hence, mobile terminal has a role of sensing the vehicle's information and providing optimal path adapt on dynamic traffic information. As a guide, every vehicles has their own mobile terminal.

Therefore, SeNS are based on progress as follow: Certain RSU, deployed at near-by certain road segment, is continuously collecting unknown vehicle's speed. At this time, several vehicles are pass by this road segment, than vehicle's mobile terminal generate the speed data. And this speed data transform to packet and transmitted to RSU. After receive the speed data at RSU, than a RSU generate road traffic statistics (This RSU has a role of representing certain road segment and merging the traffic information, generated by another RSUs in same road segment). And transform the road traffic statistics to packet then transmit to TCC. TCC receives road traffic statistics from all of RSUs in target area, then TCC

merges those statistics for overall traffic information. Finally, certain user request routing to their mobile terminal, the mobile terminal request the TI to corresponding TCC, and searching optimal fastest path based on the received TI. Additionally, when vehicle breakaway the path or congested the path, mobile terminal schedule the rerouting, same sequence of when we routing.

84.3.3 Routing Algorithm

In this section, describe about routing algorithm of SeNS. First of all, we modelling the TI at the TCC. Than searching optimal path based on model of TI at the vehicle.

84.3.3.1 Traffic Information Modelling

To architecting proposed SeNS, the collected traffic information modeling technique is necessary. So, we choose the modelling method of the Ford-Fulkerson Algorithm (FFA) that maximize the flow of network based on Min-Cut Max-Flow Theory [16]. As follow the FFA's modelling method, TI is modelled as directed graph at TCC. Intersection as node and road segment as edge. The edge, link between two nodes, has weight of capacity and flow. Capacity represent the maximum number of vehicles in a road segment at a time. Flow represent the number of vehicles in a road segment at a time. For example, certain road segment has full with vehicles and number of vehicle is 10. When same road segment, at another time, has 4 travelling vehicles, the capacity and flow will be 10 and 4. The node, representing an intersection, will occur more than one. And these nodes are assigned with number of approach roads. For instance, if there are four-way intersection, four nodes will generated. Because, four-way intersection has four approach roads.

84.3.3.2 Searching Optimal Path

When user request routing, mobile terminal request the TI from TCC and searches optimal path. Searching several path to destination, then calculate and compare their estimated travel time, as flow, each other. Finally the fastest one will selected. First of all, mobile terminal receive the TI then find the source node and destination node, input from user. Next, searches all of possible path on received TI model, using depth-based search algorithm. Finally, calculate those path's estimated travel time and select fastest one for provide to user.

84.3.3.3 Calculate Estimated Travel Time

Generally, estimated travel time is summation of average link delay of each road segment on certain path. But, this method cannot handle the traffic, which will dynamically change in future. Therefore, we can handle the dynamic traffic by applying prediction methods. In this paper, we use k-NN methods for calculate the estimated travel time based on stored TI. If the estimated travel time is greater than capacity, then estimated travel time will be assign to capacity. Through this, SeNS can searching optimized path, reflecting dynamically changed traffic.

$$Predicted\ flow = Average\left(flow(t)_i + \sum_{j=1}^{k} flow(t-j)_i\right) \quad (84.1)$$
$$(t = current\ time\ stamp, i = ID\ of\ road\ segment)$$

84.3.4 Re-routing Algorithm

In the mobile terminal keep receiving traffic information, re-routing should be initiated by driver's necessary. Re-routing can be initiated when deviation from the path and decline of satisfaction of path. The former one operate by comparing between vehicle's current location and path information. And the latter one operate when the estimated travel time drop down the certain threshold. This means that estimated travel time goes longer. System designer can define the relationship between satisfaction and estimated travel time by using utility function (e.g. log, sigmoid), according to their intention [14]. For instance, long estimated travel time monitored and input to utility function, than it means low satisfaction of driving. Consequently, system will initiate the re-routing. At this time, mobile terminal will find the new optimal path from current location to destination as follow the Sect. 84.3.3.2.

84.4 Implementation and Evaluation

84.4.1 Simulator

Proposed SeNS is running on environment, Intel Core2Quad Q9550, RAM 4 GB, JAVA JRE 1.7. And this simulator is roughly divided on three part. First, constructing road traffic networks and managing running vehicles. Second, attaching the infrastructures on road traffic networks. Finally, routing the path for vehicle.

Basically, we assume that those road segments are implemented as single lane. And every vehicles are injected and routed to one of outer intersection to heading another one of outer intersection. It can be modified to change the parameter of

injection in period and quantity. In the simulation we inject 14 vehicles (equal to number of injectable intersections), every 10 s. For making specific congestion, we generate the Gaussian distributed random variable and use for one of outer intersection to be more selected than other outer intersections. And we implement the speed of vehicles for the making a traffic information model, aspect of average speed, by considering the vehicle's time duration of transition in queue (e.g. road segment), ranged 0.7–1.0. Through this, we can achieve various average speeds on each road segments, average occupancy of roads and variance of occupancy in target area.

84.4.2 Scenario

Basically, following scenarios will be simulated under the same parameters except routing algorithm. First scenario is current time stamp-based routing without re-routing and second on is prediction-based routing with re-routing.

- Scenario 1—Current time stamp-based routing

The current time stamp-based scenario decides the estimated travel time of whole path by summation of average travel time of each road segments. Table 84.2 is the example of calculating whole estimated travel time, when a vehicle routing based on current time stamp. In edge 01–23 have 10 s of average travel time and in edge 23–34 have 7.5 s of average travel time, than estimated travel time of node 01–34 will be 17.5 s. We can get total estimated travel time by doing this sequence recursively.

- Scenario 2—Proposed routing

The prediction-based scenario decides the estimated travel time of whole path by summation of current and past average travel time of each road segments. Table 84.3 is the example of calculating total estimated travel time, when a vehicle routing based on prediction. In edge 01–23 have 10 s of average travel time and in edge 23–34 have three past average travel time. So, predicted average travel time will be 5.925 s, than estimated travel time of node 01–34 will be 15.925 s. We can get whole estimated travel time by doing this sequence recursively. After getting path, if there are some road that full occupancy, than initiate re-route sequence.

$$Occupancy = \frac{flow}{capacity} * 100\,(\%) \tag{84.2}$$

84.4.3 Evaluation

The simulation results of those two algorithm for 30 min is on Fig. 84.2. The variance, means how much spread out the vehicles, is calculated on occupancy of

Table 84.2 Example of calculating current time stamp-based travel time

Edge	01–23 (s)	23–34 (s)	34–25 (s)	25–18 (s)
Avr. Tt	10	7.5	6.24	7.7
Est. Tt	10	17.5	23.74	31.44

Table 84.3 Example of calculating prediction-based travel time

Edge	01–23 (s)	23–34 (s)	34–25 (s)	25–18 (s)
Avr. Tt	10	7.5	6.24	7.7
Avr. Tt-1	–	6.2	9	7
Avr. Tt-2	–	5	9	5
Avr. Tt-3	–	5	6	4
Est. Tt	10	15.925	23.485	29.41

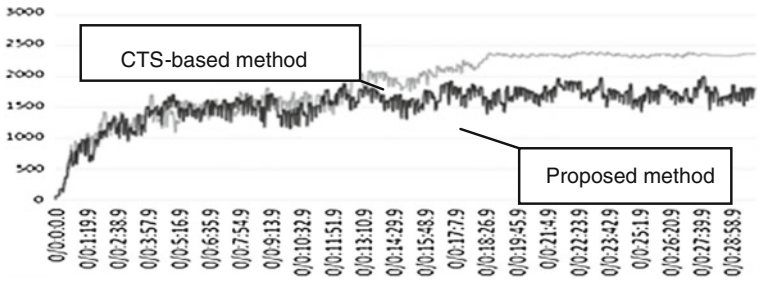


Fig. 84.2 Graph of variance for occupancy of roads in target area

roads. In case of variance, current time stamp-based routing has result of 1,875.482 in average, and our proposed method has result of 1,549.405 in average. This result show that our proposed method is more efficient in spreading out the vehicles about 17.39 %. And in case of average speed, current time stamp-based routing has result of 0.5040 in average, and our proposed method has result of 0.7691 in average. This result show that our proposed method is faster about 52.60 %. According to these results, we could prove the prediction-based routing and re-routing is a more effective way.

84.5 Conclusion

In this paper, we propose the self-adaptive navigation system (SeNS) in order to provide the optimal path to user by keep adapting the dynamic traffic information based on DSRC. This system constructs the traffic information of target area by collecting vehicular data (e.g. speed) via DSRC and generating road traffic

statistics. Through this system, we made a baseline for researching about how the self-adaptive software can affect to ITS, and we adapt to current traffic information in real-time. As the comparison of simulation results, between current time stamp-based scenario and prediction-based scenario, we could prove our proposed system is better. Our prediction-based algorithm decrease about 17.39 % of variance of occupancy of target area and increase about 50.60 % of average travel speed. As future work, we will reflect realistic feature of ITS to simulator and apply various self-adaptive methods to this system for experiment and analysis.

References

1. Wang J, Wakikawa R, Zhang L (2010) DMND: collecting data from mobiles using named data. In: Vehicular networking conference (VNC), 2010. IEEE, pp 49–56
2. Jung W, Kang C, Yoon C, Kim D, Cha H (2012) DevScope: a nonintrusive and online power analysis tool for smartphone hardware components. In: Proceedings of the 8th IEEE/ACM/IFIP international conference on hardware/software codesign and system synthesis, pp 353–362
3. Boriboonsomsin K, Barth MJ, Weihua Zhu, Vu A (2012) Eco-routing navigation system based on multisource historical and real-time traffic information. Intell Transp Syst IEEE Trans 13(4):1694–1704
4. Waze http://www.waze.com/
5. TOMTOM http://tomtom.com/
6. TrOASIS http://www.its.go.kr/opGuide/troasis.jsp
7. Tmap http://www.tmap.co.kr/tmap2/
8. Yoon J, Ji Y, Huh K, Cho DD, Park JH (2005) Optimal route searching algorithm of dividual base station for intelligent transportation systems. In: Proceedings of the conference of Korea society of automotive engineers, pp 597–603
9. Kim J, Lee S (2009) Reliable routing protocol for vehicle to infrastructure communications in VANET. J Korea Inf Commun Soc 34(8):839–845
10. ITS Maunal 2013.01. Korea Information & Comm. Contraction Association
11. Jeong JP, He T, Du DHC (2013) TMA: trajectory-based multi-anycast forwarding for efficient multicast data delivery in vehicular networks. Comput Netw 57(13):2549–2563
12. Kang Y, Kim H (2010) A study on standardization foundation through comparative analysis of visual design elements in car navigation. J Digital Des 10(4):109–118
13. Kim H et al (2011) The effects of driving performance during driving with sending text message and searching navigation: a study among 50 s taxi drivers. Korean J Sci Emotion Sensibility 14(4):571–580
14. Esfahani N, Elkhodary A, Malek S (2013) A learning-based framework for engineering feature-oriented self-adaptive software systems. Softw Eng IEEE Trans 39(11):1467–1493
15. Road service Manual (2013) Ministry of land, infrastructure and transport
16. Ye P, Chen C, Zhu F (2011) Dynamic route guidance using maximum flow theory and its mapreduce implementation. In: 14th international IEEE conference on intelligent transportation systems (ITSC), pp180–185
17. Morgan YL (2010) Notes on DSRC & WAVE standards suite: its architecture, design, and characteristics. Commun Surv Tutorials IEEE 12(4):504–518
18. Kim Seung-Cheon (2011) An evaluation of the performance of wireless network in vehicle communication environment. J Korea Inf Commun Soc 36(10):816–822

Chapter 85
A Two-Tier Routing Algorithm Based on Geographical Zone in VANET

Seung-Hwan Lee and Seung-Joon Seok

Abstract VANET Routing configures a data transmission path between vehicles sending and receiving message by using car node as relay node. Since car node is of mobility, it is difficult to maintain communication connectivity between two neighbor vehicles. A lot of studies have addressed VANET routing for improving communication qualities so far. This study proposes GZTR Protocol fit to VANET environment to tackle the difficulty. The essence of the protocol is to divides urban space into zones, selects one zone with high geographical conditions as next zone, operate relay node selection algorithm for forwarding packets to the next zone. The performance test of this system turned out to be superior to that of existing protocol.

Keywords VANET · Ad-hoc routing · Zone based routing · GZTR

85.1 Introduction

In VANET (Vehicular Ad-hoc Network), routing configures a data transmission path between vehicles sending and receiving message by using car node as relay node. Since the vehicle that plays a role of relay node is of mobility, it is hard to maintain communication connectivity between neighbor vehicles [1]. Therefore pre-computational end-to-end routing is not suitable for VANET because it has a problem of maintaining the end-to-end routing path due to the change of car location at the point when actual packet transmits. On the other hand, it is widely

S.-H. Lee · S.-J. Seok (✉)
Department of Computer Engineering, Kyungnam University, Changwon, South Korea
e-mail: sjseok@kyungnam.ac.kr

S.-H. Lee
e-mail: ivfetree@kyungnam.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_85, © Springer Science+Business Media Dordrecht 2014

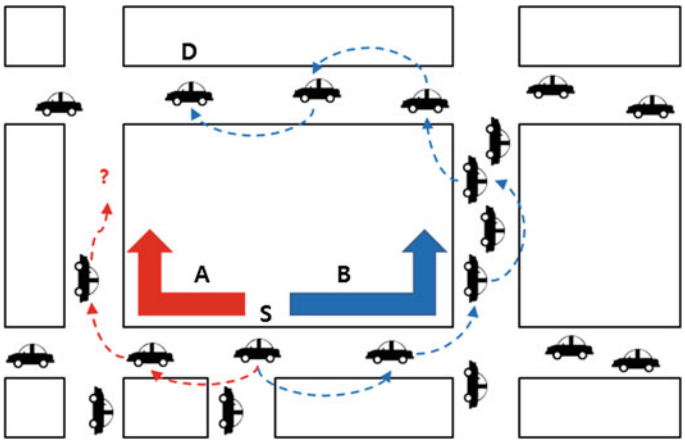


Fig. 85.1 The problem of VANET routing

known that 1-hop routing has a high level of performance because it can decide a next relay node using geographical information such as the location, direction, and speed of adjacent nodes.

The existing routing protocols that use geographical information place only distance in computation, so they select the closest node to destination node. The problem of this algorithm lies in transmission delay and path disconnection in case the density of node is not high enough and so the chance for finding an appropriate next node is low. The algorithm that depends on density has problem of path disconnection due to the lack of information about density after relay node. It is because it selects relay node according to vehicle local density information without considering the entire path to a destination. Figure 85.1 shows routing process from sending node S to destination node D. On one hand, when routing selects the closest node to the destination (Direction A) only on the basis of distance, a path is disconnected because car density is low. On the other hand, when routing selects the path of high car density (Direction B), though the number of hops increases, so it causes delay. Accordingly, this study proposes GZTR Protocol that considers both distance and density to configure an appropriate path from sending node to destination node and thus it can minimize the loss of packets.

85.2 Related Work

It is geographical zone-based routing protocol that most studies have been conducted on regarding VAMET routing. Because node location changes very fast in VAMET environment, it is advantageous to use information of neighboring nodes to set a path. This study will examine two representative ones (GPSR and GBSR) of routing protocols using geographical information.

85.2.1 GPSR Protocol

GPSR (Greedy Perimeter Stateless Routing) protocol [2] is one of the most representative routing protocols that use geographical information. It selects the physically closest node to destination node as relay node by using location information of nodes within the boundary of sending node. Because individual node retains information of adjacent nodes only in GPSR protocol, it performs better in VANET environment in which topological change often occurs.

GPSR motions are divided into greedy forwarding and local maximum processing. Greedy forwarding, which is a common data transmission mode, selects the physically closest node to destination node within the boundary of current sending node as relay node. However, it can have problem of local maximum when it fails in sensing a closer neighboring node to destination node than sending node itself. To tackle this problem, sending node selects the closest node to sending node among the nodes located within transmission range until it finds a closer node to destination node than itself. It increases transmission delay by increased number of nodes because packets are transmitted to detour direction regardless of destination.

Figure 85.2 shows the process that relay node is selected through greedy forwarding motions. When source Node S transmits packet to destination node D, Node S selects the closest node to Node D among neighboring nodes as relay node. Node C selects the closest node to Node D among the nodes within transmission range as relay node. This process repeats until packet reaches Node D.

Figure 85.3 shows the local maximum that sending node can't find a node closer to destination node than itself in the process of node selection. A problem occurs if there is no node closer to Node D than Node S when source Node S sends packet to destination node D. To solve this problem, Node S selects the closest Node F to Node S among neighboring nodes as relay node. However, in this process, there occurs transmission delay due to the increased numbers of nodes. This kind of local maximum keeps repeating until Node S finds Node B, which is closer to Node D than itself. The dotted line in Fig. 85.3 shows the process that local maximum is solved while the solid arrow line shows greedy forwarding.

85.2.2 GBSR Protocol

GBSR (Greedy Border Superiority Routing) [3] is a routing protocol proposed to solve the increased number of nodes by local maximum of GPSR. GBSR Protocol sends packet to the farthest node among neighboring nodes from the current sending node to avoid local maximum as quickly as possible.

Figure 85.4 shows GBSR Protocol motions. Local maximum occurs because there is no closer node to destination node D than sending node D within the transmission range of the source sending node (D). Node S selects Node E, which is located the farthest from the transmission range, as relay node. Node E selects a

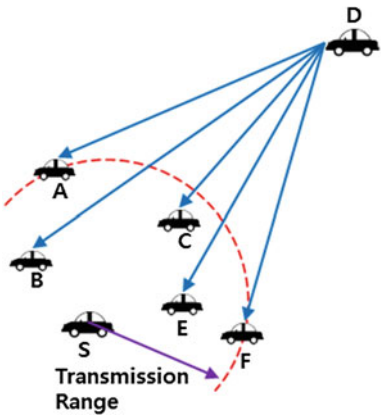


Fig. 85.2 Greedy forwarding motions

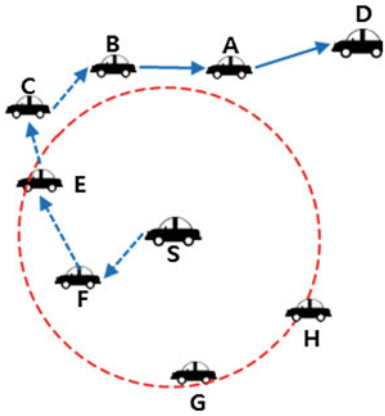


Fig. 85.3 Procedure of solving local maximum

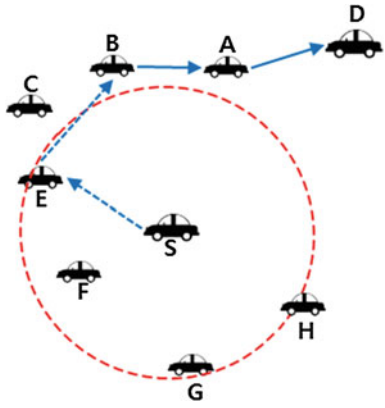


Fig. 85.4 GBSR protocol operation

node positioned the farthest from the range to escape from local maximum quickly. GBSR Protocol can escape from local maximum only with two relay nodes while GPSR Protocol has four relay nodes to avoid it as seen in Fig. 85.3.

When packet is sent in greedy forwarding mode, both GPSR and GBSR have similar performance rate but GBSR has a better result to handle local maximum than GPSR. However, swift data transmission and reliability must be secured in VANET environment. Therefore decreasing the number of relay nodes can't be an ultimate solution to local maximum.

85.3 GZTR Protocol

In VANET environment that consists of mobile terminals, there also exists topography in which a car is limited to mobility due to buildings or parks as well as jammed roads or crossroads with traffic lights. Therefore, routing algorithm is required that sets up a safe path in consideration of such virtual environment. This study proposes GZTR (Geographical Zone Based Two-tier Routing) algorithm that divides local geographic environment into zones, selects a zone that is the most fit to routing at first, and then selects an appropriate relay node to send packet between zones. Information of location such as roads, buildings, and parks is used to divide space in a manner that communications between cars go smooth within a local zone.

GZTR algorithm is classified into two algorithms: one is to select next zone and the other is select relay node. First a node that has packet should decide whether the destination node is in the zone where it is or not. If the zone includes the destination node, the node runs relay selection algorithm to move to the destination mode, otherwise next zone selection algorithm is first run to select a zone with superior environment for communication among neighboring zones. After carrying out next zone selection, it performs relay node selection algorithm to decide a next node toward next zone.

Figure 85.5 shows two selection algorithms: next zone and next relay node. Node S that has packet selects one zone of optimal routing condition among neighboring A, B, C, and D zone as next zone. After the next zone has been selected, Node S selects the closest node to the boundary surface of the next zone, among the nodes in the zone, as relay node to send packet to it. Let's look into the two algorithms more in detail.

85.3.1 Next Zone Selection Algorithm

Next zone plays a role of intermediate area to destination zone. The zone with the optimal geographic condition for communication among neighbor zones is selected as a next zone in consideration of the density of vehicles within each zone

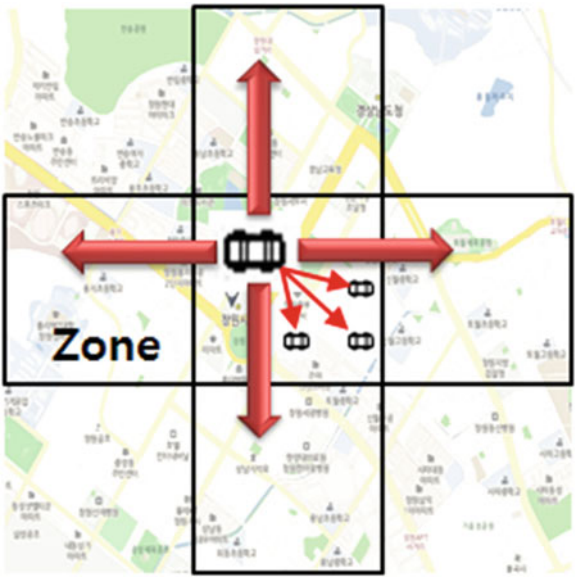


Fig. 85.5 Next zone selection and relay node selection

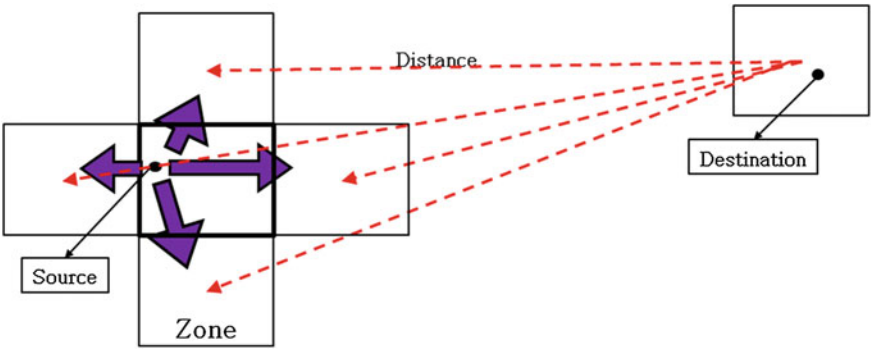


Fig. 85.6 Next zone selection algorithm

and the distance to the destination zone. When a node receives a packet, it checks the next zone information in the header of the packet. If the current location of the node is the next zone, next zone selection algorithm is run because the current node is the first node in the zone.

Figure 85.6 shows the process of selecting a next zone. First, the algorithm calculates the physical distance from the central point of the zone, where destination node is located, to the central point of the neighboring zones near the source node (current node). If car density in the neighboring zones is similar, closer zone is advantageous in reducing the number of the entire hops. Since there may occur

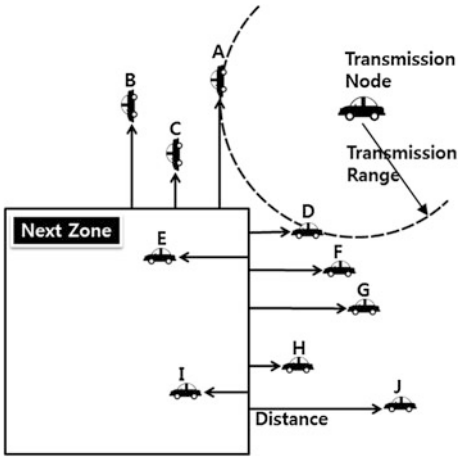


Fig. 85.7 In case next zone is beyond transmission range

path disconnection in the zone of low car density if next zone is determined only in consideration of physical distance, both density and distance are in calculation. In this study, the number of cars in a zone is defined as zone density and then the zone density of the current zone where source sending node and neighboring zones are computed. The zones of more than a certain level of density are selected as candidate zones. Among them, the closest one to the center point of destination zone is selected as next zone.

85.3.2 Relay Node Selection Algorithm

After next zone is decided, relay node selection algorithm is carried out to send packet to next zone. Since both density and distance were computed to select next zone, only distance is calculated for relay node selection. The process of relay node selection is that sending node uses location information system to decide first if the boundary surface of next zone is within transmission range. In case that sending node is positioned beyond the range, the algorithm calculates the physical distance of neighboring nodes within the range from the surface and selects the closest node to the boundary. It is because the closest node to the boundary is highly probable to send packet to next zone the fastest.

Figure 85.7 shows the case that next zone is beyond transmission range. From the point of the sending node, Node D is the closest to the boundary surface of next zone. Therefore, Node D is selected as relay node. When sending node is positioned in the boundary surface of next zone and within the range, the algorithm decides if next zone is destination zone and a node within the zone. In case that next zone is destination zone, the closest node, among neighboring nodes within transmission range, to destination node. At this time, only distance is included in calculation.

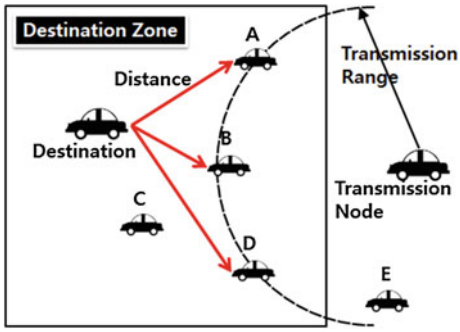


Fig. 85.8 In case next zone is destination zone

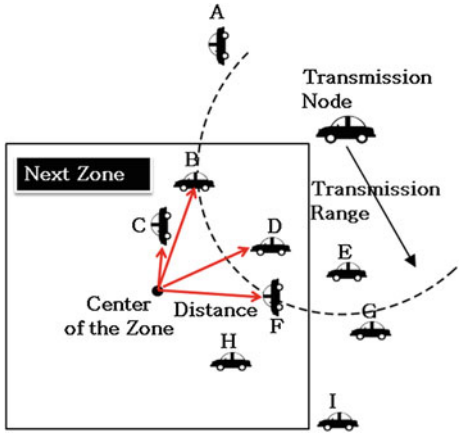


Fig. 85.9 In case next zone is within transmission range

Figure 85.8 shows that Node B is selected because next zone is destination zone. Even if it is not destination zone, the closest node to the central point of next zone is selected as relay node. The reason why the closest node to the center of next zone is selected as relay node is that doesn't have any information of the zone after next zone, so there is possibility that relay node can be distant from the boundary surface of next zone if distance from a certain fixed boundary surface is used to decide relay node.

Figure 85.9 shows the process that the closest node to the central point of next zone is selected as relay node. Then this node becomes an axis and relay node selection of algorithm is processed again.

85.4 Performance Evaluations

This paper evaluates the performance of GZTR algorithm as compared with the existing routing algorithm that uses geographical information through computer

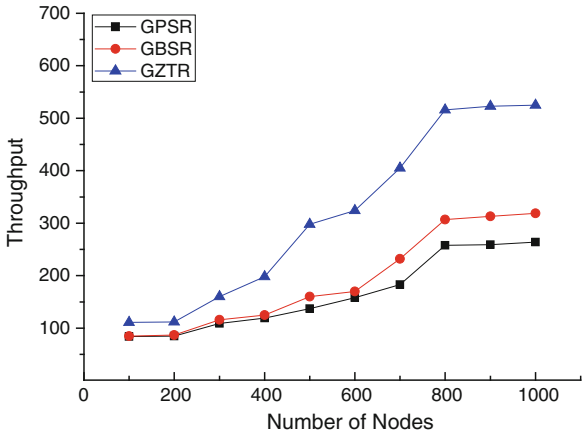


Fig. 85.10 Throughput rate by car density

simulation. C/C++ language was employed to make a simulator for performance evaluation. Throughput rate and packet loss rate by vehicle density are compared. For simulation environments, Wi-Fi IEEE 802.11b are considered as link communication technology and ns2 VANET network model is used. An assumption was established in this experiment that information of location of the entire zones and nodes was obtainable at all vehicles.

Figure 85.10 depicts throughput rate by car density when packet is transmitted from sending node to destination mode. As there are more cars, throughput rate increases. It can be explained in that the number of packets equally increases as the number of nodes increases, so the number of packets to be handled (throughput rate) also increases. And when the number of nodes increases up to a certain level, it is more probable that the closest car to destination mode is selected. Therefore, it was confirmed that the degree of change is not significant.

Figure 85.11 depicts the packet loss rate by the number of nodes when packet is transmitted. When the number of packets is below a certain level, there is a high probability of non-communication, so the loss rate turns out to be high, as confirmed in the graph. In general, as the number of nodes increases, car density increases. Next, it can be observed that the loss rate decreases as the number of nodes that can be selected as relay node increase. When the number of nodes increases up to a certain level, density increases, leading to constant rate of packet loss rate. Therefore, it was found that the technology suggested in this study has lower rate of packet loss than the existing ones.

Table 85.1 shows the hop counts of end-to-end path from source and destination nodes according to the variable density of vehicles. In the case of low density, the hop counts of end-to-end path is increased because a relay node may not be in the direction of its destination node. However, the more number of vehicle (vehicle density), the less hop counts between source and destination nodes. It is reason why nodes, which are in the direction of its destination, possibly selected as relay nodes.

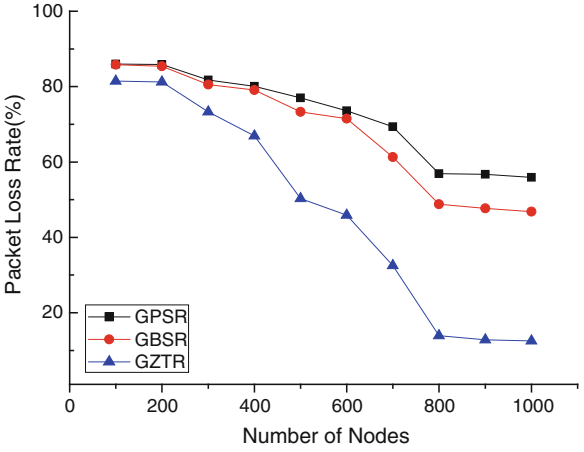


Fig. 85.11 Packet loss rate by car density

Table 85.1 Average hop counts between source and destination according to vehicle density

Vehicle density	GPSR	GBSR	GZTR
200	80	76	58
400	60	57	42
600	45	42	30
800	28	26	16
1,000	27	22	12

85.5 Conclusions

In this paper, GZTR Algorithm, which is suitable for to urban conditions in VANET environments, was proposed. This study divided geographic local area into zones and selected a next zone by using the distance to destination and vehicle density within each zone. Relay node was selected after deciding if the boundary surface of zone is within or beyond transmission range. When it is beyond the range, the closest node to the boundary surface of next zone was selected as a next relay node. When it is within the range, it was first decided whether next zone is destination zone or not. When next zone is destination, the closest node to destination node was selected as a next relay node while the closest node to the center of the next zone was selected as a next relay node when next zone is not destination. For future research topic, this study propose that it is necessary to study on the solution to local maximum problem which can occur when there is no clear node to destination node than sending node.

References

1. Jerbi M, Senouci SM (2009) Towards efficient geographic routing in urban vehicular networks. IEEE Trans Veh Technol 58:5048–5059
2. Karp B, Kung HT (2000) GPSR: greedy perimeter stateless routing for wireless network. In: Proceedings of ACM/IEEE MOBICOM 2000, pp 243–254
3. An D, Cho GH (2009) KORE. http://www.dbpia.co.kr/Journal/ArticleDetail/1455819
4. Ali S, Bilal SM (2009) An intelligent routing protocol for VANETs in city environments. In: Computer, control and communication, IC4 2009, pp 1–5
5. Li H, Guo A, Li G (2010) Geographic and traffic load based routing strategy for VANET in urban traffic environment. In: 3rd international conference on wireless, mobile and multimedia networks (ICWMNN 2010), IET, pp 6–9
6. Hsieh M, Wei C (1998) Channel estimation for OFDM systems based on comb-type pilot arrangement in frequency selective fading channels. IEEE Trans Consum Electron 44(1):217–225

Chapter 86
Analysis of Vehicle Surroundings Monitoring System Based on Vehicle-to-Vehicle Communication

Hong Cho and Byeong-Woo Kim

Abstract Using conventional technology an intelligent vehicle system can monitor the area within 200 m of a vehicle using sensors attached to the vehicle. In this study, a vehicle-to-vehicle (V2V) communication method is introduced to overcome the limitation of driving area detection. By applying V2V communication to an autonomous longitudinal control system, the possibility to replace existing longitudinal control systems based on sensors is suggested. In addition, it can be verified by applying V2V communication to an actual autonomous emergency braking system.

Keywords Vehicle-to-vehicle (V2V) · Vehicle-to-infrastructure (V2I) · Intelligent transportation system (ITS) · Cooperative safety system · PreScan · CarSim

86.1 Introduction

Vehicle safety is one of the most important issues in the car industry. This is reflected in recently reinforced legal safety measures and increasing consumer interest in safety features. In line with this trend, the Advanced Driver Assistance System (ADAS) has been attracting attention due to its focus on accident prevention by detecting risk factors in the driving area with various sensors such as vehicle-mounted cameras and radar/lidar [1, 2]. However, such sensor-based ADAS has a limited capability of monitoring only vehicles within the detection

H. Cho (✉) · B.-W. Kim
Department of Electrical Engineering, University of Ulsan, Ulsan, South Korea
e-mail: hong9life@ulsan.ac.kr

B.-W. Kim
e-mail: bywokim@ulsan.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_86, © Springer Science+Business Media Dordrecht 2014

range of the sensor, and its driving area risk detection becomes inaccurate or impossible for obstacles in a blind spot or during adverse weather conditions.

Intensive efforts are currently underway to introduce the intelligent transportation system (ITS) using rapidly evolving radio communication and state-of-the-art information technologies—in particular, the cooperative safety system which integrates vehicle safety communication, such as vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication into the existing ITS concepts [3, 4].

The advanced emergency braking system (AEBS) has been established as an international standard, and a variety of studies have been conducted on it [5]. However, these studies have relied on the performance of the vehicle-mounted sensors, and thus have limited detection range. The current trend is therefore to shift from the ADAS based on vehicle-mounted sensors to a V2V communication system.

By using V2V communication technology as the means for monitoring a vehicle's surroundings, the limitations associated with blind spots or adverse weather conditions in recognizing risks from vehicles in the driving area can be overcome. Therefore, this paper proposes a V2V-based driving area risk vehicle monitoring system.

86.2 Vehicle Surroundings Monitoring System

86.2.1 Analysis Model for the Vehicle Surroundings Monitoring System

In order to analyze the V2V-based driving area vehicle risk monitoring system proposed in this paper, it is necessary to construct an analysis model equipped with devices enabling V2V communication, such as a camera sensor, positioning system, and inter-vehicle wireless communications module as illustrated in Fig. 86.1. Thus, we constructed a model to analyze the V2V-based vehicle surroundings monitoring system using PreScan, CarSim, and MATLAB/Simulink.

Our analysis model was constructed with the following elements: (1) a detailed model of a vehicle-mounted camera sensor, positioning system, and wireless communication module, and driving environment using PreScan, (2) a vehicle dynamic analysis using CarSim, (3) a vehicle surroundings monitoring system coordinating PreScan and CarSim with MATLAB/Simulink.

A vehicle surroundings monitoring system comprises a camera sensor, a positioning system, a vehicle-mounted sensor, and a wireless communication module for V2V communication. The camera sensor recognizes lanes and provides information on lane conditions, and the positioning system provides information on vehicle position and heading. Additionally, various in-vehicle sensors provide data on velocity, acceleration, and yaw angular velocity via an in-vehicle communication system. Moreover, the wireless communication module enables inter-vehicle

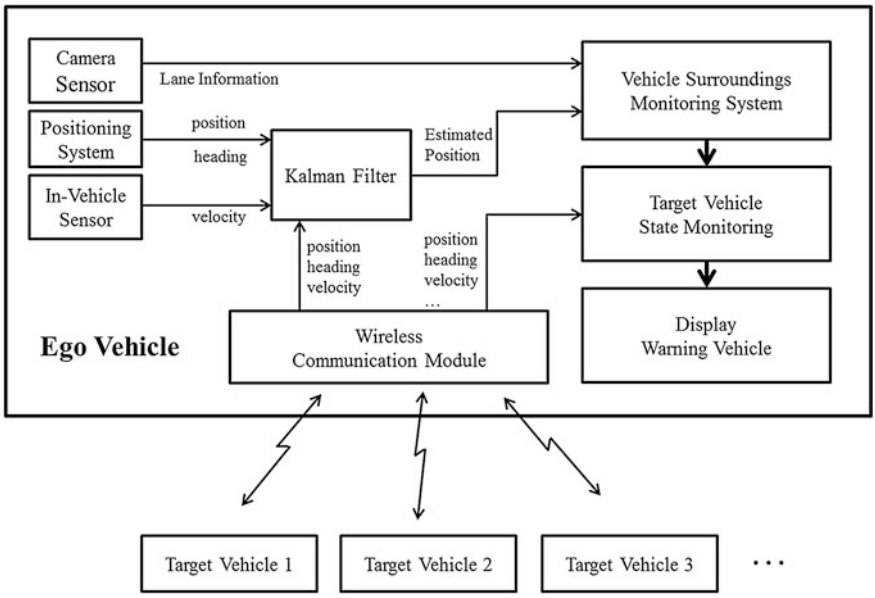


Fig. 86.1 Block diagram of the monitoring system

communication within the driving area and thus information exchange. We applied the basic safety message (BSM) defined by the SAE J2735 to the inter-vehicle message frame. A Kalman filter was used to adjust for positioning system errors and tracing the ego vehicle trajectories [6]. We expressed the Global Positioning System (GPS) data received from the positioning system for the ego and surrounding vehicles in Cartesian coordinates that are defined here as the tangential plane in GPS coordinates of the infrastructure system within 1 km, with the positive x and y directions indicating east and north, respectively [7]. Additionally, given that the simulation does not provide GPS error data, we included a noise generator matching the error specification of a commercialized differential global positioning system (DGPS).

86.2.2 Mechanism of the Vehicle Surroundings Monitoring System

The V2V-based vehicle surroundings monitoring system proposed in this paper is a system for generating a coordinate system around the ego vehicle and calculating its distance from surrounding vehicles (Fig. 86.2). To analyze the system, we defined a new Cartesian coordinate system (CS_{ego}) with the current ego vehicle position as the origin (x_{ego}, y_{ego}). The longitudinal and lateral directions of the ego

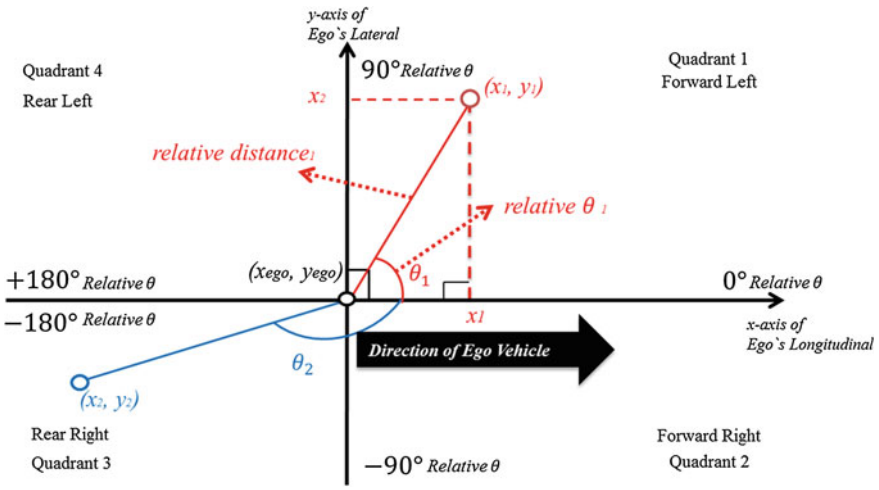


Fig. 86.2 Principles of the monitoring system

vehicle's driving direction are along the x and y axes, respectively, in the CS_{ego} coordinate system defined here. The positions of surrounding vehicles are expressed as relative coordinates (x_n, y_n) $(n = vehicle\ id)$ in the CS_{ego} system estimated by comparing the positions of ego and surrounding vehicles based on V2V communication data. A relative angle θ_n is calculated by comparing the driving direction of the detected vehicle at azimuth φ, to the driving direction of the ego vehicle. As shown in Fig. 86.2, the vehicle surroundings monitoring system can recognize the positions of surrounding vehicles based on the changes in the relative angle θ_n along the quadrants. The azimuth φ changes according to the vehicle's heading. The azimuth φ in the eastward (longitudinal) direction is defined to be 0° (Fig. 86.3).

In the CS_{ego} coordinate system, changes in a vehicle's Azimuth φ entails an axial rotation, as shown in Fig. 86.3. In consideration of this rotation, the ego vehicle's longitudinal direction was designated always to run parallel to the x axis using a translation–rotation matrix. We established a warning zone on the basis of the relative angles obtained in the process described above (Fig. 86.4). θ_r and θ_f are error ranges of θ_n, assuming that all vehicles are driving in the center of their respective lanes.

Important parameters for determining a warning zone are relative angles, distances, and headings of the surrounding vehicles. The heading is an important parameter for determining a vehicle's driving direction. We calculated the relative angles between the headings of ego and target vehicles to determine whether the target vehicle was driving in the same, intersectional, or opposite direction. Further, the relative distance could be obtained using the CS_{ego} coordinates of the ego vehicle-centric coordinate system. Because the relative distance thus obtained does not reflect the vehicle size, we assumed it to be round and calculated the relative

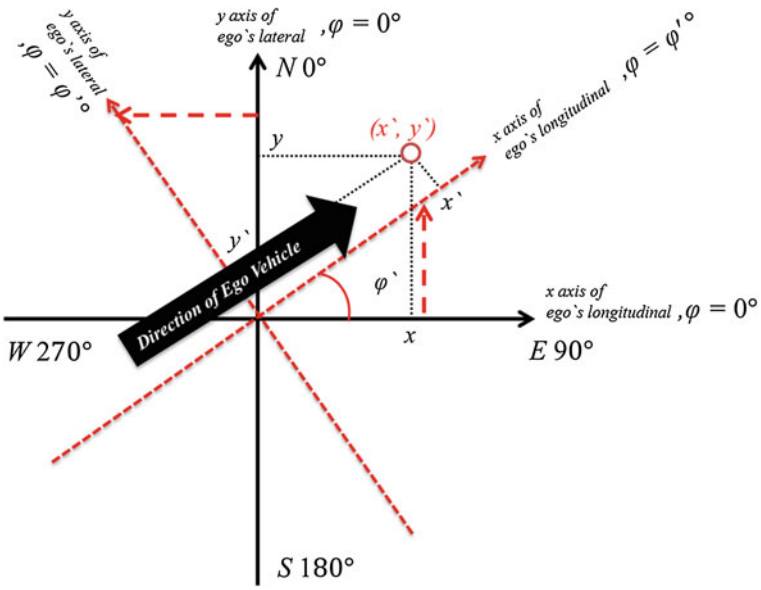


Fig. 86.3 CS_{ego} Transformation by the azimuth change

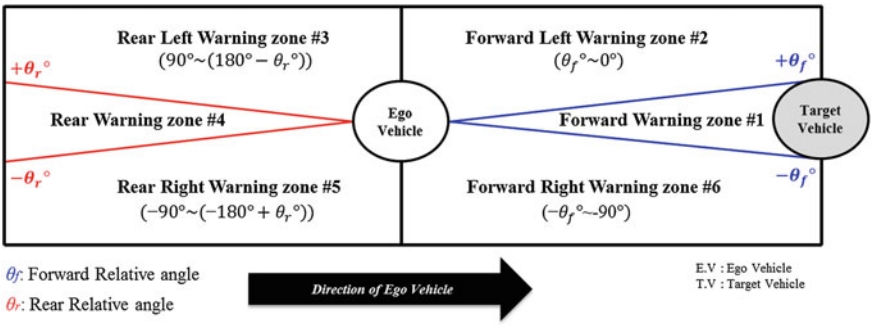


Fig. 86.4 Warning zone corresponding to the relative angle

distance [8]. The relative distance calculated was used for determining the collision risk by calculating the time to collision (TTC) using Eq. (86.1) based on the relative distances and speeds of the ego and surrounding vehicles.

$$TTC(s) = \frac{Relative\ distance}{Relative\ speed}. \tag{86.1}$$

Table 86.1 Braking force of AEBS

Braking force (g)	TTC (s)
0.3	≤2.0
0.6	≤1.6
1.0	≤0.7

86.3 Verification of the Vehicle Surroundings Monitoring System

86.3.1 Simulation Scenario

For a verification test of the V2V-based vehicle surroundings monitoring system proposed in this paper, we applied the system used for the AEBS verification test. AEBS is an active-type safety system designed to prevent or avoid accidents by implementing automatic breaking on the basis of TTC between the ego and forward vehicles. Therefore, we conducted the verification test by substituting the TTC inputs of AEBS with those of the proposed system.

The instants for applying AEBS braking force depending on the TTC change were determined as outlined in Table 86.1 [9]. The detailed simulation scenario for the verification test of a vehicle monitoring system is as follows: (1) The ego vehicle is driving first at 50 km/h, then accelerates to 60 km/h, and continues at that speed; (2) Vehicle 1 is driving at a constant speed of 60 km/h and changes lanes after recognizing trouble with a forward vehicle; (3) Vehicle 2 is driving first at 50 km/h and then gradually slows until it stops due to vehicle trouble; (4) Vehicle 3 is driving in the opposite lane at a constant speed of 70 km/h. Figure 86.5 illustrates this simulation scenario.

86.3.2 Simulation Results

Figure 86.6 shows the results of the vehicle surroundings monitoring system simulation with the ego vehicle as the reference. In Fig. 86.6d, the relative speed of the trouble vehicle (Vehicle 2, red line) is shown to rise abruptly starting from the 9th second. This is because vehicle trouble caused it to decrease speed until it completely stopped. At the moment Vehicle 2 stopped, the longitudinal direction risk index TTC_x was reduced as the ego vehicle approached Vehicle 2. As shown in Fig. 86.6b, a braking force was generated according to the AEBS brake application instruction time based on TTC_x. This is an example of the TTC_x-based normal operation of AEBS. Additionally, it could be demonstrated that a safe distance was maintained by AEBS, and collision did not occur in this simulation environment, as shown in Fig. 86.7.

The reason that Vehicle 3 in Fig. 86.6a does not appear in the longitudinal collision risk index TTC_x is because the system recognized it as a vehicle driving in the opposite lane on the basis of its heading information.

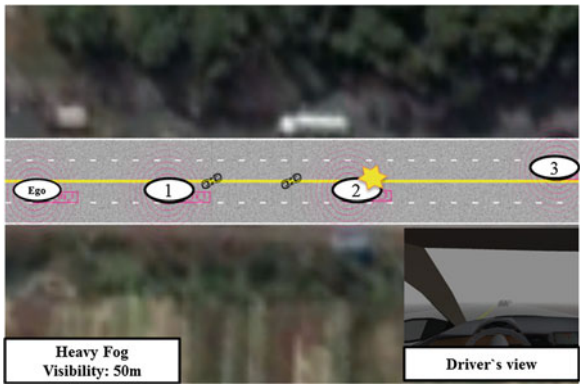


Fig. 86.5 Simulation scenario

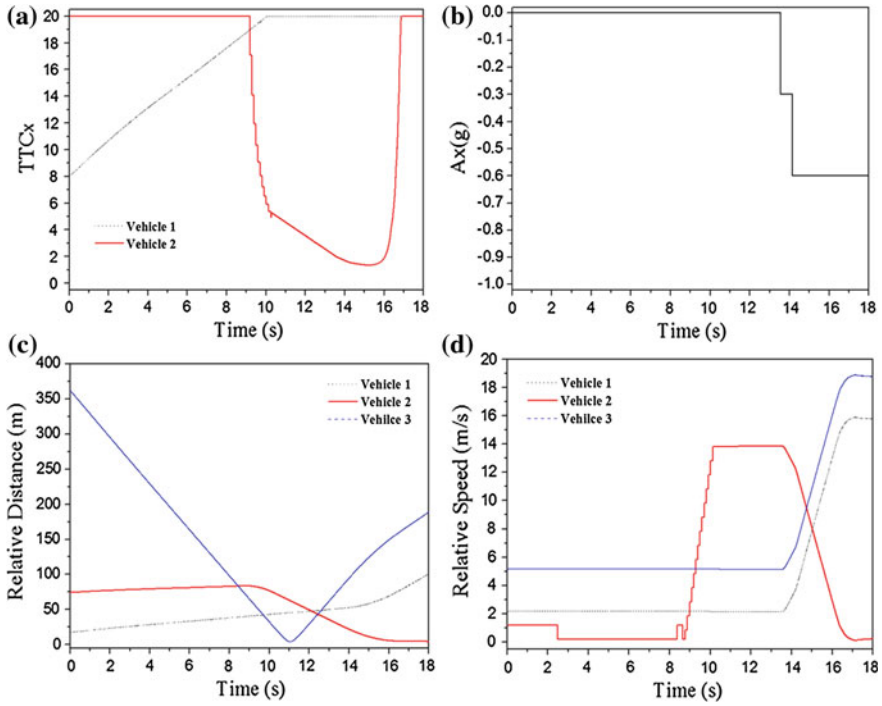


Fig. 86.6 Simulation results for verification. **a** TTCx. **b** AEBS command. **c** Relative distance. **d** Relative speed

Figure 86.8 illustrates the warning zones depending on driving time. Vehicle 1 was forward of the ego vehicle, and its warning zone changed from 1 to 6 because it changed lanes after recognizing the trouble vehicle. The change of the warning

Fig. 86.7 Safe distance secured by AEBS

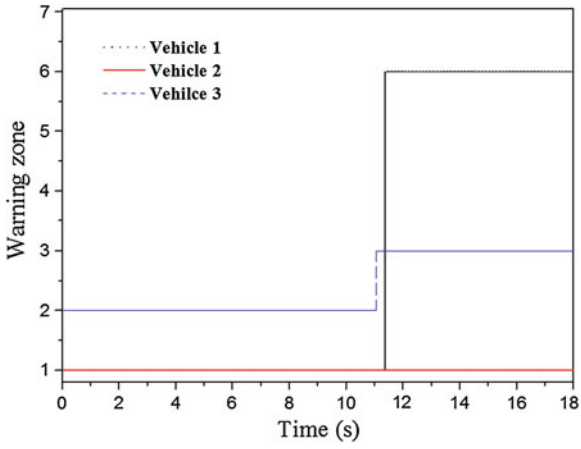


Fig. 86.8 Warning zone changes in the simulation

zone for Vehicle 3, which was driving in the opposite lane, from zone 2 to zone 3 at about the 11th second is explained by its changed position from the forward to the rear quadrant. This can be verified by the relative distance of Vehicle 3 approximating to 0 and increasing again, as shown in Fig. 86.6c.

86.4 Conclusions

This paper investigated an analysis method of a V2V-based vehicle surroundings monitoring system. To this end, it proposed an analysis model for AEBS using PreScan, CarSim, and MATLAB/Simulink.

Additionally, it quantified collision risk by introducing the collision risk index TTC extracted from the relative speed calculated on the basis of relative distance obtained using an ego-vehicle-centric coordinate system and inter-vehicle communication. In this study, we also proposed a method for quantitatively analyzing the collision risk by applying the collision risk index TTC to a real AEBS device. The results of applying the proposed TTC to the AEBS device verified that collision risk can be quantitatively analyzed by real-time detection of an obstacle at a greater distance ahead of the forward vehicle.

In this study, we used the proposed vehicle surroundings monitoring system to quantify the V2V-based longitudinal collision risk detection index while driving on a straight road. Furthermore, we verified that the proposed system recognized a risk vehicle in the driving area without a problem even in extreme weather conditions that pose problems to existing ADAS based on a vehicle-mounted sensor.

In our future research, we intend to focus on system analysis under various road situations using the proposed V2V-based vehicle surroundings monitoring system and a new risk index extraction method by modeling various driving situations.

Acknowledgments This Research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the Convergence Information Technology Research Center (CITRC) support program (NIPA-2013-H0401-13-1008) supervised by the NIPA (National IO Industry Promotion Agency).

References

1. Kato T, Nimomya Y, Masaki I (2002) An obstacle detection method by fusion of radar and motion stereo. IEEE Trans Intell Transp Syst 3(3):182–188
2. Atev S, Arumugam H, Masoud O, Janardan R, Papanikolopoulos NP (2005) A vision-based approach to collision prediction at traffic intersections. IEEE Trans Intell Transp Syst 6(4):416–423
3. Caveney D, Dunbar WB (2012) Cooperative driving: beyond V2V as an ADAS sensor. In: IEEE intelligent vehicles symposium (IV), pp 529–534, June 2012
4. Jung D-H (2009) Intelligent vehicle development and application projects in USA. Auto J 31(3):70–74
5. Kopetz H, Poledna S (2013) Autonomous emergency braking: a system-of-systems perspective. In: IEEE/IFIP conference, dependable systems and networks workshop, pp 1–7, June 2013
6. Ammoun S, Nashashibi F, Laurgeau C (2006) Real-time crash avoidance system on crossroads based on 802.11 devices and GPS receivers. In: IEEE intelligent transporation systems, pp 1023–1028, Sept 2006
7. PreScan R6.6.0 Help Manual, pp 832, May 2013
8. Wang Y (2011) Vehicle collision warning system and collision detection algorithm based on vehicle infrastructure integration. In: Advanced Forum on Transportation of China (AFTC 2011), 7th, pp 216–220, Oct 2011
9. Lee J, Jo S, Kwon J, Hong T, Park K (2013) Development of V2I-based intersection collision avoidance system. In: Proceedings of conference for Korea Institute of ITS, pp 90–96, May 2013

Chapter 87
The Energy-Efficient Probabilistic Routing in MANETs

Sang-Hyun Park and Jung-Ryun Lee

Abstract In this paper, we propose Energy-efficient Probabilistic Routing (EEPR) that control routing request packets stochastically to increase network lifetime and decrease packet loss in flooding algorithm. We adapt energy efficient probabilistic control using residual energy of each node and ETX metric for original AODV protocol. In simulations, we verify that the proposed algorithm has longer network lifetime and consume residual energy of each node more evenly than the original AODV protocol.

Keywords Energy-efficiency · Energy-aware · ETX · Residual energy · Network lifetime · MANET

87.1 Introduction

In MANET (Mobile Ad-Hoc Network) environment, most of nodes are mobile devices and use their battery as an energy source. In distributed network environment likes MANET, all nodes organize network themselves and communicate each other [1]. Therefore node's battery is a key factor for mobile nodes in MANET. Nodes which are worn out its battery are not able to function in the network. This problem causes a network partitioning or something other problems that decrease network performance.

Basic routing protocols like AODV (Ad hoc On-demand Distance Vector) and DSR (Dynamic Source Routing) are designed to find a shortest path and use this

S.-H. Park · J.-R. Lee (✉)
School of the Electrical Engineering, Chung-Ang University, Anseong, Republic of Korea
e-mail: jrlee@cau.ac.kr

S.-H. Park
e-mail: shpark3@cau.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_87, © Springer Science+Business Media Dordrecht 2014

path for the packet delivery [2, 3]. Nodes that are on the shortest path are rapidly worn out their battery and not able to communicate. This problem leads to decreasing network lifetime and network performance. Reactive routing protocols like AODV and DSR find the routing path without information of all network nodes. Also, reactive routing protocols use the flooding algorithm that forwards request packets to its 1-hop neighborhood nodes to find the routing path. But, in density network, when nodes forward request packets to its 1-hop neighborhood nodes using the flooding algorithm, packet loss or network congestion will be occurred because of excessive request packets in the network. Therefore limiting excessive request packets is required.

87.1.1 Related Work

Wang proposed energy efficient routing protocol using node's residual battery based on the DSR [4]. Wang proposed a algorithm that controls the probability of forwarding request packets according to a node's residual energy. If node *i*'s residual energy is larger than predefined residual energy threshold, forward the request packet. If node *i*'s residual energy is smaller than predefined residual energy threshold, compute forwarding probability using the proposed equation, and make a decision to forward or discard it. Because nodes having more residual energy are selected in routing process, using this algorithm, the more energy efficient routing path is selected. But Wang's proposed algorithm has disadvantages. Using only the proposed algorithm, when all nodes' residual energy are worn out, all nodes in the network do not forward the request packet, and routing request process will be failed. Also, using only residual energy for controlling request packets, throughput will drop down because of ignoring link quality. Also basic routing protocols like AODV and DSR, finding shortest path for routing path, are not considering the throughput of data packets. Because of not considering link quality, algorithms finding shortest path for routing path cannot guarantee the performance about throughput. For complementing basic routing protocols, it's necessary that a link quality metric is contained in routing path setup process.

Patil explained that the reactive routing protocol AODV is found to be the most energy efficient and proposed energy efficient routing protocol based on AODV considering Transmission power and remaining energy capacity of mobile nodes [5]. The author classified energy efficient routing protocols in MANETs as four sets. The first set is protocols that use the amount of transmission power of mobile nodes as a routing cost metric and save energy consumption of nodes. The second set is protocols that use remaining energy capacity of mobile nodes as a routing cost metric and target to gain fairness of energy consumption. The third set is protocols that use lifetime of mobiles nodes as a routing cost metric and target to gain fairness of energy consumption likes the third set. And the fourth set is protocols that use at least two of above cost metrics as routing cost metrics. The last set is hard to find suitable tradeoff between cost metrics. The algorithm

chooses the best path that nodes' transmission power is smaller and remaining energy capacity is bigger. The algorithm is well organized in terms of cost metrics that consider transmission power and remaining energy capacity at the same time. But Patil's proposed algorithm has several disadvantages. First, effort to reduce routing request message is not in the proposed algorithm. Second, considering only energy efficient metrics in routing process, throughput will drop down because of ignoring link quality between nodes.

Nand explained that broadcast by flooding algorithm waste network resources like channel bandwidth and node energy [6]. Thus the author proposed Probability Based Improved Broadcasting algorithm that control RREQ message based on probability using node energy. A new field, 'Node Remaining Energy Strength' is added to routing request packet frame. When a node broadcast RREQ message to other nodes, the node store its node energy in the new field. When, a node receive first RREQ message from other node, the node starts relay delay timer and receives RREQ messages from other nodes. If the delay timer ends, the node calculates a ratio, $E(S)$, between the maximum value of node energies in RREQ messages and the minimum value of node energies in RREQ messages. Then, the broadcasting probability is defined as a value dividing $E(S)$ into the number of RREQ messages. The node broadcasts RREQ message according to the broadcasting probability. The algorithm is well organized in terms of reducing RREQ messages using a broadcasting probability that consider node energies. But Nand's proposed algorithm has several disadvantages. First, node energies are only used in reducing RREQ message and not considered in routing path selection algorithm. Second, the algorithm is not convincing, because choosing broadcasting or not according to neighbor's node energies is hollow.

De Couto proposed the ETX (Expected Transmission Count) metric. ETX is a link quality metric and used in routing request process [7]. ETX metric means the number of transmission that source node have to forward probe packets to destination node to success the forward process in the link. Exchanging probe packets each other, each node predicts the quality of link between them and use the ETX metric to find routing path. De Couto adapted the ETX to DSDV (Destination-Sequenced Distance-Vector) and DSR routing protocol. From simulations, De Couto verified that the ETX-DSDV and ETX-DSR surpass previous shortest path DSDV and DSR in throughput and hop-count.

In this paper, we propose the Energy-efficient Probabilistic Routing (EEPR) based on reactive routing protocol, AODV. EEPR use ETX metric and residual energy of each node as routing metrics. By using ETX metric, EEPR compose the routing path with links in good link quality. By using residual energy of each node, EEPR choose nodes in good energy condition as members of the routing path and all nodes in the network use their residual energy more evenly. Also, EEPR controls the flooding of request packets in routing process contrary to blind flooding in original AODV. As a result, EEPR finds the energy-efficient routing path and reduces overhead in routing process.

87.2 Proposed Algorithm

In this paper, the EEPR algorithm controls request packet forwarding process for reducing packet loss and network congestion in flooding process of original AODV. The main routing path setup process of EEPR algorithm is same as one of original AODV. A source node which has data packets to transmit forwards route request packets to its 1-hop neighbor nodes. In original AODV routing flooding process, forwarder nodes which receive request packets forward request packets to all its 1-hop neighbor nodes. But, in the EEPR, forwarder nodes do not forward request packets all the time. Forwarder nodes calculate the forwarding probability via the proposed forwarding probability equation and decide stochastically whether to forward or discard it. In this point, the residual energy of each node and ETX metric are used.

87.2.1 Metrics in EEPR Algorithm

Residual energy. In this paper, we use the residual energy of each node as a metric to consider the energy efficient factor of the network. The residual energy of nodes is one of the most energy efficient related factors among network energy factors. We define the residual energy of each node and maximum residual energy of each node as below.

E_i The residual energy of node i
E_{max} The maximum residual energy of node i

ETX metric. We use the ETX metric as a metric to consider the link quality between each node. ETX metric is known as the most reliable routing metric among metrics about link quality [8]. We define the ETX metric of each link and maximum ETX value as below. ETXmax is the ETX value of a link which has maximum transmission range.

$ETX_{i\text{-}1,i}$ The ETX value between node $i-1$ and node i
ETX_{max} The maximum ETX value of links in the network

Probe packets are used to calculate ETX metric. Probe packet has very small packet size. And each node periodically broadcasts probe packets to its 1-hop neighbor nodes. ETX metric is defined as below.

$$ETX = \frac{1}{pq} \tag{87.1}$$

In Eq. (87.1), p means forward packet delivery ratio. q means reverse packet delivery ratio. We assume that each node remember the number of probe packet from the other node within w seconds. When each node periodically broadcast probe packets in τ cycles, the probe packet delivery ratio of one node at time t is as below.

$$r(t) = \frac{count(t-w,t)}{w/\tau} \tag{87.2}$$

The denominator of Eq. (87.2) is value that w is divided by period τ seconds. It means the number of probe packets that one node has to be received in w seconds. The numerator of Eq. (87.2) means the number of probe packets that one node is received from $(t - w)$ seconds to t seconds. Therefore, from the Eq. (87.2), each node can calculate delivery ratio by counting the number of probe packets. Each node periodically calculates ETX metric between itself and neighbor nodes and save it.

Like above process, Probe packets are used to calculate ETX metric. In this paper, we use BER (Bit Error Rate) to calculate ETX metric. We setup path loss model, and calculate RSS (Receive Signal Strength), the signal strength that received node senses, using distance between nodes. RSS is calculated as below.

$$RSS_{dB}(x) = P^{tx}_{dBm} - P^{loss}_{dB}(x) \tag{87.3}$$

$RSS_{dB}(x)$ Received Signal Strength at a node which be away x km from the source node (dB scale)
P^{tx}_{dBm} Transmission power from the source node (dBm scale)
$P^{loss}_{dB}(x)$ Path loss at x km from the source node (dB scale)

And then, SNR can be calculated by using RSS as below.

$$SNR(x) = \frac{2 \times RSS_W(x)}{P^{noise}_W} \tag{87.4}$$

$SNR(x)$ SNR value at a node which be away x km from the source node
$RSS_W(x)$ RSS at a node which be away x km from the source node ($Watt$ scale)
P^{noise}_W Noise power ($Watt$ scale)

Using the above SNR value, we calculate BER when packets pass through ITU Ped A channel [9]. We calculate PER (Packet Error Rate) using BER and the size of a probe packet.

$$E_{pp} = 1 - (1 - E_b)^{L_{pp}} \tag{87.5}$$

E_{pp} PER of a probe packet
E_b BER
L_{pp} The size of a probe packet

We calculate ETX of each link by counting the number of probe packets that a node is received when the total number of probe packets is 10.

87.2.2 The EEPR Algorithm

To define forwarding probability, we define each factor as below. And the forwarding probability p of node i is defined as below.

$$p = \left[p_{min} + E_i A \left[1 + \frac{\left(ETX_{i-1,i} - ETX_{max} \right)}{\left(1 - ETX_{max} \right)} \right] \right]^{\frac{1}{\alpha}}$$
$$A = \frac{1 - p_{min}}{2 \times E_{max}} \tag{87.6}$$

p_{min} Predefined minimum forwarding probability
α The weighted factor for varying forwarding probability

We composed the forwarding probability Eq. (87.6) to change the forwarding probability variously according to ETX metric and residual energy of each node. When a node has high residual energy and the link has low ETX metric, the forwarding probability is high. Even a link has far lower ETX metric because of having good link quality, when a node has inadequate residual energy, the forwarding probability is low. Figure 87.1 shows the forwarding probability using $ETX_{max} = 45$, $\alpha = 1$, $p_{min} = 0.7$.

When forwarder node i is set to forward request packets by using the forwarding probability p, node i forward request packets to its 1-hop neighbor nodes like original AODV. On the other hand, forwarder node i is not set to forward request packet by using the forwarding probability p, node i discard the request packet. An example of this algorithm is shown in Fig. 87.2. When source node S has data packets to transmit, node S forwards request packet to its neighbor node 1 and 2. Node 1 has higher residual energy and ETX value between node S is good. In this case, node 1 has high probability to forward. Similarly, node 2 calculates the forwarding probability. But node 2 has lower residual energy, and ETX value between node S is bad. So node 2 has lower probability to forward.

87.2.3 Advanced EEPR Algorithm

Adapting only the forwarding probability of the proposed algorithm, a problem will be occurred when all nodes in the network has low residual energy. According to Eq. (87.6), a node with lower residual energy has lower forwarding probability. If all nodes in the network have low residual energy, most of forwarder nodes in the middle of the network discard request packets because of low forwarding probability. In this case, routing process can be failed continuously.

To solve the above problem, we proposed advanced EEPR algorithm considering the residual energy of its 1-hop neighbor nodes locally and the average value

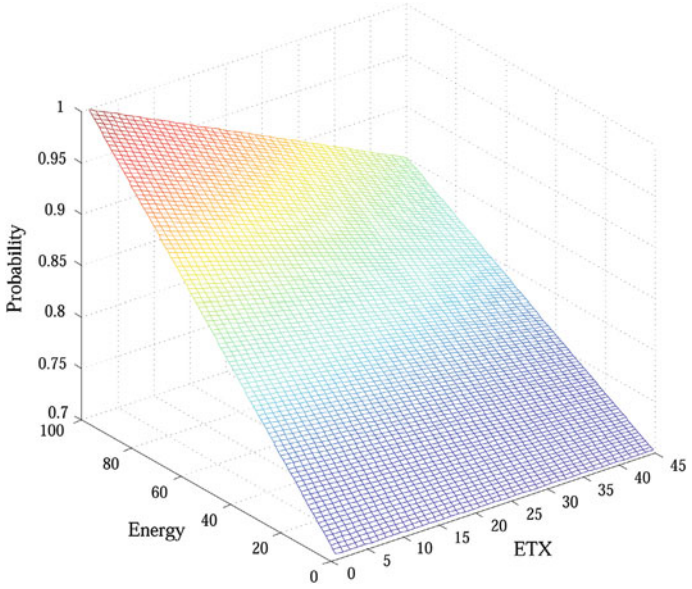


Fig. 87.1 The forwarding probability via ETX and residual energy

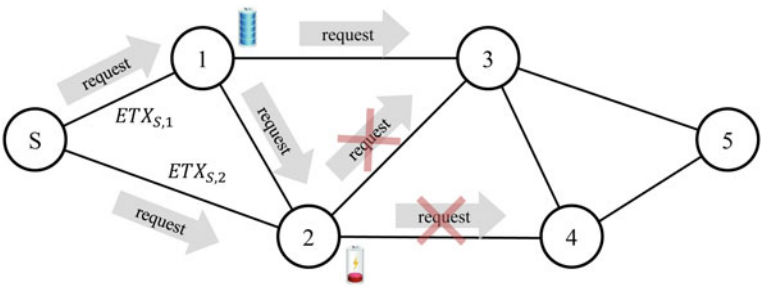


Fig. 87.2 An example of the EEPR algorithm

of residual energy of all nodes in the network. To propose advanced EEPR algorithm, we assume two factors. First, we assume that each node knows the average value of residual energy of all nodes in the network, E_{avg}. Second, we assume that each node usually knows the residual energy of its 1-hop neighbor nodes by using hello packet.

We define new factor, E_{th}, to propose advanced EEPR algorithm. E_{th} is predefined threshold value. According to the EEPR algorithm, when source node needs a routing path, source node forwards request packets to its 1-hop neighbor nodes. A forwarder node that receives the request packet calculates forwarding probability p using its residual energy and ETX value. Prior to calculate forwarding probability, the node implements the advanced EEPR algorithm. First, the

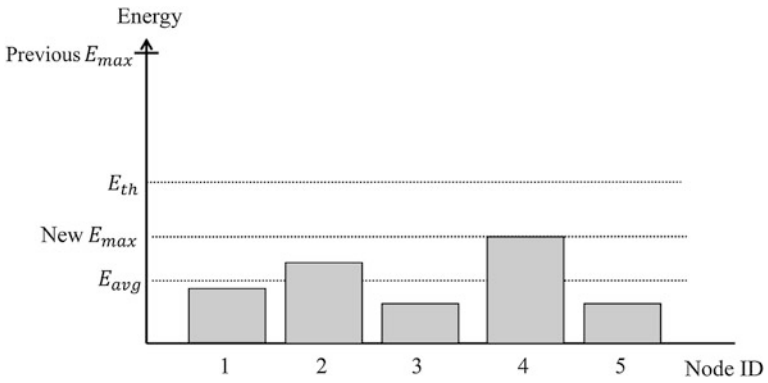


Fig. 87.3 An example of the advanced EEPR algorithm

node compares the average value of residual energy of all nodes, E_{avg}, with the predefined residual energy threshold, Eth. If, E_{avg} is bigger than E_{th}, the node thinks that the network is in a good energy condition and it is not necessary to make the forwarding probability higher. So, the node calculates the forwarding probability same as Eq. (87.6). If, E_{avg} is smaller than E_{th}, the node thinks that the network is in a low energy condition and implements the advanced EEPR algorithm. A node makes the forwarding probability higher by implementing advanced EEPR algorithm. Each node defines the maximum value of its neighbor node's residual energy as a new E_{max}, E_{max}^{new}, in place of previous E_{max}, E_{max}^{previ}. And each node calculates the new forwarding probability. This algorithm is shown as below.

$$p = \left[p_{min} + E_i A \left[1 + \frac{\left(ETX_{i-1,i} - ETX_{max} \right)}{\left(1 - ETX_{max} \right)} \right] \right]^{\frac{1}{\alpha}}$$
$$A = \begin{cases} \frac{1-p_{min}}{2 \times E_{max}^{previ}}, If, E_{avg} > E_{th} \\ \frac{1-p_{min}}{2 \times, E_{max}^{new}}, If, E_{avg} \leq E_{th} \end{cases} \tag{87.7}$$

An example of the advanced EEPR algorithm is as Fig. 87.3. Node 3 has to calculate the forwarding probability p and has 1-hop neighbor nodes, node 1, 2, 4, 5. Before calculating forwarding probability p, node 3 has to compare the average value of residual energy of all nodes in the network, E_{avg}, with E_{th}. In this example, E_{avg} is lower than E_{th}. So, node 3 has to implement advanced EEPR algorithm. According to the residual energy of 1-hop neighbor nodes, node 4 has the best residual energy. Therefore, node 3 replaces the previous E_{max} value with the residual energy value of node 4. Then, node 3 calculates the forwarding probability p and determines whether to forward or not the request packet.

In case of using the advanced EEPR algorithm, E_{avg} is global factor to globally consider the energy condition of the whole of the network. Also, the algorithm locally considers the biggest residual energy among its 1-hop neighbor nodes and

Table 87.1 Simulation factors used in the simulation

Simulation factor	Value
Topology	1,000 m by 1,000 m grid random
Number of nodes	50
Path loss model	$128.1 + 37.6 \log_{10}(\text{dist.(km)})(\text{dB})$
Noise power	$10^{-11} W$
Transmission range	300 m
Packet size	1,000 bytes
Initial node energy	$10 \sim 100$ *J*, uniform distribution
Transmission power	0.1 mW
Trans. consumption power	1.65 W
Recv. consumption power	1.1 W
E_{max}	100
E_{th}	40
ETX_{max}	45
P_{min}	0.7
α	1

defines the new E_{max}. The algorithm confirms the difference of the residual energy among local nodes, and reacts to local network energy condition.

87.3 Simulation

In this paper, we simulated our EEPR algorithm and compared with original AODV protocol about 4 simulation performance factors. As we aimed to achieve energy efficiency in network, we simulated about energy related 2 simulation performance factors.

87.3.1 Simulation Setup

Simulations were performed by NS-2 simulator version 2.35 on the Linux Fedora 13. Simulation parameters that are used in simulation are shown as Table 87.1. In one simulation iteration, a source node request routing path and transmit 10 data packets that each size is 1,000 bytes size. Simulation has 1,500 iterations and terminated when all nodes wear out residual energy of nodes in the network. Initial residual energy of each node is set to uniform random distribution between 10 and 100 J.

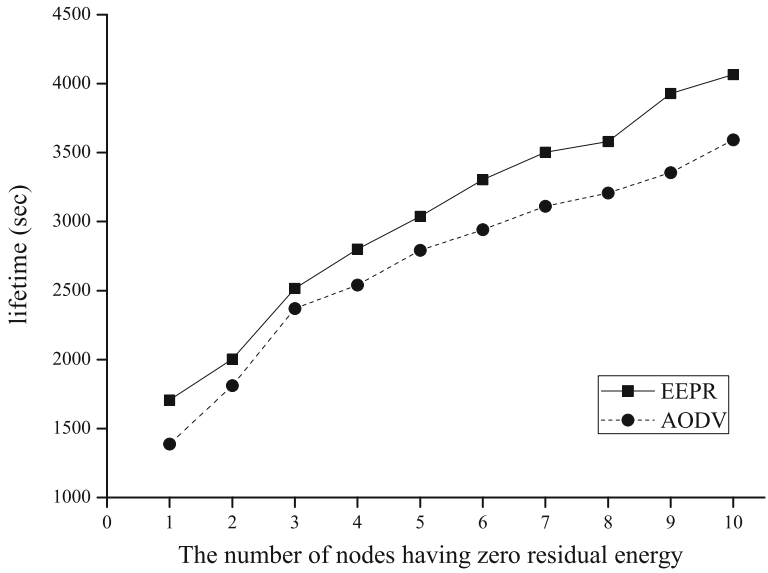


Fig. 87.4 The network lifetime using different routing protocols

87.3.2 Performance Metrics

Network lifetime. We define the network lifetime as the time that a node has zero residual energy and cannot forward packets anymore. We measure the network lifetime from the time that 1st node is worn out its residual energy to the time that 10th node is worn out its residual energy.

The result of the network lifetime is shown as Fig. 87.4. Nodes using the EEPR algorithm have about 12.57 % higher network lifetime than nodes using original AODV protocol. As a result, EEPR algorithm uses residual energy of all nodes in the network more evenly than the original AODV.

Variance of residual energy. We measure the residual energy of all nodes and calculate the variance of residual energy at the end of the communication process. It means the difference of residual energy of nodes in the network. And the smaller the variance is, the more evenly the algorithm use residual energy of nodes. Therefore we verify the performance of the EEPR algorithm by comparing with original AODV about the variance of residual energy of all nodes. We repeat this process and average these variances over the number of simulation iteration.

The result of the variance of residual energy of all nodes in the network is shown as Table 87.2. The variance of residual energy of nodes using EEPR algorithm is 62.4 smaller than the result using original AODV protocol. The simulation of original AODV protocol and EEPR algorithm had same initial residual energy and the variance of residual energy. According to the simulation

Table 87.2 Simulation Results as to performance Metrics

Performance Metric	EEPR	AODV
Variance	751.3705	813.7303
Routing delay	4.3319 ms	3.9216 ms
Routing success probability	91.7647 %	93.5335 %

result that difference was occurred by using different algorithms, the EEPR algorithm brought nodes in the network spending the residual energy more evenly than original AODV protocol.

Delay time in routing process. According to the EEPR algorithm that stochastically controls the number of request packets, it might be happen that forwarder nodes in the middle of the network do not forward request packets frequently. This result in larger routing delay compare to original AODV protocol. In this paper, we defined the routing delay as time difference between the time that a source node forwards request packets and the time that a destination node receives the first request packet.

The result of the routing delay time is shown as Table 87.2. Routing setup process using EEPR algorithm has about 0.4 ms higher routing delay time than the result using original AODV protocol. Because the EEPR algorithm stochastically control the number of request packets, it is definite that routing setup process using the EEPR algorithm always have higher routing delay time than the original AODV protocol that forward routing packets all the time. But 0.4 ms routing delay time more than the time using the original AODV protocol has no a decisive effect in actual routing setup process.

Routing success probability. The EEPR algorithm stochastically controls the number of request packets. Therefore, like the simulation about delay time in routing process, it might be happen that forwarder nodes in the middle of the network do not forward request packets frequently. It is important that the EEPR algorithm have to verify similar results compare with original AODV protocol even if the EEPR algorithm has energy efficient performance. The result of the routing success probability is shown as Table 87.2. Because the EEPR algorithm stochastically controls the number of request packets, it is definite that routing setup process using the EEPR algorithm always have lower routing success probability than the original AODV protocol that forward routing packets all the time. But the routing success probability of the original AODV protocol is 93.5335 %, and the routing success probability of the EEPR algorithm is 91.7647 %. It is about 1.8% lower than one of the original AODV protocol. But this result has no a decisive effect in actual routing setup process.

87.4 Conclusion

In this paper, we confirmed that the energy efficiency is key factor in MANET environment and proposed Energy-efficient Probabilistic Routing (EEPR) using residual energy and ETX metric as routing metrics. The EEPR algorithm stochastically controls the number of request packets using residual energy and ETX metric. Therefore the EEPR algorithm reduces the number of routing request packets and packet collisions in density network environment. And the EEPR algorithm facilitates packet delivery that has energy efficiency and considers link quality. In simulations, we verified that the EEPR algorithm has longer network lifetime and consumed residual energy of each node more evenly than the original AODV protocol while the routing delay time is slightly raised and the routing success probability is slightly decreased.

Acknowledgments This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2012-H0301-12-4004) supervised by the NIPA (National IT Industry Promotion Agency). This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No.2012R1A2A2A01014170). This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No.2011-0024132).

References

1. Allard G (2006) Evaluation of the energy consumption in MANET. In: Proceedings of the 5th international conference on Ad-Hoc, Mobile, and Wireless Networks. ADHOC-NOW'06, pp 50–69, Aug 2006
2. Perkins CE, Royer EM (1999) Ad-hoc on-demand distance vector routing. In: Proceedings of the Second IEEE workshop on mobile computing systems and applications, WMCSA '99, pp 99–100, Feb 1999
3. Johnson DB (1996) Dynamic source routing in Ad Hoc wireless networks. Mob Comput 353:153–181
4. Wang X (2004) An energy-aware probability routing in MANETs. In: Proceedings of the IEEE workshop on IP operations and management, pp 146–151, Oct 2004
5. Patil AP (2011) Design of an energy efficient routing protocol for MANETs based on AODV. Int J Comput Sci Issues 8(4), Jul 2011
6. Nand P (2011) Probability based improved broadcasting for AODV routing Protocol. In: International conference on computational intelligence and communication networks (CICN), pp 621–625, Oct 2011
7. De Couto DSJ (2005) A high-throughput path metric for multi-hop wireless routing. Wireless Netw 11(4):419–434
8. Zhu J (2009) An energy balanced reliable routing metric in WSNs. Wireless Sens Netw 1(1):22–26
9. ITU-R Recommendation M.1225 (1997) Guidelines for evaluation of radio transmission technologies for IMT-2000. International Telecommunication Union

Chapter 88
Functional Annotation of Proteins by a Novel Method Using Weight and Feature Selection

Jaehee Jung, Heung Ki Lee and Gangman Yi

Abstract The definition of the automatic protein function means designating the function with the automation by utilizing the data that already revealed unknown protein function. The demand for analysis on the sequencing technology such as the next generation genome analysis (NGS) and the subsequent genome are on the rise; thus, the need for the method of predicting the protein function automatically has been more and more highlighted. As for the existing methods, the studies on the definition of function between the similar species based on the similarities of sequence have been primarily conducted. However, this paper aims to designate by automatically predicting the function of genome by utilizing InterPro (IPR) that can represent the properties of the protein family, which similarly groups the protein function. Moreover, the gene ontology (GO), which is the controlled vocabulary to describe the protein function comprehensively, is to be used. As for the data used in the experiment, the analysis on properties was conducted in the sparse state that is deflected to one side. Thus, this paper aims to analyze the prediction method for protein function automatically through selecting the features, assigning the data processing and weights and applying a variety of classification methods to overcome that property.

Keywords Gene ontology · GO · InterPro · IPR · Functional annotation · Gene annotation · SVM · SMO · Adaboosting

J. Jung · H. K. Lee
Samsung Electronics, Suwon, South Korea
e-mail: cleme76@gmail.com

H. K. Lee
e-mail: heungkilee@gmail.com

G. Yi (✉)
Gangneung-Wonju National University, Gangwon, South Korea
e-mail: gangman@cs.gwnu.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_88, © Springer Science+Business Media Dordrecht 2014

88.1 Introduction

With the rapid development of the next genome sequencing technology (Next Generation Sequencing), the proteins whose function has not been defined are on the rise and it takes a lot of time and economic costs to assign the function manually. Therefore, the requirements for the automatic annotation processing (functional gene annotation) that can designate the protein function automatically have been gradually increased. As for the traditional method, the biologists defined the function by conducting the experiment manually in order to define the protein function. But there are some limitations in terms of time and efforts due to the explosive increase in data with the development of sequencing equipment in order to define the protein function by a manual method. As for new proteins that have not been revealed experimentally, the protein functions could be automatically annotated by the automatic annotation processing if the model was created by using the known protein function; therefore, it would be possible to predict the function through the much more reduced time than the existing method based on experimentation. Many of the existing studies have conducted the researches by utilizing the DB (protein functional site, gene expression, etc.) in which the pattern of base sequence, similarities and protein function, in order to define the protein function. Of those, InterPro (IPR) [9] defines the protein group (family), domain, etc. as the database to be used to define those unknown new protein functions. In the case of newly revealed proteins through InterProScan [15], IPR can be easily obtained. Thus, it would be appropriate to use it as a feature to define a function. In addition, Gene ontology [2] can be regarded as a language to represent the properties of genes, and it consists of the three independent hierarchical structures that include molecular function, biological process and cellular component. In the consortium, the database is still in progress at present in order to describe the comprehensive feature of genome. And many of the search tools and automation classification tools, which are based on this, have been researched. This paper aims to analyze the methods that allow for the automatic annotation processing through comparing the relationship between IPR and Gene ontology by a variety of methods as utilizing the known data.

88.2 Related Studies

In the case of the functional annotation that identifies the function of genome automatically, various studies related to the database and automatic annotation are in progress in order to define the function of genome of various species from human beings to small microorganisms. The studies are in progress to allow for the automatic prediction of the protein function by an easy access through web or automatic installation, and mostly the database to manage this systematically is also in progress as it has continued to be updated. However, the method of defining

the protein function automatically is still at the initial phase; thus, the accuracy is not very high. In the case of using the interpro2go that was manually mapped as in GOA [3], mapping is relied for defining the function; thus, the accuracy is not high. Most studies are conducted by small species to increase the accuracy, and the prediction methods for the protein function based on the calculation that has been researched makes a judgment mostly by utilizing the similarities of sequence.

As for the most frequently used tools, they include Gotcha [13], OntoBLAST [16], Blast2GO [4], AutoFACT [11], etc. Gotcha [13] can have the similarity of sequence and the Directed Acyclic Graph (DAG) of Gene ontology, in other words, the parent node can have several offspring nodes. Thus, it is the method of utilizing the property in which the parent node means the functions of more comprehensive meaning. It is the method of automatically naming GO by assigning a score to GO owned by the genome that is determined to be similar by judging the similarity of sequence. Blast2GO [4] is the method of annotating new protein functions that cannot be known by Gene ontology that is owned by a similar sequence after judging the similarity of sequence by utilizing BLAST [1]. It is the prediction model for the accuracy by assigning weights in accordance with the evidence code that is the annotation code of GO at this point. The evidence code means the code of GO to indicate whether it is automatically named (IEA) and it is determined by the similarity (ISS). OntoBLAST [16] is the method of finding possible protein functions from GO, which are obtained also from BLAST search. AutoFACT [11] proposed a fast annotation method by utilizing BLAST with the relevant database.

88.3 Methods

88.3.1 Features of Data

The data to be used is Saccharomyces Cerevisiae; it is one of yeast fungus; thus, it belongs to the fungus class and it is the most well-known data by the experiments. Since it forms a relatively small dataset as compared with the other species and it already comes out its related function; thus, it would be an appropriate data for establishing a model for the automatic annotation processing. For the extraction of this data, 4,370 proteins could be obtained as a result of searching and extracting Saccharomyces Cerevisiae only from SWISS-PROT.

The property to be used as a feature to create a model of data is IPR. IPR has the appropriate features for the reference data that include the protein family binding the protein functions in a similar way and the functions of Prosite, Prints, Pfam, Prodom, SMART, TIGRFams and PIR SuperFamily that play the central role to refer to the functional domain database. GO is utilized as the reference data for defining the function automatically. GO forms a hierarchical structure and divided

into the three big classes—cellular component, molecular function and biological process.

When counting the total number of IPR and GO term possessed by the 4,370 extracted proteins of Saccharomyces Cerevisiae, it was found to have 2,624 IPRs and 2,438 GO terms. When this data had one of the properties of IPR or GO term for each protein, it was represented in a binary form. It is represented in a large matrix (4,370 × 2,438) of GO in a binary form by representing "1" when the proteins have one term of particular GO terms and "0" when they do not have one as parsing Gene ontology at Ontology in the data section of SWISS-PROT. Also as for IPR, the IPR data was configured in a matrix form of 4,370 × 2,624 by a matrix of binary form as representing whether each protein has it through listing IPRs possessed by Saccharomyces Cerevisiae proteins after extracting InterPro in the family and domain database section with the same method as described above. A diagram for representing GO of IPR for each protein in a matrix and lining up the quantity of GOs that can be represented by "1", in other words, the quantity owned by the proteins would be the same as shown in Fig. 88.1. As shown in Fig. 88.1, it has the problem that it does not have a sufficient quantity for each GO to conduct the learning. When viewed from the perspective of one single GO, the number of case in which it has only one single GO is 414. This means that only one protein owns the relevant GO; thus, it would not appropriate to utilize it as the learning data. In addition, the validity was tested through the 10-fold cross validation; thus, GO that has fewer quantity than a certain level would not be appropriate for the use as the learning data.

However, the biggest problem of the data is that the data exists sparse even though it has a relatively sufficient quantity to be utilized as the learning data. And the fact that the protein data not having the relevant IPR or GO are inclined to one side than the data having the relevant IPR or GO when viewed from a particular IPR or GO term is also a problem. For instance, only 50 proteins have a particular GO out of 4,370 proteins when viewed from the perspective of a particular single GO; therefore, they are represented by "1" and the remaining 4,320 proteins are represented by "0" since they do not have it. In the case of conducting the learning and experiment with such data, it is quite often predicted that most do not have it since the learning is conducted as being excessively inclined to "0" that is not owned by the learning result; thus, it cannot become an effective model for the automatic function prediction and command processing that has to assign new functions. There are many cases represented by "0" representing "not having" in the case of IPR in addition to GO. This cannot be utilized as an effective feature. Due to such properties of these two features, this paper aims to apply the method as to the feature selection and balanced dataset. Moreover, it aims to analyze the results by converting the binary form of the data into a non-binary form (weighted IPR) by utilizing the correlation coefficients since the data to be processed is not a binary form.

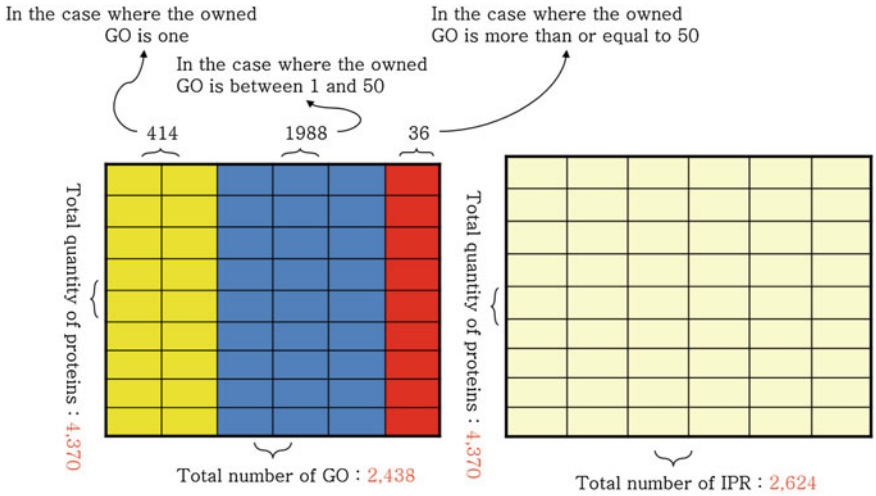


Fig. 88.1 Features of data

88.3.2 Prediction Method of Protein Function

This paper aims to compare and analyze the prediction method of protein function by utilizing the data having a sufficient quantity of data as the data of learning and experiment of this paper among the data described in 3.1. Before the analysis, it would be essentially required to have a process of reconfiguring it as a balanced dataset due to the feature of not being balanced with the sparseness of data. It aims to compare the case of applying the feature selection by the three mutually different learning methods and the case of conducting the weighted IPR that adds weights to the data respectively.

This paper shows to comparison and analysis of the prediction methods by the methods presented in Table 88.1. First as for the learning methods, adaboosting [7] is the method of creating an optimal classification through several times of learning by assigning weights as to the instances wrongly classified by the method of weak learner. SVM is the method of seeking a boundary that makes the error of margin that can differentiate the class to be classified at the hyperplane; thus, it is one of the learning methods of machine learning. SMO [10] is the most well-known tool of libsvm [5]; thus, it can be regarded as the method that has simplified the complexity of SVM by the sequential minimal optimization. As for the methods to be presented in Sect. 88.3.2.2., the case of using the feature selection method and the case of not using it were compared and analyzed as W/O in Table 88.1 meant Without and W meant With. Furthermore, it compared the case of using the method called weighted IPR to be stated in Sect. 88.3.2.3 with the case of using the original data as it was.

Table 88.1 Learning methods

Data\method	Original IPR		Weighted IPR	
Adaboosting	W/O feature selection	W feature selection	W/O feature selection	W feature selection
SVM	W/O feature selection	W feature selection	W/O feature selection	W feature selection
SMO	W/O feature selection	W feature selection	W/O feature selection	W feature selection

88.3.2.1 Dataset Reconfiguration to Adjust Balance

As shown in Fig. 88.2, there are more proteins not having the relevant GO than those having it when viewed based on a particular GO. However, there are more proteins not having a particular GO, in other words, negative proteins, when learning with such data; therefore, there would be a high degree of probability for the modeling that most of learning results turned out to be not having it. However, it is only possible to find it out by creating a model having the relevant GO rather than a model not having GO. When experimenting with proteins that are not able to perform the function, it is impossible to obtain the desired result. Thus, balanced sampling approach is employed to overcome this handicapped data property.

There are the under-sampling method and the over-sampling method in terms of reconfiguring the data that consists of balanced proteins; the over-sampling [6] is the method of making the number equal by generating the data that become the major in terms of quantity as many as the quantity at which the relatively fewer data becomes the major in a random way. The under-sampling [12] is the method of meeting the ratio by selecting more data randomly based on the data whose quantity is few. In this experiment, the data that is relatively few in quantity is more important information; therefore, this paper reduces that quantity by utilizing the under-sampling. As shown in Fig. 88.2, the data indicated by "1" is to be named as positive protein, whereas the data indicated by "0" is to be named as negative protein. And it is supposed to learn with proteins that are fewer than 4,370 in terms of the quantity of protein by reconfiguring the data for the learning model at each GO through selecting the negative proteins just as many as the quantity of positive proteins.

88.3.2.2 Feature Selection

As shown above Fig. 88.1, IPR has the matrix of many binary features of 2,624 when viewed based on one GO. It is the well-known fact that learning and experimenting by selecting only meaningful features would reduce the time to be taken and have a better result as compared with learning and experimenting the method presented above by these many matrices [8, 14]. When representing the case in which "1" representing that each IPR has protein by positive data and

Fig. 88.2 Data reconfiguration using under-sampling

the case of not having it by "0", the positive negative data is to be counted for each protein. The positive data is represented as "IPos", "GPos" and the negative data is represented as "INeg", "GNeg" at IPR and GO in respectively and it is possible to classify the state of IPR and GO for each protein. They can become 4 states as shown in Table 88.2.

It is possible to calculate the four probabilities (N_{GPos_IPos}/N_{Pos}, N_{GNeg_IPos}/N_{Pos}, N_{GPos_INeg}/N_{Neg}, N_{GNeg_INeg}/N_{Neg}) by utilizing the 4 data, where N_{Pos} stands for the total number of positive proteins and N_{Neg} means the total number of negative proteins. These probabilities represent a conditional probability, which the Gene ontology term may possess depending on the conditions of each IPR. When viewed the property by adding these conditional probabilities as an example of GO:0000329, the diagram as shown below could be viewed. The x-axis means several IPRs that are being experimented and the y-axis is the value of adding the conditional probabilities. It is possible to see the phenomenon of which most are concentrated in Area 1. On that account, 99 % of them are those IPRs having negative IPR term and also negative Gene ontology (Fig. 88.3). This paper selected the features based on the IPRs that are concentrated in Area 2 as excluding these IPRs. In other words, this is the learning method of utilizing only the selected index as a feature by selecting only the index of IPRs in Area 2 among the 2,624 IPRs by calculating the conditional probabilities above for each GO.

Table 88.2 Number of cases in accordance with the status of IPR and GO

	Positive IPR term	Negative IPR term
Positive gene ontology	N_{GPos_IPos}	N_{GPos_INeg}
Negative gene ontology	N_{GNeg_IPos}	N_{GNeg_INeg}

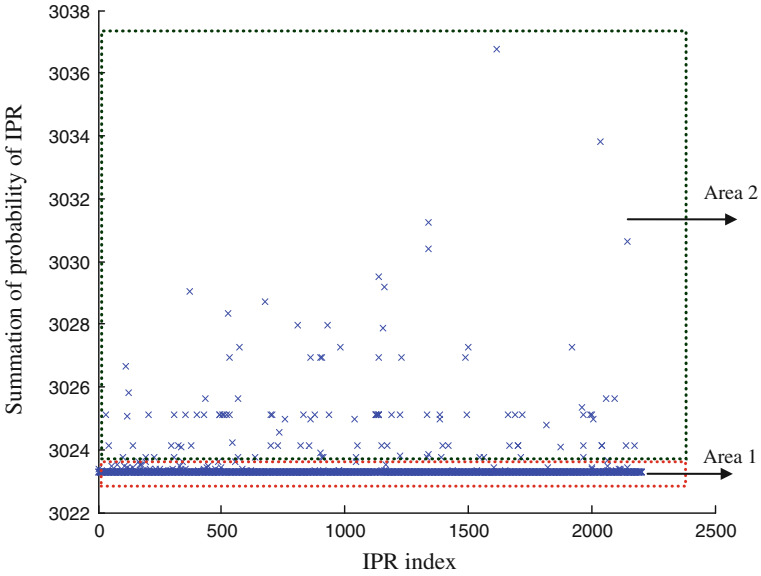


Fig. 88.3 Plot the summation of probability of IPR in terms of GO:0000329

88.3.2.3 Weighted IPR

IPR that is utilized as the feature is the binary data that consists of 0 and 1. When converting this data into a continuous form rather than binary form by utilizing a correlation coefficient, IPR feature data would be expected to select a feature without partiality. This paper aims to analyze the performance between the two methods by the differences between the feature extraction using the binary data that consists of 0 and 1 and the weighted IPR of a continuous form as naming this data as the weighted IPR.

For instance, as shown in Table 88.3, the table that is composed of 0 and 1 would be modified into a table that utilizes a correlation coefficient (Table 88.4). A correlation coefficient becomes a value closer to 1 with a higher degree of correlation, whereas it is represented by a value close to 0 when there is no correlation. In addition, it becomes a negative value when there is a mutually contradicting correlation.

This paper aims to change to weight coefficients as proposed by Formula 1. Based on this correlation coefficient. First, each protein p possesses IPR from 1 to n.

Table 88.3 Original IPR

	IPR1	IPR2	IPR3	IPR4	IPR5	IPR6
Protein1	0	1	1	1	0	0
Protein2	1	1	0	0	1	0
Protein3	0	0	0	0	0	1
Protein4	1	1	1	0	0	1
Protein5	0	1	0	1	0	0

Table 88.4 Correlation Coefficient among the IPRs

	IPR1	IPR2	IPR3	IPR4	IPR5	IPR6
IPR1	1.0000	0.4082	0.1667	−0.6667	0.6124	0.1667
IPR2	0.4082	1.0000	0.4082	0.4082	0.2500	−0.6124
IPR3	0.1667	0.4082	1.0000	0.1667	−0.4082	0.1667
IPR4	−0.6667	0.4082	0.1667	1.0000	−0.4082	−0.6667
IPR5	0.6124	0.2500	−0.4082	−0.4082	1.0000	−0.4082
IPR6	0.1667	−0.6124	0.1667	−0.6667	−0.4082	1.0000

Table 88.5 Weighted IPR

	IPR1	IPR2	IPR3	IPR4	IPR5	IPR6
Protein1	0.5251	0.2076	0.1462	0.1462	0.1376	−0.1628
Protein2	0.2008	0.1295	−0.2501	0.3750	0.1697	0.3750
Protein3	0.1389	0.3934	0.0889	−0.1334	0.0122	0.5000
Protein4	0.2633	0.0725	0.2633	0.2500	0.2500	-0.0991
Protein5	0.7990	0.2500	−0.0633	0.2500	−0.1724	−0.0633

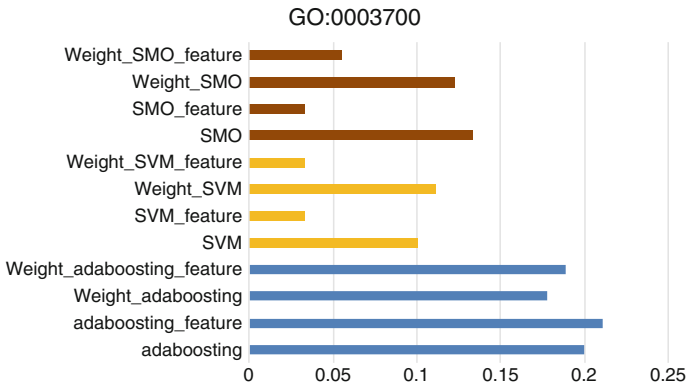


Fig. 88.4 Error rate applying several methods for GO:0003700

Table 88.6 Error rate for each GO term

GO	Error rate using SVM with feature selection	Error rate using SMO with feature selection	Error rate using Weighted IPR SVM with feature selection	Errorrate using weighted IPR SMO with feature selection
GO:0000324	0.030303	0.045455	0.030303	0.166667
GO:0000329	0.090909	0.136364	0.075758	0.257576
GO:0000398	0.025641	0.025641	0.025641	0.128205
GO:0003677	0.055556	0.055556	0.051852	0.096296
GO:0003700	0.033333	0.033333	0.033333	0.055556
GO:0003723	0.026316	0.026316	0.035088	0.096491
GO:0003735	0.033333	0.038889	0.053333	0.06
GO:0005515	0.128571	0.142857	0.144444	0.266667
GO:0005524	0.166667	0.177778	0.144444	0.266667
GO:0005730	0.116667	0.116667	0.108333	0.166667
GO:0005732	0.060606	0.060606	0.060606	0.106061
GO:0005743	0.208333	0.263889	0.208333	0.416667
GO:0005783	0.22069	0.234483	0.224138	0.275862
GO:0005789	0.075758	0.075758	0.106061	0.242424
GO:0005829	0.090909	0.090909	0.1	0.254545
GO:0005886	0.123077	0.123077	0.115385	0.182692
GO:0005935	0.111111	0.111111	0.041667	0.152778
GO:0006281	0.151515	0.181818	0.151515	0.227273
GO:0006355	0.133333	0.15	0.133333	0.166667
GO:0006365	0.075758	0.090909	0.075758	0.166667
GO:0006412	0.05303	0.05303	0.056818	0.079545
GO:0006457	0.030303	0.030303	0.045455	0.166667
GO:0006468	0.009804	0.039216	0.009804	0.058824
GO:0006511	0.030303	0.030303	0.030303	0.121212
GO:0006888	0	0	0	0.083333
GO:0006897	0.013889	0.013889	0.013889	0.152778
GO:0006950	0.090909	0.106061	0.060606	0.212121
GO:0007047	0.083333	0.125	0.092593	0.148148
GO:0009060	0.066667	0.066667	0.066667	0.316667
GO:0009277	0	0	0	0.016667
GO:0016020	0.05	0.05	0.05	0.166667
GO:0016021	0.029412	0.029412	0.019608	0.078431

All the proteins to be experimented are represented by IPR of n units. A particular protein having IPR would be represented by 1, whereas those not having IPR would be represented by 0.

For instance, Protein 1 in Table 88.3 is represented as not having IPR 1, IPR 5 and IPR 6, which are 0, whereas IPR 2, IPR 3 and IPR 4 are represented by IPR possessed by the relevant protein. At this point, there is a relationship between IPR 5 and IPR 6 since IPR 1 is not a property that is not owned when viewed by each IPR of Protein 1. In reference with Table 88.4, the weight (IPR1) value of

IPR 1 of corr(IPR1, IPR5) = 0.6124 and corr(IRP1, IPR6) = 0.1667 Protein 1 is 0.6124 + 0.1667 = 0. 7791. Moreover, the value of weight sum (IPR 1) is represented by IPR 1 = 0; therefore, the value of adding all the correlation coefficients of IPR 5 and IPR 6 becomes 0.7418. Essentially the value was the binary form of 0 and 1 in order to calculate the weighted sum (IPR 1) as to IPR 1 of Protein 1 of the calculated value; therefore, there finally generates the value of 0.5 × 0.7791/0.7418 = 0.5251 by giving the weighted value 0.5. A new data defined in the new weighted IPR would be generated by such method. Table 88.5 can be regarded as one of such cases.

$$P = \{IPR_1, \ldots, IPR_n\}$$

$$Weight(IPR_i) = \sum_{j=1}^{|P|} corrcoeff(IPR_i, IPR_j), \qquad \text{where } i \neq j$$

$$Weight_{sum}(IPR_i) = 0.5 \times \frac{Weight(IPR_i)}{\sum_{j=1}^{|P|} Weight(IPR_j)}, \quad \text{where } IPR_i = IPR_j$$

Formula 1. Converting weighted IPR by the correlations and weights

Figure 88.4 is the diagramed result by a graph of the error rate of applying the 12 methods presented in Table 88.1 to GO:0003700. The 4 mutually different methods presented in Table 88.1 were represented by different colors. And a better performance would be shown with fewer error; thus, selecting the weighted method and features suggested by SVM or SMO resulted in a similar or less error rate. As to the 4 methods utilizing the feature selection and weighted method among the proposed 12 methods, the results of comparing and experimenting each GO term are as shown in Table 88.6. As shown in Table, they are found to have a high prediction rate of more than a range of 97–99.

88.4 Conclusion

This paper compared and evaluated the performance that could define the protein function by applying the classification algorithm by utilizing the feature selection and data transformation. As for the data to be processed, the data having GO term has been composed in much less quantity than the protein not having GO term when viewed by individual GO term. In addition, IPR that is set as the feature point is sparsely distributed; thus, it becomes difficult to learn all the protein data through the general classification algorithm. Due to such limitations, the performance as to the automatic annotation was compared by various classification methods through extracting only the GO term having the standard level or more as the learning subject. Moreover, the performance with the original data was also analyzed by the method of using the binarized data as the correlation coefficient through converting it into a newly weighted coefficient.

However, as for the data sampling and feature selection processed in this paper, the GO term learned primarily the data of protein having a certain amount or more for the learning at Saccharomyces Cerevisiae; thus, there is the limitation that the quantity of learned data of GO term is small. If it is to learn by utilizing the data that includes a variety of species such as SWISS PROT in order to overcome this limitation, it will be possible to expect to utilize the automatic function prediction by learning more GO terms with the use of large quantity of data. Thus, this paper aims to study a learning method that is appropriate for this. In addition, it aims to prepare a base to allow for the automatic annotation by seeking for different features that can be utilized as a keyword in addition to IPR when trying to find out unknown protein functions by identifying the correlation with GO.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2013R1A1A2063006).

References

1. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402
2. Ashburner M, Ball CA, Blake JA et al (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25:25–29
3. Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R (2004) The gene ontology annotation (GOA) database: sharing knowledge in Uniprot with gene ontology. Nucleic Acids Res 32: D262–D266
5. Chang CC, Lin CJ (2011). LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol 2(3):27
6. Chawla N, Bowyer K, Hall L, Kegelmeyer P (2002) SMOTE: synthetic minority over-sampling technique. JAIR 16:321–357
4. Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18):3674–3676 Epub, Aug 4 2005
7. Freund Y, Schapire R (1996) A short introduction to boosting. J Japan Soc Artif Intell 14(5):771–780
8. Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182
9. Hunter S, Jones P, Mitchell A et al (2011) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40:D306–D312
10. John CP Sequential minimal optimization: a fast algorithm for training support vector machines
11. Koski LB, Gray MW, Lang BF, Burger G (2005) AutoFACT: an automatic functional annotation and classification tool. BMC Bioinf 6:151
12. Kubat M, Matwin S (1997) Addressing the curse of imbalanced training sets: one-sided selection. In: Proceedings of the fourteenth international conference on machine learning, pp 179–186
13. Martin DM, Berriman M, Barton GJ (2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. BMC Bioinf 5:178

15. Quevillon E, Silventoinen V, Pillai S et al (2005) InterProScan: protein domains identifier. Nucleic Acids Res 33:W116–W120
14. Shahib A Al, Breitling R, Gilbert D (2005) Feature selection and the class imbalance problem in predicting protein function from sequence. Appl Bioinf **4**(3):195–203
16. Zehetner G (2003) OntoBlast function: from sequence similarities directly to potential functional annotations by ontology terms. Nucleic Acids Res 31(13):803–3799

Chapter 89
Gaze Behavior Analysis System Based on Objects Using Trajectory Information

Jung-Min Oh, Sangjin Hong and Nammee Moon

Abstract There have been diverse attempts to analyze user preference in terms of gaze behavior on specific object from various viewpoint in the fields of HCI, eye tracking area, image processing technology, etc. However, it is not easy to get clear about what user looks at instead of where user watches so far. Because there are issues such as object representation method in multimedia content, eye tracking error compensation and analysis method, efficient data profiling method in a situation that both user and image data keep being increased. Thus, in this paper, we propose eye tracking error compensation method using trajectory information and data compression method using object information of content to solve these problems mentioned above. Finally, we verify data accuracy of the proposed system and reduction ratio of data in a situation if there is a random error.

Keywords User's behavior analysis · Object representation · Error compensation · Data compression · Eye tracking

89.1 Introduction

Recently, there are significantly increase of image content and video source which contains a lot of information such as special story, specific scene, important object all over the world. In this trend, there is an increasing demand to find out how user consumes the image content in terms of an object which implies the user's preference. The reason why it is growing is because there is a limitation to understand

J.-M. Oh · S. Hong
Stony Brook University, Stony Brook, USA
e-mail: aliibaba@naver.com

N. Moon (✉)
Hoseo University, Asan, South Korea
e-mail: mnm@hoseo.edu

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_89, © Springer Science+Business Media Dordrecht 2014

user's internal perception behavior or actual intention of usage based on external action information, namely, social activity information, content selection, and channel switching by using a remote controller. Although the need to also analyze nonverbal gaze behaviors of users is gradually increasing, it is not easy to search for contents foundation and analysis platforms based on objects that are supposed to support this analysis method. Merely, various preceding researches are suggesting specific consumption behavior analyses results by using eye trackers in the process that users consume contents [1–4]. However, regretfully, due to the high price of eye trackers and the limitation that they need to be supported by certain tools in restrained regions, it is not easy for them to diversely support real-time interactions with users. Moreover, even though gaze information of users can be saved through eye trackers, it can be hardly guaranteed that eye trackers would provide 100 % accurate location information, and accordingly there exists the possibility that the results of combining contents information and gaze information in real-time would leave errors [5].

Thus, this paper tries to suggest user's gaze behavior analysis system that acquires location information of real-time gaze data and then combines it with the image object representation structure, and it attempts to include not only the data processing method based on gaze trajectory to compensate errors that occur in the stage of gaze data acquisition, but the data compression processing process to efficiently lighten image data and user information that will get larger in the future.

The structure of this paper is as follows. First, the purpose and need of this study will be suggested in Sect. 89.1, and related theories will be examined in Sect. 89.2. The system model suggested in this paper will be explained in Sect. 89.3, and test results through experiments will be suggested in Sect. 89.4. Lastly, a conclusion will be drawn in Sect. 89.5.

89.2 Related Works

89.2.1 Image Object Representation Method

In order to analyze object data and gaze behaviors by linking them, image representation technology and image annotation technology are needed in advance. In the fields of computer vision and image processing, a lot of studies are being conducted to develop an automated annotation method by applying the learning technique and the similarity comparison method. At the same time, as digital images have been produced in large quantities, techniques with which anyone can identify and represent objects on images are being studied. For example, the video contents annotation using MPEG-7 [6], information detection based on subregions [7], polygon-based object annotation method [8], etc. were suggested. Of them, LabelMe studied in MIT is a data set that polygonizes the object area by clicking the outer line of the object on the image with a mouse, and inputs and saves the name of the object, and it is easy to use it in various application areas by

forming the annotation data into XML files [8]. However, it still only represents objects by targeting individual images, and so representation is limited when same objects exist throughout many image frames.

89.2.2 User's Gaze Behavior Analysis Method

Since studies on user's gaze behaviors started to be performed in 1870s, many researchers have tried to analyze them. Especially, separate from gaze information detection studies from the physiological perspective, user's gaze behavior analysis technology from the psychological perspective is technology widely used as the usability evaluation tool in the HCI area [1]. As user's gaze behaviors have come to be more easily detected due to the rapid development of visual sensor systems such as smart devices or eye trackers, user analysis researches are being continuously conducted from various points of perspectives such as the visual information recommendation [7], visual-based searching [9], analysis on consumers' decision-making [2], gaze information analysis based on areas of interest [3], analysis on the difference of gaze behaviors between the east and the west, etc.

However, due to the limitation of the image representation technology explained above, user's gaze behavior analysis has focused on 'Where' based on the location information, rather than on 'What'. In addition, despite the issue of the high price, the majority of the preceding researches selected the gaze information analysis method totally relying on eye trackers in the stage of user data collection. However, the eye tracking technology does not allow perfect gaze recreation, and thereby the possibility of error occurrence exists until today. That is, unless compensation for the risk of error occurrence is realized, fundamentally there is weakness that the analysis results cannot become reliable [5]. Furthermore, no clear solving method for the issue to process and save data with a large size that shall be used for analyses when the number of users and the amount of image data increase has been suggested yet.

89.3 Object Centered Gaze Behavior Analysis System

89.3.1 Structure of the System Model

As shown in Fig. 89.1, the system consists of six modules, that is, for securing data source, identifying image data structure, securing and processing gaze data, gaze behavior profiling, checking and monitoring rules, and data request. Each module exchanges data in real-time, closely interacting one another. First, the module for securing data source collects contents data and user's gaze data simultaneously. The module for identifying image data structure identifies and save data on the image defined in advance in accordance with the object representation model

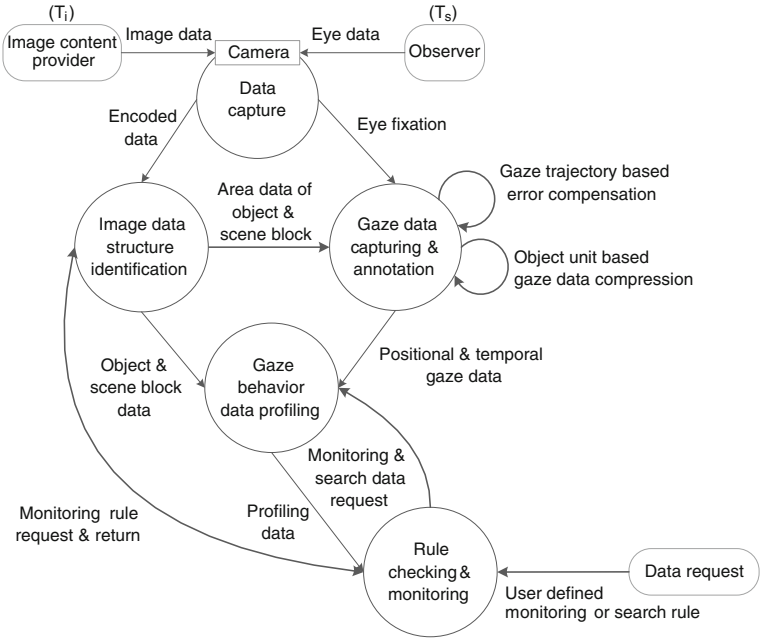


Fig. 89.1 Overall structure of the system

structure determined in this paper. And at the same time, the module for securing and processing gaze data recognizes user's gaze data in real-time fitting for the sampling time, and forms the secured gaze data into object-based gaze data by linking it to the location information. In this stage, if there exist any errors in the gaze tracing technology, gaze data whose errors are compensated by applying the analysis method considering the gaze trajectory is secured. And through executing data compression by the unit of the object, the profiling data structure that enables analyzing the optimum gaze behavior with a minimum amount of data is designed. Conclusively, the module for gaze behavior profiling saves user's gaze behavior data with the unit of the object whose errors are compensated and compression process is finished. The module for checking and monitoring rules provides event results by checking on user monitoring rules and executing real-time monitoring. Lastly, the module for data request supports more flexible analysis by enabling users to additionally define real-time and non-real-time monitoring.

89.3.2 Content Object Representation Structure

In this stage, the method that a contents producer represents the image contents structure to flexibly define an object in accordance with the size and characteristic

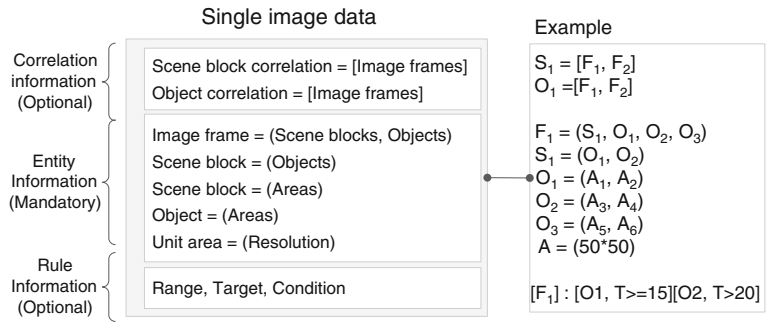


Fig. 89.2 Representation structure of image content

of it is explained. On the image frame, an object comes to contain more than one pixel, and for efficiency of calculation, the unit of the area that contains multiple pixels is used as the basic unit for representing and tracing the object. The size of the area is defined through resolution, and the appropriate size of the area is determined being influenced by the object size. When multiple objects have equal characteristics, the unit of the scene block is added to the contents structure in order to use the characteristics as one analysis unit. Because when there are multiple objects that users desire to see, they tend to bind them to one group and look at the center [10], the scene block comes to be an object group unit that can reflect this behavioral tendency. Thus, as shown in Fig. 89.2, image contents data is represented with image frame information, scene block information, object information, and area information. In addition, video contents contain frame connection information on the basis of the object and the scene block to describe that same objects are contained in many image frames and to represent the connection relation among image frames. And video contents also contain analysis rule information to perform the user's gaze behavior analysis in real-time. The analysis rule information consists of an analysis range, analysis object, and analysis condition, and provides the environment to check rules and execute the monitoring module at desired moment in real-time, from the point when contents start to be played. The analysis object and condition can be variously applied depending on the combination method, from one to many.

89.3.3 Gaze Data Analysis Model in Real Time

The eye movement information of users basically has two elements, fixations and saccades [11]. Gaze duration is calculated by multiplying the number of fixations that remain on the object by sampling time. On the image contents, fixations provide coordinate information of axis X and axis Y. This coordinate information is saved as temporary gaze data, and goes through the combination process with information of the object area or the scene block unit. Here, if the number of users or the amount of

video image information increases, the amount of data increases geometrically, and in order to prevent data processing from slowing down, the method of processing the gaze information compressed by the unit of the object comes to be selected. Thus, when temporary gaze information data and the object are combined, the gaze time information is not immediately saved, and the final saving process is delayed until gaze moves from one object to another. Figure 89.3 shows the gaze data processing process in brief by exemplifying seven gaze coordinates. While gaze moves from O_1 to O_2, the system temporarily saves the gaze point information and delays saving it as profiling data even during the combination process with the object. Later, it puts together all the time information from the point P_1 that gaze was first recognized on O_1 to when gaze moved out of O_1. Also for O_2, the final time information processed with P_5 for compression processing is saved.

At that time, if the gaze information processed in the eye tracking module is 100 % accurate, additional error compensation process is not needed. However, if there any errors exist in the gaze information coordinate processing, it is difficult to judge whether the gaze coordinate saved in the error area of the external area of the object is actually error data of looking at the object as shown by P_{n+1} in Fig. 89.4, or is accurate data of looking at another object breaking away from the former one. Thus, the gaze trajectory-based information processing method that additionally considers whether a user has an intention of 'Stay' to continuously look at the original object or of 'Transition' to turn his or her gaze to another spot is needed. As shown in Fig. 89.4, if P_{n+1} is in the range of an error, processing of the data is delayed, and if the gaze coordinate is re-detected on the existing object, namely O_1, P_{n+1} is saved as gaze data of that object. However, if gaze moves to another object later, namely O_2, P_{n+1} is processed as data irrelevant to the object.

89.4 Evaluation

89.4.1 Experiment Environment

In order to verify performance of this system, five image frames and data of five users are applied. Each image frame is exposed for one minute, totally five minutes, and contains minimum of one to maximum of four objects. Image frame 1 and 2, 2 and 3, and 3 and 4 are mutually related frames that contain same objects. On the image frame 1 and 3, objects are positioned close to each other, and on the image frame 2 and 4, objects are positioned distant from each other, and the object size gets larger as the image frame progresses from 1 to 5. It is assumed that the coordinate data for user data was collected from outside through the system module for securing data, and then random gaze data is generated. U_1 composes data with a pattern to calmly examine the whole image, U_2 to examine the whole image and move around objects, U_3 to focus on one object, U_4 to gaze the background instead of objects, and U_5 to let gaze move away and back again intermittently. Finally, in order to confirm the excellence of the error

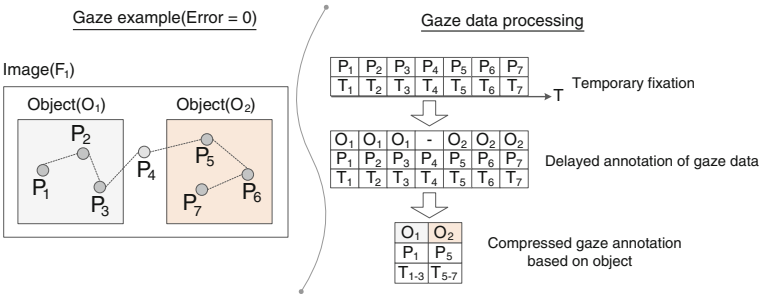


Fig. 89.3 Gaze data compression process depending on gaze examples

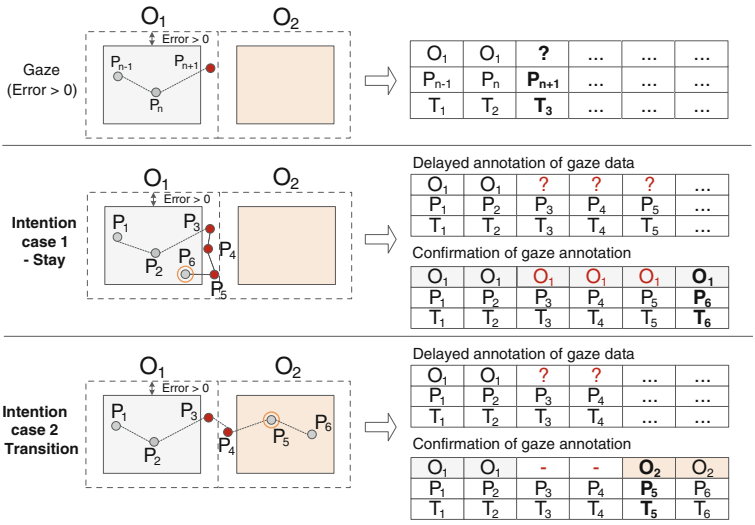


Fig. 89.4 Trajectory information-based gaze data processing for error compensation

compensation method of the system, random errors are applied after designating them to the range of 5, 25, 50, and 75 pixels in the stage of securing gaze data. In order to check real-time monitoring, the four rules, R1([−]:[O1, T ≥ 15][O2, T > 20]), R2([−]:[O3][O4][O5]), R3([−]:[O3][*][O4][*][O5]), and R4([−]:[O1,O2](50 %)) are applied.

89.4.2 Experiment Result

When there is no error, for the all five users, expected intention time and actual gaze time turned out to be equal as shown in Fig. 89.5, and as a result of executing real-time monitoring rules, the event results were precisely returned at the

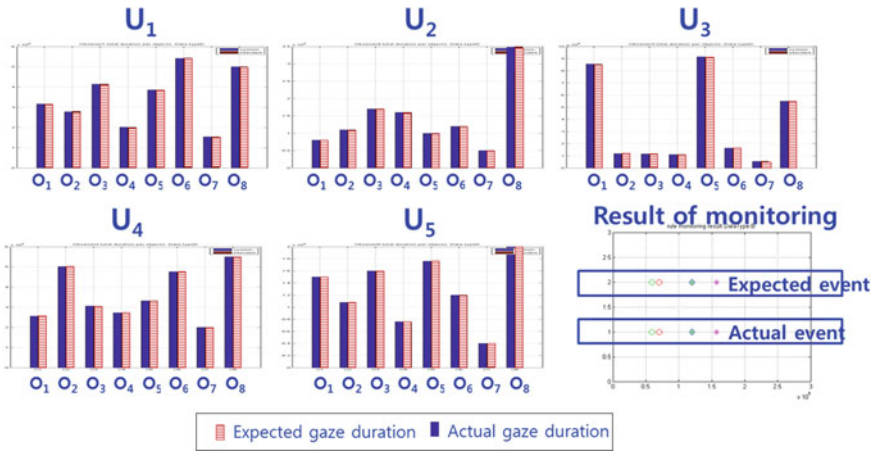


Fig. 89.5 Basic gaze data and monitoring results

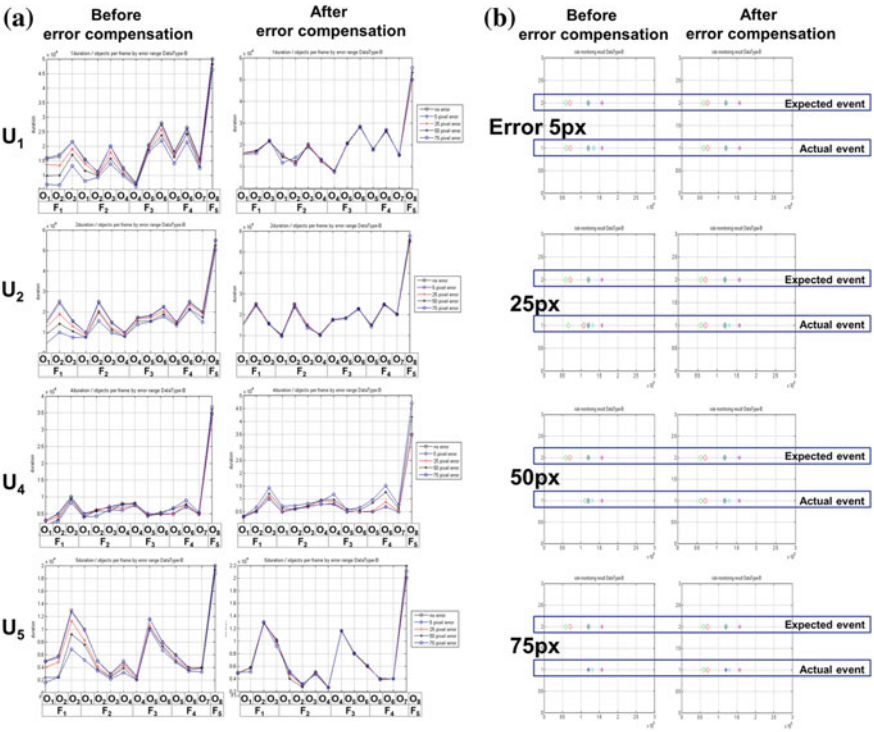


Fig. 89.6 Gaze time and monitoring according to pre/post-error compensation

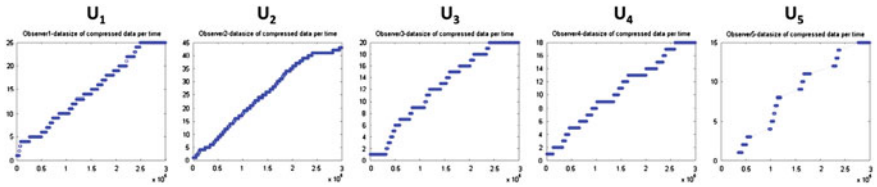


Fig. 89.7 Changes of data compression patterns according to users' gaze behaviors

expected time. Because, as shown in Fig. 89.5, gaze time was drawn by the unit of the object, it can be confirmed that securing gaze time and processing for the object were accurately performed even when equal objects were positioned through many image frames.

And when errors occurred, as a result of examining differences of object gaze data for each image frame by dividing the case that applied error compensation process and the case that did not, for U_1, the average of standard deviation depending on errors based on normal data was compensated from 10.1 to 1.9 %, U_2 from 10.9 to 1.2 %, U_3 from 11.9 to 5.6 %, U_4 from 7.3 to 16.2 %, and U_5 from 11.0 to 1.8 %. It can be confirmed in Fig. 89.6 that all of them except for U_4 were compensated close to the normal data, and the monitoring results also returned to the result values similar to those of expected events through error compensation.

Furthermore, as a result of applying object-based compression processing process for lightening the profiling data size, the average size of five users' data turned out to be 0.99 %, which is drastic size reduction. Looking into users' gaze patterns and data compression patterns, it can be found that, for U2, data compression time is shorter than those of other users, and as time for focusing on one object or for gazing at objects gets longer, compression processing can be realized more efficiently.

89.5 Conclusion

The gaze behavior analysis system suggested in this paper, even when there existed errors in the eye tracking data, reduced the standard deviation from normal data to maximum of 1.2 %, by applying the data compensation method considering gaze trajectory information, even in the environment that large error width, 5–75 pixels, existed. Merely for U4, it was found that the deviation of gaze data from normal data came to be enlarged after error compensation processing, indicating that the method suggested in this paper cannot work efficiently for users who gaze at backgrounds more than objects. In the aspect of the data size, the profiling data size was reduced to below 1 % on average compared to the early data that wasn't applied processing. Especially, users who have high concentration on objects show lower reduction rate of the data size (Fig. 89.7).

It is expected that, this suggested system, by enabling efficient real-time monitoring for object-based gaze behaviors through inclusion of all the connection information, element information, and rule information into the image representation structure, and at the same time, by realizing accurateness of gaze data and lightening of the data size in preparation for eye tracking errors, will be widely used in environments such as mobile, internet of things, etc. in the future. This author intends to study eye tracking patterns customized depending on users' gaze speed and object sizes, and more user-friendly analysis platform methods.

References

1. Jacob RJK, Karn KS (2003) Eye tracking in human–computer interaction and usability research: ready to deliver the promises. Mind's Eye: Cogn Appl Aspects of Eye Mov Res 5:573–607
2. Castagnos S, Pearl P (2010) Consumer decision patterns through eye gaze analysis. In: IUI, pp 1–10
3. Rajashekar U, van der Linde I, Bovik AC, Cormack LK (2008) GAFFE: a gaze-attentive fixation finding engine. IEEE Trans Image Process 17(4):564–573
4. Chua HF, Boland JE, Nisbett RE (2005) Cultural variation in eye movements during scene perception. Proc Natl Acad Sci USA 102(35):12629–12633
5. Hornof AJ, Halverson T (2002) Cleaning up systematic error in eye tracking data by using required fixation locations. Behav Res Methods, Instrum Comput Spec Issue Eye Mov Res 34:592–604
6. Bloehdorn S, Petridis K, Saathoff C et al (2005) Semantic annotation of images and videos for multimedia analysis. The semantic web: research and applications lecture notes in computer science, vol 3532, pp 592–607
7. Yoshitaka A, Wakiyama K, Hirashima T (2006) Recommendation of visual information by gaze-based implicit preference acquisition. Lecture Notes in Computer Science, vol 4351, pp 126–137
8. Russell BC, Torralba A, Murphy KP, Freeman WT (2008) LabelMe: a database and web-based tool for image annotation. Int J Comput Vision 77(1–3):157–173
9. Rajashekar U, Cormack LK, Bovik AC (2004) Point of gaze analysis reveals visual search strategies. Human Vision Electron Imaging 4(2004):1–12
10. Fehd* HM, Seiffert AE (2008) Eye movements during multiple object tracking: where do participants look? Cognition 108(1):201–209
11. Jacob RJK (1995) Eye tracking in advanced interface design. Virtual Environ Adv Interface Des

Chapter 90
The Crack Width Calculation Program for Structural Concrete Members Based on a Limit State Design

HwaMin Lee, DooSoon Park and GiYeol Lee

Abstract Serviceability performance verification of concrete structural members predicts whether the crack width satisfies requirements at the service load step. The correct calculation of steel stress is important for serviceability performance verification. In this paper, we define the material characteristics of concrete by a parabolic-rectangular stress–strain relation and proposed algorithms that calculate steel stress accurately using a trial-and-error method. We developed a program that automatically calculates the crack width for serviceability performance verification by a design code such as the CEB-FIP Model Code 1990, EUROCODE 2, JSCE PBD, and ACI 318 CODE 1995. We tested our program using experimental data to verify the accuracy of the results.

Keywords Crack width · Limit state design · Parabolic-rectangular stress–strain curve · Serviceability performance

90.1 Introduction

In structural design, limit states are conditions of potential failure that make a design unfeasible. A limit state is a set of performance criteria that must be met when the structure is subject to loads. Structural limit states tend to fall into three

H. Lee (✉) · D. Park
Department of Computer Science and Engineering, Soonchunhyang University, Asan, South Korea
e-mail: leehm@sch.ac.kr

D. Park
e-mail: parkds@sch.ac.kr

G. Lee
Department of Civil Engineering, Suncheon First College, Suncheon, South Korea
e-mail: gylee@suncheon.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_90, © Springer Science+Business Media Dordrecht 2014

categories: serviceability limit states, strength limit states, and ultimate limit states [1]. Serviceability limit states are those conditions that are not strength-based but still may make the structure unsuitable for its intended use. Serviceability limit states are determined by vibration, and cracking. The strength limit states are potential modes of structural failure. To satisfy the ultimate limit state, the structure must not collapse when subjected to the peak design load for which it was designed.

In general, concrete structural members are designed by verification of strength or ultimate limit states. The designs that result from this verification method use high stress-reinforced steel and concrete. High stress can lead to crack generation that harms the functionality and exterior of structures and shorten their operational lifetimes. Serviceability performance verification of concrete structural members predicts whether crack width satisfies requirements at the service load step [2]. For correct verification of the serviceability limit states, this prediction must reflect the strength of concrete and steel and the deformation characteristics. The crack width verification should be performed using the steel stress and strain and the nonlinear stress–strain relations of concrete. To calculate stresses in steel using the nonlinear material relations of concrete, we should apply a trial-and-error method that calculates repeated changes in the strain of the compressive top fiber. For this reason, existing design codes based on the limit states design method have not adopted trial-and-error methods. Instead they still verify the serviceability performances of structures using empirical formulae based on experimental results [3–6].

In this paper, we propose and implement algorithms that can calculate the stress in steel by a trial-and-error method for correct serviceability performance verification of concrete structural members. Our application automatically calculates stress and strain in steel, and crack width depending on the design codes. In addition, we added a user-defined function to reflect the new verification theory or philosophy of the designer. Our application is designed to apply various design codes or new theories for serviceability performance verification and to give more accurate verification.

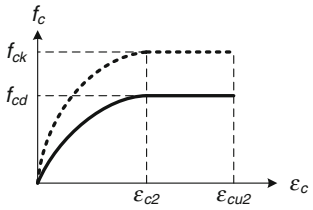
90.2 Algorithm for Steel Stress Calculation

90.2.1 Nonlinear Material Characteristics of Concrete

In Fig. 90.1, the parabolic-rectangular stress–strain curve (*p-r curve*) shows the nonlinear material relations of concrete to analyze member sections for serviceability performance verification [4].

$$0 \le \varepsilon_c \le \varepsilon_{c2},\ f_c = f_{cd}\left[1 - \left(1 - \frac{\varepsilon_c}{\varepsilon_{c2}}\right)^n\right] \tag{90.1}$$

$$\varepsilon_{c2} \le \varepsilon_c \le \varepsilon_{cu2},\ f_c = f_{cd} \tag{90.2}$$



f_{ck} (MPa)	≤ 40	50	50	50	50
ε_{c2} (10^{-3})	2.0	2.2	2.3	2.4	2.5
ε_{cu2} (10^{-3})	3.0	2.9	2.8	2.7	2.6
n	2.0	1.8	1.6	1.45	1.4

Fig. 90.1 Parabolic-rectangular stress–strain relationships

Here, f_c is the concrete compressive stress; ε_{c2} is the strain reaching the maximum strength according to Fig. 90.1; ε_{cu2} is the ultimate strain according to Fig. 90.1; n is the exponent according to Fig. 90.1; f_{cd} is the design compressive strength, defined as $\gamma_c(0.85f_{ck})$; γ_c is the partial safety factor for concrete; and f_{ck} is the concrete compressive strength.

90.2.2 Computation of Steel Stress

In previous studies, the stress–strain relation of concrete has been clearly defined. For this reason, steel stress was calculated empirically based on many experiments to figure out the compressive resultant force and point of action. However, the limit states design method allows a definite compressive stress–strain relationship for concrete. We can calculate the compressive resultant force of concrete and point of action as

$$C = \int_0^c f_c(\varepsilon) b dx \tag{90.3}$$

$$\beta c = c - \frac{\int_0^c x f_c(\varepsilon) b dx}{\int_0^c f_c(\varepsilon) b dx}, \tag{90.4}$$

where b is the width of the compression zone, c is the neutral axis depth, and x denotes the neutral axis depth (Fig. 90.2).

However, as we mentioned in the previous section, we must use a trial-and-error method to calculate the actual steel stress. The weakness of this method is that it must calculate the resultant force every time. Because of this, it is necessary to use several kinds of dimensionless coefficients to ease the calculations. The average stress of the compression area $f_{c,ave}$ around the size of the resultant compression α is defined as the design strength f_{cd} ratio. β is defined as the ratio of the distance between the compression podium and the point of action and the depth of neutral axis. When all of these parameters are determined to be

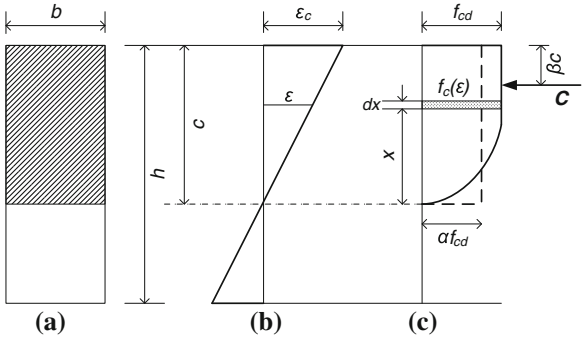


Fig. 90.2 Section with compressive strain and stress diagram using *p-r* curves. **a** Section. **b** Strain. **c** Stress

dimensionless coefficients, the calculation becomes easy. In this study, α and β were defined as functions of the compression podium strain ε_c as

$$0 \le \varepsilon_c \le \varepsilon_{co},\ \alpha = 0.5\varepsilon_c - \frac{\varepsilon_c^2}{12},\ \beta = \frac{8 - \varepsilon_c}{4(6 - \varepsilon_c)} \tag{90.5}$$

$$\varepsilon_{co} \le \varepsilon_c \le \varepsilon_{cu},\ \alpha = 1 - \frac{2}{3\varepsilon_c},\ \beta = \frac{4\varepsilon_c - 3(\varepsilon_c)^2 - 2}{6\varepsilon_c(0.67 - \varepsilon_c)} \tag{90.6}$$

where ε_c is the actual maximum concrete strain in the extreme fiber of the concrete in compression and in ‰.

Figure 90.3 illustrates the rectangular concrete structural members of width b and effective depth d that have a form section that is strengthened by a tension reinforcing bars A_s receiving a bending moment M. In early stages when the section is not cracked, we can predict the behavior of reinforced concrete using the theory of linear elasticity. If a sealed concrete platform reaches bending strength f_r, flexure cracks appear. At this point, the flexural moment is put into the cracking moment M_{cr}.

After flexure cracks appear, we can ignore the tension strength of the concrete tension area by assuming that the strain of the section is arranged linearly in a triangle as shown in Fig. 90.3b by the plane preservation assumption. Applying the *p-r* curve relationships in the compression zone from Eq. (90.1), the stress distribution appears as shown in Fig. 90.3c.

However, after flexure cracks appear in the beam, a resultant force occurs on the seal reinforcing rod A_s and compression area by flexural moment M as

$$C = \alpha f_{cd} bc, \tag{90.7}$$

$$T = \gamma_s A_s f_s = \gamma_s \rho f_s bd, \tag{90.8}$$

where γ_s is the partial safety factor for steel, f_s is the steel stress (N/mm^2), ρ is the steel ratio $(= A_s/bd)$, and d is the effective depth.

These two resultant forces are equal from the laws of equilibrium and the assumption that plane cross sections remain plane. Also, the arm length z, being

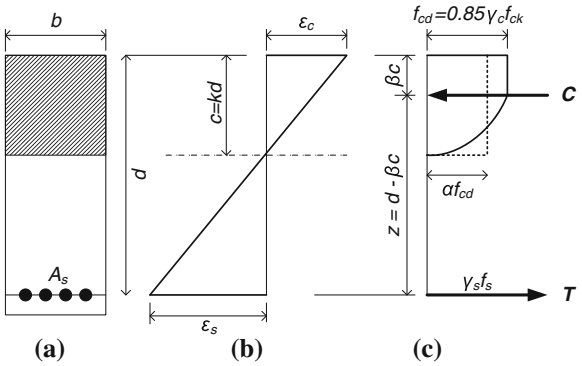


Fig. 90.3 Strain and stress distribution in the ultimate limit state for a single reinforced rectangular beam. **a** Section. **b** Strain. **c** Stress

defined the distance of the couple of the force C and T, $z = d - \beta c = (1 - \beta k)d$, where k is the ratio of the neutral axis depth $(= c/d)$.

The moment of resistance is therefore

$$M = Cz = \alpha f_{cd} k(1 - \beta k)bd^2 \tag{90.9}$$

$$M = Tz = \gamma_s f_s A_s(1 - \beta k)d. \tag{90.10}$$

Section analysis requires three dimensionless values: α, β, and k. Until these coefficients satisfy all equilibrium conditions of section and transformation condition, the trial-and-error method must be used. Therefore, the dimensionless coefficient m, which expresses the relative size of the bending moment that acts on a section

$$m = \frac{M}{f_{cd} bd^2}, \tag{90.11}$$

is used.

If we readjust by substituting this coefficient into the bending moment Eq. (90.9), we get the quadratic equation $m = \alpha k(1 - \beta k)$, where k is a variable. The value that satisfies this equation is

$$k = \frac{1 - \sqrt{1 - 4m\beta/a}}{2\beta}. \tag{90.12}$$

Supposing a priority compression platform strain ε_c to calculate the steel stress, the values of α and β can be calculated using Eqs. (90.5) and (90.6). If these values are substituted for Eqs. (90.11) and (90.12), we get a quadratic equation where k represents a variable. After the substituted ε_c and calculated k are adapted to the geometry condition of the section strain distribution of Fig. 90.3b, the reinforcing rod stress f_s is given as

$$f_s = E_s \varepsilon_c \left(\frac{1 - k}{k} \right) \leq \gamma_s f_y, \tag{90.13}$$

where E_s is the modulus of elasticity for steel and f_y is the yielding stress of steel.

If the steel stress calculated by Eq. (90.10) is equal to the value that calculated the stress from Eq. (90.10), the steel stress that calculate a supposed strain ε_c is calculated by applied flexural moment, as follow

$$f_s = \frac{M}{\gamma_s A_s (1 - \beta k) d}. \tag{90.14}$$

However, if the reinforcing bar stress is unequal, the top fiber strain is changed continuously and the trial-and-error method is applied until the values match.

90.2.3 Crack Width Formulation

Based on the limit state design method for crack width to verify the serviceability performance design code provisions, the CEB-FIP Model Code 1990 (i.e., MC 90) [3], EUROCODE 2 (i.e., EC 2) [4], and the Structural Performance Verification of Japan Society of Civil Engineering (i.e., JSCE PBD) were developed [5]. As specified in ACI 318 CODE-1995 (i.e., ACI 318-95) [6], until recently many researchers calculated crack width using the empirical formula proposed by Gergly-Lutz [7].

The basis of the crack width calculation to design codes is presented here. First, the simplified case of a reinforced concrete direct tension member is considered, as in Fig. 90.4a. The member will first crack when the tensile strength of the weakest section in reached. Cracking leads to a local redistribution of stresses adjacent to the first crack, as indicated in Fig. 90.4b. At the crack, the entire tensile force is carried by the reinforcement. With increasing tension, a second crack forms at the next weakest section. This does not occur within a distance l_t of the first crack due to the reduction in concrete stresses caused by bond stress in the region, as shown in Fig. 90.4c. With further increase in tension, more cracks develop until the maximum crack spacing anywhere is $2l_t$. At this point, no more cracks then form, but further loading causes the existing cracks to widen, as shown in Fig. 90.4d. That is called stabilized cracking. The crack width formulations are based on the above discussions. The crack width is defined from the difference in deformation of the concrete and steel over a length equal to the crack spacing. This basic concept for crack width formulation has been proposed and modified in each design code provision.

90.2.3.1 CEB-FIP Model Code 1990

The crack width formulate in MC 90 stems from the difference in extension of concrete and steel over a length equal to the cracking spacing. The crack width is thus given as

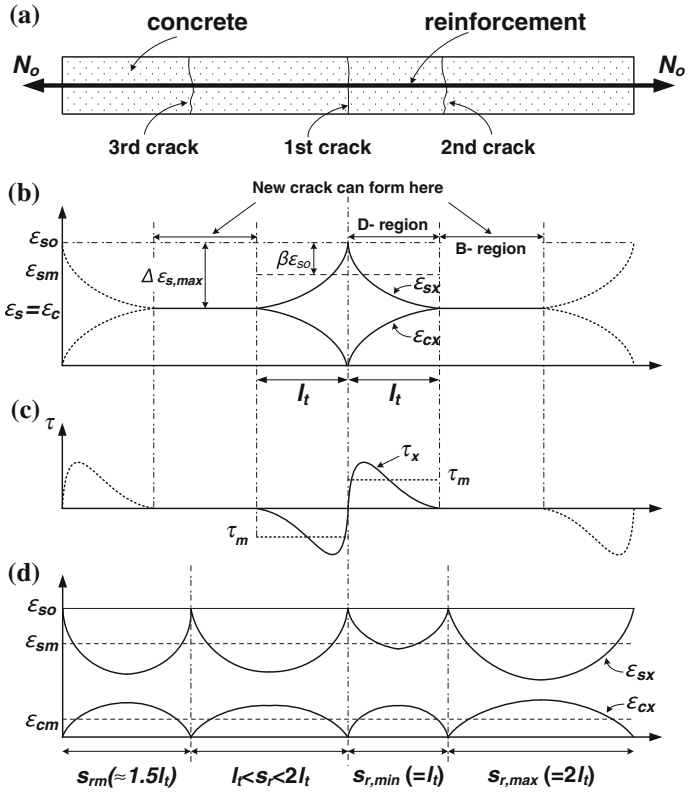


Fig. 90.4 Strain and bond stress distribution of tension members: **a** direct tensile member, **b** strain distribution at first cracking, **c** bond stress, **d** strain distribution at stabilized crack

$$w_{MC90} = \frac{d_b}{3.6\rho_{ef}}\left[\frac{f_s}{E_s} - \beta_t \frac{(1+n\rho_{ef})f_{ct}}{E_s\rho_{ef}}\right], \tag{90.15}$$

where d_b is the bar diameter (mm), ρ_{ef} is the effective steel ratio $(= A_s/A_{c,ef})$, $A_{c,ef}$ is the effective tension area of concrete surrounding the tension reinforcement of depth $h_{c,ef}$, and $h_{c,ef}$ is the lesser of $2.5(h-d)$, $(h-x)/3$ or $h/2$. Here, f_s is the stress in tension reinforcement assuming a cracked section, n is the modulus ratio E_s/E_c, β_t is an empirical factor to assess the average strain within a transmission length (0.6/0.4 for short/long-term loading), and f_{ct} is the tensile strength of concrete (N/mm^2).

90.2.3.2 Eurocode 2

In EC 2, the crack width formulation resembles Eq. (90.10) of MC 90, but the maximum crack space is

$$w_{EC2} = \left(3.4c + 0.425k_1k_2\frac{d_b}{\rho_{ef}}\right)\left[\frac{f_s}{E_s} - 0.4\frac{(1+n\rho_{ef})f_{ct}}{E_s\rho_{ef}}\right], \tag{90.16}$$

where c is the cover to the longitudinal reinforcement, k_1 is a coefficient that takes account of the bond properties of the bonded reinforcement (0.8 for high bond bars and 1.6 for bars with an effectively plain surface), and k_2 is a coefficient that takes account of the distribution of strain (0.5 for bending and 1.0 for pure tension).

90.2.3.3 Structural Performance Verification of the Japan Society of Civil Engineering

In JSCE PBD, the crack spacing including a cover thickness, the center-to-center distance between the reinforcing bars, and the bar diameter are different from MC 90 and EC 2. The crack width is calculated as

$$w_{JSCE-PBD} = 1.1k_1\frac{15}{f_{cd}+20}\frac{5(n_l+2)}{7n_l+8} + 0.7[4c + 0.7(c_s - d_b)]\left(\frac{f_s}{E_s}\right), \tag{90.17}$$

where k_1 is a constant that accounts for the effect of surface geometry reinforcement on crack width. It may be taken as 1.0 for deformed bars and 1.3 for plain bars. n_l is the number of layers of tensile reinforcement, c is the concrete cover (mm), c_s is the center-to-center distance of tensile reinforcements (mm), and d_b is the diameter of tensile reinforcement (mm).

90.2.3.4 ACI 318 Code 1995

In ACI 318–95, statistical analysis of experimental data following Gergly and Lutz [7] were used to calculate the crack width

$$w_{ACI95} = 1.08\beta_c f_s \sqrt[3]{d_c A} \times 10^{-5}, \tag{90.18}$$

where β_c is the ratio of distances from the tension face and from the steel centroid to the neutral axis, equal to $(h-kd)/(d-kd)$, d_c is the thickness of concrete cover measured from the tension face to the center of the bar closest to that face, and A is the concrete area surrounding one bar, equal to the total effective tensional area of concrete surrounding the reinforcement with the same centroid divided by the number of bars $(= (b \times 2d_c)/n_b)$.

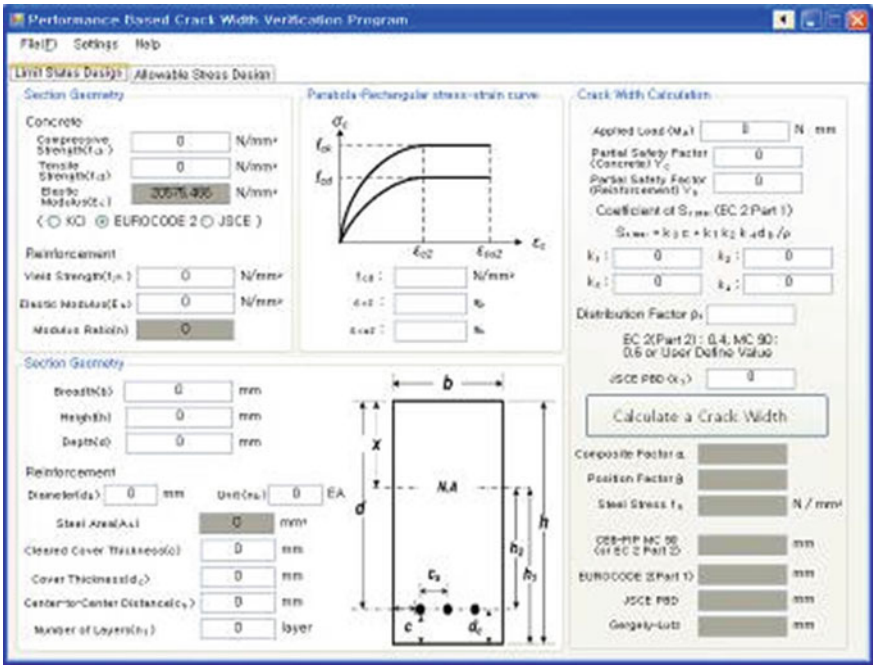


Fig. 90.5 Crack width calculation program

90.3 Implementation and Experiments

We developed a serviceability performance verification application using Visual C#.Net. Our application is compatible with Windows operating systems Windows 98, Windows 2000, Windows XP, Windows Vista, and Windows 7.

The calculation of the crack width based on the limit states design method consists of five steps: inputs of material characteristics, inputs of concrete p-r curve properties, input of section dimension and steel placement characteristics, input of applied load, materials coefficient and crack spacing coefficient, and calculation of crack width.

Figure 90.5 shows our performance-based crack width verification program. First, the user inputs the material characteristics of concrete and steel to calculate crack width. Next, the user refers to Fig. 90.1 and inputs the base value of the *p-r* curve in the nonlinear stress–strain relation of concrete. Then the user inputs the section dimension and steel placement for calculating crack width by each design code. Finally, if the user inputs the applied load, material coefficients, coefficients for calculating crack spacing, and strain distribution coefficient, s/he can check the calculation results automatically.

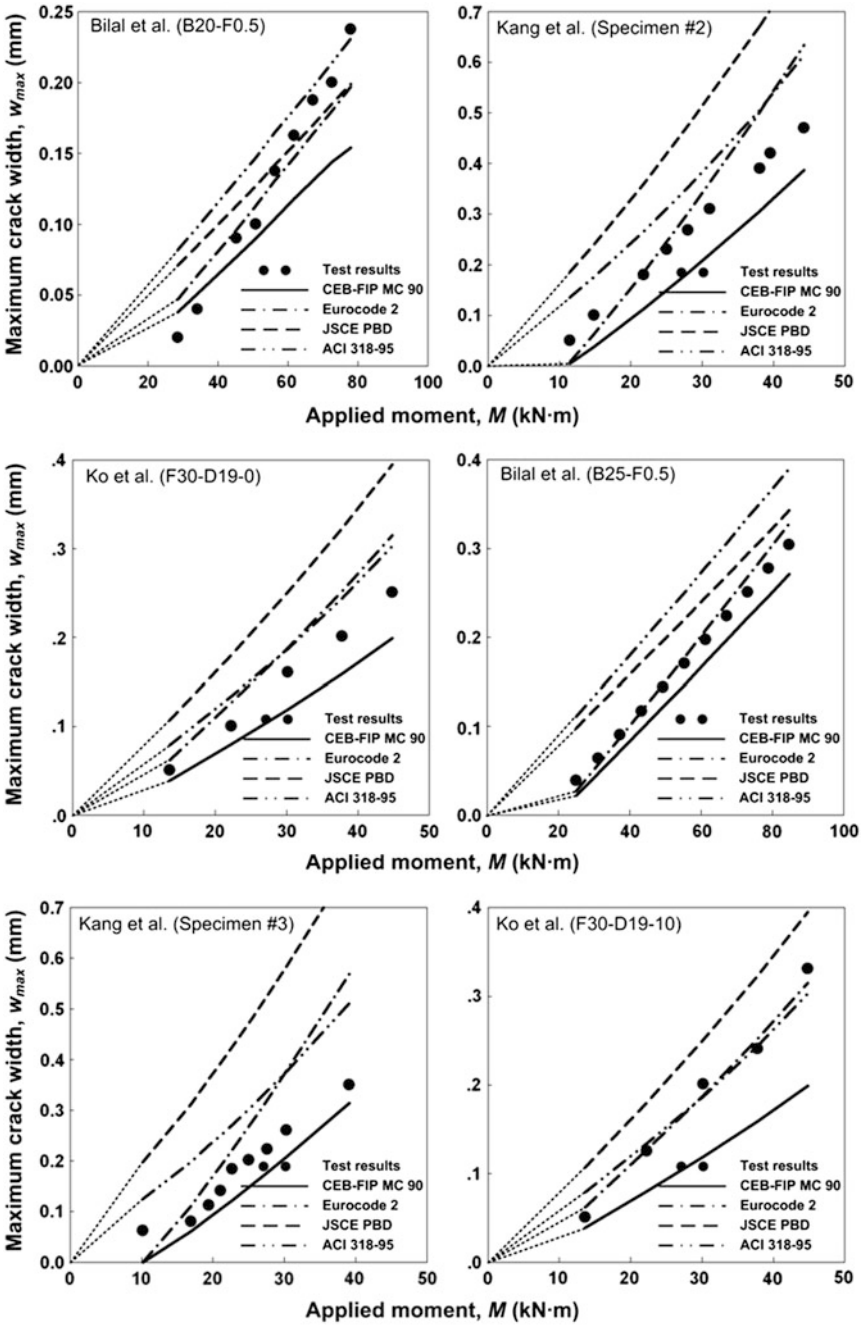


Fig. 90.6 Comparison of crack width calculations versus test results of our program

We tested our program using six datasets (Bilal et al. [8], Kang et al. [9], Ko et al. [10]) to verify the accuracy of the crack width calculation. Figure 90.6 compares the crack width calculations versus the test results of our program.

As shown in Fig. 90.6, we confirmed that the crack width calculation results of our program for serviceability performance verification were accurate. In the test results, small differences in crack width were caused by each design code. In all of the tests, EC 2 was the design code that could calculate crack width most accurately, while MC 90 underestimated crack widths. In contrast, JSCE PBD verified serviceability performance with heavy reliance on crack width. Because this produced incorrect results, we suggest that JSCE PBD should be applied with caution. Like JSCE PBD, ACI 95 had wide variation in verification results according to requirements.

90.4 Conclusion

We developed a program that calculates crack width for serviceability performance verification of structural concrete members based on limit state design. The correct calculation of steel stress is important in serviceability performance verification. Therefore, we defined the material characteristics of concrete by parabolic-rectangular stress–strain relations and a proposed algorithm that calculates steel stress accurately using a trial-and-error method. Our program automatically calculates crack width by a design code based on the limited states design method and supports verification and comparison of serviceability performance. Using previously published datasets, we confirmed that the results that our system produces for serviceability performance verification are accurate.

Acknowledgments This work was supported by National Research Foundation of Korea Grant funded by the Korean Government (2010-0022773).

The English in this document has been checked by at least two professional editors, both native speakers of English. For a certificate, please see: http://www.textcheck.com/certificate/xIZPFN.

References

1. Fib (1999) Structural Concrete, vol 2. International Federation for Structural Concrete, pp 75–91
2. CEB (1997) Serviceability models-behavior and modeling in serviceability limit states including repeated and sustained load. Comite Euro-International Du Beton, pp 12–18
3. CEB-FIP (1991) CEB-FIP model code 1990. Comite Euro-International Du Beton, pp 87–109
4. European Committee for Standardization (2002) Eurocode 2-design of concrete structures. European Committee for Standardization, pp 130–136
5. Concrete Committee of JSCE (2003) Standard specifications for concrete structures-2002. Japan Society of Civil Engineers, pp 107–130

6. ACI Committee 318 (1995) Building code requirements for structural concrete and commentary (ACI 318-95). American Concrete Institute, pp 135–137
7. Gergely P, Lutz LA (1968) Maximum crack width in reinforced concrete flexural members, causes, mechanism and control of cracking in concrete. ACI Special Publication SP-20, pp 87–117
8. Bilal SH, Mohamad HH (2001) Effect of fiber reinforcement on bond strength of tension lap splices in high-strength concrete. ACI Struct J, pp 638–647
9. Kang YJ, Oh BH (1985) Determination of crack width and crack spacing in reinforced concrete flexural members. KSCE J, pp 103–111
10. Ko WJ, Park SG (2002) Estimation of flexural crack width considering bonding characteristics in concrete members. KSCE J, pp 825–835

Chapter 91
Novel Real-Time Automobile Detection Algorithm for Blind Spot Area

Seung-Hun Yang, Gwang-Soo Hong, Beak Ryong and Byung-Gyu Kim

Abstract Many facets of automobile technology are moving to information technology (IT)-based convergence systems. Sensors and equipment used to prevent accidents that occur due to drive negligence are under development. An automobile detection algorithm for blind spot areas while driving is proposed. Detection of candidate vehicles is based on vehicle features, and then a vehicle is subject to verification of the detected candidate vehicle area. Using a Haar-like feature for vehicle detection, an adaptive vehicle verification process based on brightness values of vehicle areas is used. The hierarchical scheme of the proposed algorithm is efficient and accurate, based on detection and verification results in the blind spot area.

Keywords Computer vision · Intelligent driver assistance system · BSD (Blind spot detection) · Feature extraction

91.1 Introduction

The traffic accident rate in France, Germany, and the UK (among the OECD members) has decreased steadily since the 1980s. However, in Korea, the accident rate greatly increased before the early 1990s, then decreased after 1996. The rate in

S.-H. Yang · G.-S. Hong · B. Ryong · B.-G. Kim (✉)
Dept. Computer Engineering, Sun Moon University, A-san, South Korea
e-mail: bg.kim@mpcl.sunmoon.ac.kr

S.-H. Yang
e-mail: smhotkdg@mpcl.sunmoon.ac.kr

G.-S. Hong
e-mail: hongzolv@mpcl.sunmoon.ac.kr

B. Ryong
e-mail: padigen@mpcl.sunmoon.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_91, © Springer Science+Business Media Dordrecht 2014

2008 was reported to be 12.1 accidents per 100,000 vehicles on the road [1]. Approximately 9 % of all vehicle accidents are caused by blind spots. A great deal of research regarding automobile detection on the road has been carried out in the past 10 years [2, 3]. For now commercialized safety and convenience systems, radar and single cameras are usually used. If only radar is used, the error detection rate for obstacles and otherwise undetected hazards increases in many real situations. To compensate for this, reliable technical development and study are needed using image-based obstacle recognition and convergence [4]. An image-based automobile detection system could provide processed information in a way that is a more intuitive for drivers [5, 6].

The biggest problem of image-based automobile detection is a quickly changing environment. It is important to define candidate vehicles in a manner that is insensitive to changes in the environment. Vehicle features used for detection of candidate vehicles are based on symmetry, color, shadow, geometrical characteristics, texture, and headlights [3].

The overall system flow of the proposed detection system is shown in Fig. 91.1. The area of a candidate vehicle is identified and verified using vehicle data that is trained in advance. The proposed automobile identification algorithm is powerful for use in any environment based on improving candidate vehicle identification via advance training.

The Sect. 91.1 of this paper presents Data Training and Detection, and Verification. In Sect. 91.2, related research is introduced, and in Sect. 91.3, the proposed vehicle detection method is defined. In Sect. 91.4, simulation results are discussed. Finally, concluding remarks will be given in Sect. 91.5.

91.2 Related Research

Related research and a definition of blind spot areas are presented. In Sect. 91.2.1, vehicle detection and verification are explained, and in Sect. 91.2.2, research that is required for data training is discussed.

91.2.1 Blind Spot Areas

The position of blind spots is shown in Fig. 91.2. A blind spot means a place that disappears from a driver's view when side and rear vehicles are in specific locations. To detect a vehicle in a blind spot, recognition and verification should be based on the approach of vehicles from a distance to close-in positions.

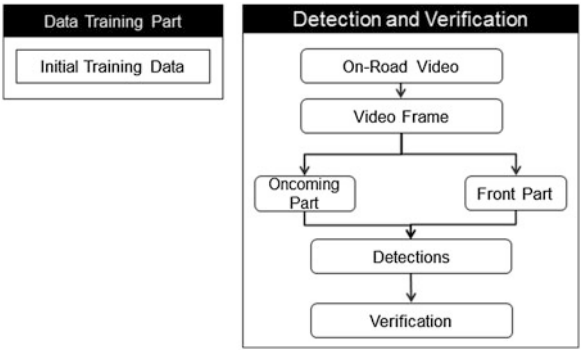


Fig. 91.1 Overall flow chart

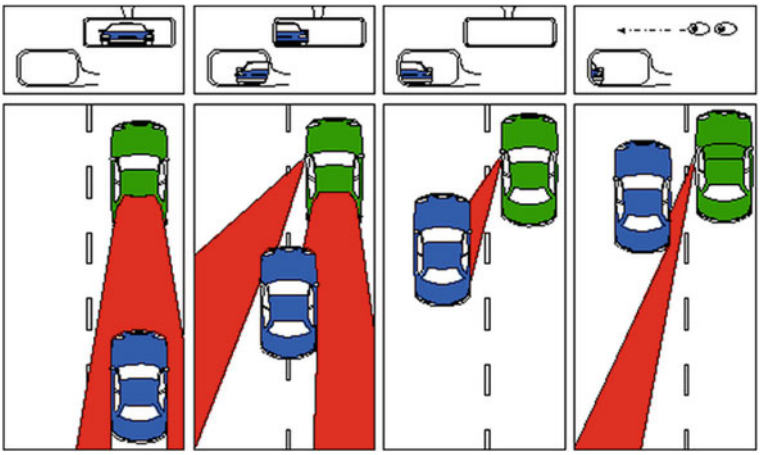


Fig. 91.2 Blind spot areas

91.2.2 Data Training

Related work on feature extraction and verification techniques is reported. Related research on data training is shown in Table 91.1. Generally, data training is used to verify vehicles after detection of candidate vehicle areas. The feature extraction technique used in this work is based on Haar-Waveletes [7, 8], and support vector machine (SVM) is widely used as a classification technique [7–9].

Table 91.1 Vehicle study research

Research	Feature extraction	Study and classification	Goal
Kapoor et al. [17]	SIFT + PCA	SVM	Various objects
Enzweiler and Gavrila [9]	Haar waveletes	SVM	Person
Roth and Bischof [7]	Haar waveletes	Online boosting	Face
Vijayanarasimhan and Grauman [8]	Local features	SVM	Various objects

Table 91.2 Related research for detection and tracking schemes

Research	Candidate vehicle detection step	Trace vehicle
Wu et al. [6]	• Lane information base • Road and vehicle shadow area segmentation	–
Lin et al. [16]	• Lane information base • Symmetry use of vehicle based edge feature	• Use motion vector estimation
Sun et al. [10]	• Lane information base • Edge information use	• Use detect and track, not general detect then track
Balcones et al. [11]	• Symmetry use of vehicle based edge feature	
Alonso et al. [12]	• Edge image detection for definition of interested area • Image division using split and merge	
Song et al. [13]	• Boundary detection for right and left side of vehicle • Detection of the contact area of tire and road	
Han et al. [14]	• Normalization perform of video brightness value • Investigation of shadow area • Vertical edge detection by haar – like feature	

91.2.3 Detection and Verification

Related research on detection and tracking techniques is presented. Related research on detection and tracking schemes is shown in Table 91.2. Generally, there are two detection methods. The first is based on use of lane information and the second is based on no use of lane information. Vehicle edge information is a vehicle feature that is generally used for identification [10–14]. It is difficult to use edge information in the detection step for candidate vehicles because edge information changes according to the conditions of illumination and weather in driving environments, even though edge information is distinctive [3]. Therefore, a technique that is applicable and adaptable to various environments is needed.

91.3 A Proposed Automobile Detection Algorithm

91.3.1 Data Training Process

In the vehicle feature extraction and training stage, initial vehicle training is applied. Then, features of the forward and side views of a vehicle after detection are trained before the detection operation is reapplied.

A far vehicle presents a different appearance from a close-in vehicle Fig. 91.3. For the close-in vehicle, the front and side are shown at the same time. However, for a far vehicle, only the front is shown. At first, both images are used simultaneously.

The Haar-like feature that is used in initial data training is shown in Fig. 91.4. In the initial image data training stage, 528 images of the vehicle and 1,240 images of non-vehicles with the Haar-like feature are used. The process of classification for vehicles and non-vehicles using the extracted Haar-like feature is shown in Fig. 91.5. In the step of the designed cascade scheme, the number of candidate vehicles is decreased. The number is finally determined in the final step. If a candidate image area remains, the image is classified as a vehicle. The feature value of Eq. (91.1), is used for fast calculation.

$$S_{acc}(i,j) = \sum_{x=0}^{i} \sum_{y=0}^{j} I(i,j) \tag{91.1}$$

The $S_{acc}(i,j)$ is the calculated summation of the corresponding area, i is the pixel value of the x axis, j is the pixel value of the y axis, and $I(i,j)$ are pixel values of real vehicle image at position (i,j) [15].

Detection results for candidate vehicle areas are shown in Fig. 91.6. Vehicle detection results for a clear day (Fig. 91.6a) and a dark environment (Fig. 91.6b) can be compared. Results for a rainy day can be seen in Fig. 91.6c. Noise is present in all detection results.

91.3.2 Verification Process

Candidate vehicle areas are detected using the proposed feature training scheme. Since noisy parts exist, a detected candidate area should be verified. The experimental environment used was a CPU of 3.4 Hz, a memory (RAM) of 8.00 GB, and the Windows 7 OS. The resolution was 128 × 128 using 528 vehicle training images, and 124 training images without vehicles.

A detected candidate vehicle requires verification because of noise.

Images of a detected vehicle and a non-vehicle are shown in Fig. 91.7. A large difference in the components of shadow areas between the vehicle and the non-vehicle is apparent. The detection result for a candidate vehicle involving little

Vehicle Feature	Example Image	
Adjacent Oncoming	The Front face and side face of a vehicle traveling roughly perpendicular to the ego vehicle.	
Oncoming	The Front face of a vehicle traveling parallel to the ego vehicle, in the opposite direction.	

Fig. 91.3 Image according the vehicle distance

Fig. 91.4 Haar-like feature

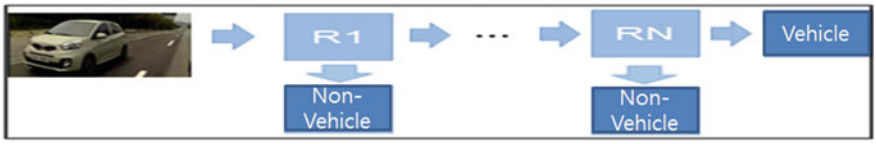


Fig. 91.5 A cascade-type classifier

Fig. 91.6 Results of vehicle detection using the proposed feature training. **a** Clear day; **b** dark environment; **c** rainy environment

noise is shown in Fig. 91.7b. Extraction results for both a real vehicle and the area of vehicle shadow are shown in Fig. 91.7b and c. Detection results for a vehicle area on a rainy day are shown in Fig. 91.7d, e.

An image prior to processing for classification of vehicle or non-vehicle in a candidate area is shown in Fig. 91.8. Squares indicate a candidate vehicle change in size with vehicle position. In the far case, the square size is detected as small. In the close-in case, the square size is detected as large. When the vehicle area and the non-vehicle area image attributes are measured, the brightness component of an exterior area is proportional to the detected square area. The maximum size of the detection area is 200 × 200 pixels and the minimum size is 100 × 100 pixels.

Equation (Formula 91.2) is used to define the area size based on detection of a bright area around the candidate vehicle area.

$$\text{bwidth} = \frac{dIwidth}{100}, \text{bheight} = \frac{dIheight}{100} \qquad \text{(Formula 91.2)}$$

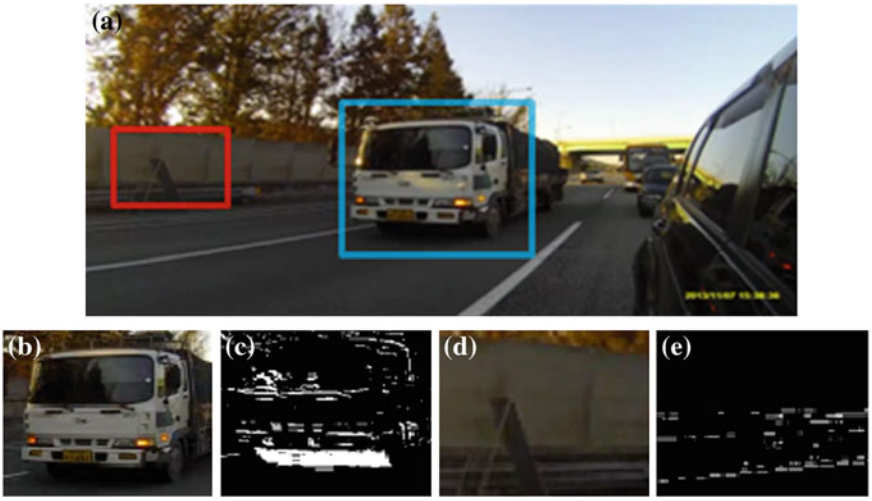


Fig. 91.7 The vehicle and non-vehicle decision based on the shadow area extraction method

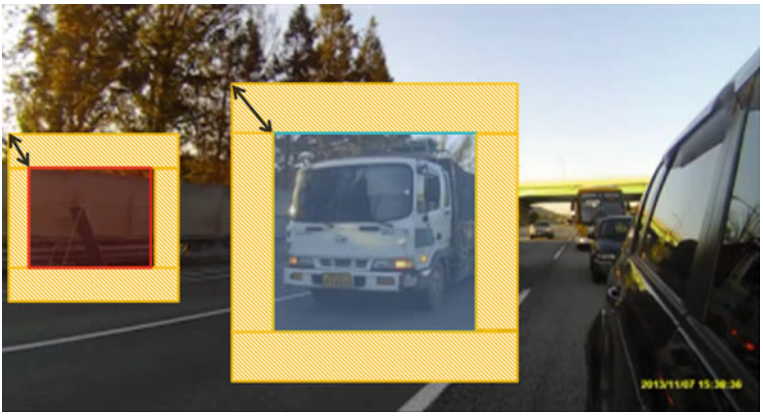


Fig. 91.8 Pre-processing image classification between a vehicle and a non-vehicle

Where *dIwidth* and *dIheight* are the width and length area of real candidate vehicles, are bwidth and bheight of the area for brightness detection. To verify candidate areas, a histogram of information from the luminance component is used. The RGB color model is converted to the YCbCr color model using the following equation:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.144 \\ -0.16874 & -0.3313 & 0.500 \\ 0.500 & -0.4187 & -0.0813 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 0 \\ 128 \\ 128 \end{bmatrix} \tag{91.3}$$

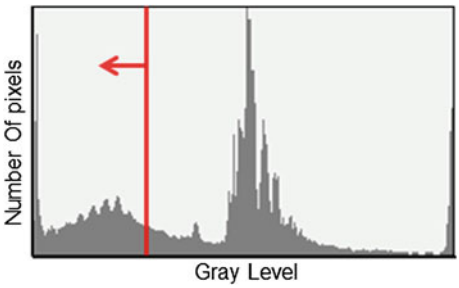


Fig. 91.9 Shadow area detection using adaptive thresholding

Since Y in the YCbCr color model indicates the brightness value, Cb and Cr contain information obtained via pixel sub-sampling of the color-difference signal. Using Eq. (91.3), the brightness value of the surrounding environment can be determined.

Using this process, the brightness value measured for each candidate vehicle area is used to compare shadow component results for candidate vehicles using Eq. (91.4):

$$\text{Threshold} = (\text{Y}, \text{d}) \tag{91.4}$$

Equation (91.4) returns threshold value information that is used to verify candidate vehicles. Y is the surrounding brightness component, and d is size information for a candidate vehicle area. The current surrounding environment can be represented by brightness information around a vehicle, and the distance between vehicles can be determined using the size of a candidate vehicle area. Thus, the recognition ratio can be increased after verification of candidate vehicles using an adaptive thresholding scheme.

Figure 91.9 shows a histogram to extract shadow area in verifying candidate vehicles. Threshold value setup of Histogram uses average of changed brightness values through formula (91.3). Vehicle's shadow area effective in night and day can be explored using adaptive threshold value according to surrounding brightness.

91.4 Simulation Results and Evaluation

Different environments using test video sequences are shown in Fig. 91.10. A general expressway environment is shown in Fig. 91.10a–c while a general city street environment is shown in Fig. 91.10b and dark environments are shown in Fig. 91.10e and f.

Results for recognition using each test video sequence are shown in Table 91.3. True Positives indicate correct detection of a real vehicle. False Positives indicate

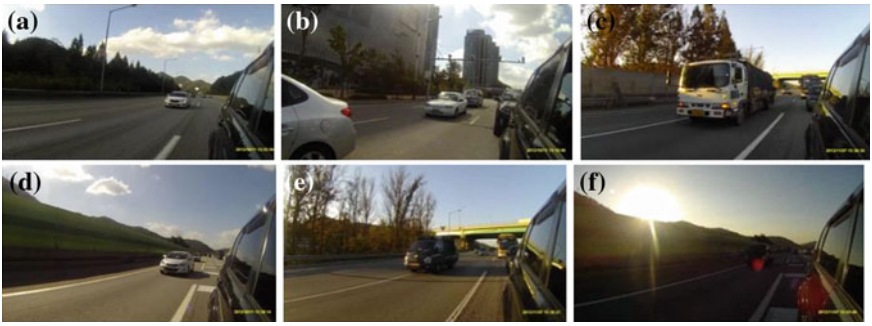
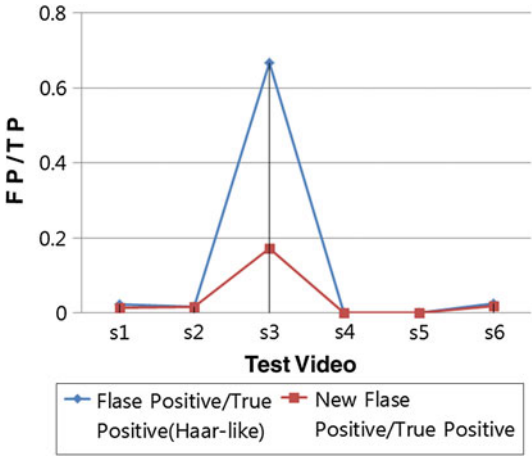


Fig. 91.10 Environments in each test video

Table 91.3 Recognition results in each test video

Test video	Frame	Real vehicle	True positive	False positive	After verification, false positive	Recognition rate (%)
Video 1	300	220	206	5	3	93.63
Video 2	300	300	298	5	3	99.33
Video 3	300	90	81	54	14	90
Video 4	300	54	50	0	0	92.59
Video 5	300	29	27	0	0	93.10
Video 6	300	162	156	4	3	96.29

Fig. 91.11 False positive/true positive



incorrect detection of a non-vehicle. The recognition ratio of true vehicles was 94.16 % indicating that the proposed algorithm works well for detection of vehicles in blind spot areas.

False detection using the proposed algorithm is shown in Fig. 91.11. The blue line shows vehicle detection prior to verification, and the red line shows the same

result after verification. The number of false positives is reduced by the verification process. The processing time used 28 frames per second with 480 × 640 (VGA) video sequences. Thus, the proposed algorithm works in real-time.

91.5 Conclusions

An efficient vehicle detection algorithm for blind spot areas is proposed. New features were defined for identification of a vehicle candidate area for feature training, then vehicles were detected using verification of the detected candidate vehicle area. After detection of a candidate vehicle using a Haar-like feature, an adaptive vehicle verification process based on a histogram of area brightness values was used. Using real video, the accuracy and effectiveness of the proposed algorithm was verified. The technique developed in this study can be used to reduce the blind spot related traffic accident rate by delivering proper information/alarms to drivers in any situation.

References

1. http://www.internationaltransportforum.org/Pub/pdf/13KeyStat2012.pdf
2. Trivedi MM, Cheng SY (2007) Holistic sensing and active displays for intelligent driver support systems. IEEE Comput 40(5):60–68
3. Sun Z, Bebis G, Miller R (2006) On-road vehicle detection: a review. IEEE Trans Pattern Anal Mach Intell 28(5):694–711
4. Alessandretti G, Broggi A, Cerri P (2001) Vehicle and guard rail detection using radar and vision data fusion. IEEE Trans on Intell Transp Syst 8:95–105
5. Trivedi MM, Cheng S (2007) Holistic sensing and active displays for intelligent driver support systems. Computer 40(5):60–68
6. Wu Y-J et al (2007) Image processing techniques for lane-related information extraction and multi-vehicle detection in intelligent highway vehicles. Int J Automot Technol 8(4):513–520
7. Roth PM, Bischof H (2008) Active sampling via tracking In: Proceedings of IEEE conference on computer vision pattern recognition, pp 1–8
8. Vijayanarasimhan S, Grauman K (2008) Multi-level active prediction of useful image annotations for recognition, In: Proceedings of neural information processing system conference, pp 1705–1712
9. Enzweiler M, Gavrila DM (2008) A mixed generative-discriminative framework for pedestrian classification. In: Proceedings of IEEE conference on computer vision pattern recognition, pp 1–8
10. Sun Z, Bebis G, Miller R (2006) Monocular precrash vehicle detection: features and classifiers. Proceedings of IEEE Transactions on Image 15(7):2019–2034
11. Balcones D, et al (2009). Real-time vision-based vehicle detection for rear-end collision mitigation systems. Comput Aided Syst Theory-EUROCAST 2009. Springer, Heidelberg, pp 320–325
12. Alonso D, Salgado L, Nieto M (2007). Robust vehicle detection through multidimensional classification for on board video based systems In: IEEE international conference on image processing ICIP 2007, vol 4

13. Song GY, Lee KY, Lee JW, (2008) Vehicle detection by edge-based candidate generation and appearance-based classification. In: IEEE intelligent vehicles symposium
14. Han S, Han Y, Hahn H (2009) Vehicle detection method using Haar-like feature on real time system. World Acad Sci Eng Techonol 59
15. Viola P, Jones M (2001), Rapid object detection using a boosted cascade of simple features. In: Processing IEEE conference computer vision pattern recognition vol 1, pp 511–518
16. Lin CC, et al (2007) Development of a multimedia-based vehicle lane departure warning, forward collision warning and event video recorder systems. In: Multimedia workshops. ISMW'07. 9th IEEE international symposium on IEEE
17. Kapoor A, Grauman K, Urtasun R, Darrell T (2007) Active learning with Gaussian processes for object categorization, In: Proceedings of IEEE international conference on computer vision, pp 1–8

Chapter 92
Environment Control Model of Intelligent Plant Factory Using Context Based of Sensing Data

NamJin Bae, KyungHun Kwak and YongYun Cho

Abstract Recently, according to rapid development and distribution of IT technology, various researches that adopt IT technologies to all the existing technologies have been conducted. However, among current IT converged agricultural technology, system which controls environment of plant factory is subordinate to particular system and has a weakness that it is hard to handle intelligently the various exceptional situations which may occur in plant factory environment without intervention of human. Thus, this research provides environmental control service which can handle intelligently the possible exceptional situations which may occur in various situations as building ontology by defining each relation such as environmental factor and control factor necessary for cultivating crops in vertical farm.

Keywords Context-aware · Ontology · Vertical farm · Environment control

92.1 Introduction

Due to great science development, the whole world has been urbanizing and environmental pollution is in serious condition. As a result, decrease in land size for food production by urbanization and environmental pollution and problems of food shortage by increasing food demand according to population growth became

N. Bae · K. Kwak · Y. Cho (✉)
Department of Information and Communication Engineering, Sunchon National University, 255 Jungang-ro, Suncheon, Jeollanam-do, Republic of Korea
e-mail: yycho@sunchon.ac.kr

N. Bae
e-mail: bakkepo@sunchon.ac.kr

K. Kwak
e-mail: spersdar@hotmail.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_92, © Springer Science+Business Media Dordrecht 2014

international problems, also Korea is in serious situation because self-sufficiency rate of grain is lower than 30 %.

Recently, in order to solve these problems, various demo businesses and researches are tried to raise added value and productivity of agriculture applying IT technologies such as many types of sensors and wireless communication technology. Among these, service which is provided by using IT convergence technology can be divided into monitoring service and control service. These services collect data of environmental factors affecting growth and development of crops through such as temperature sensor, soil sensor and moisture sensor, and control monitoring and environmental factors for growth and development environment of crops using environment controller. Also, it increases production efficiency of agriculture as making possible for analyzing and providing growth and development environment of crops, predicting growth and development based on accumulated data [1].

However, in case of existing systems, it is mostly belonged to particular system and takes the form that users directly or remotely control and monitor or set value of growth and development environment.

This has a weakness that it is hard to handle positively the exceptional situations of system and environment where diverse variables are working. In order to solve these weaknesses, this thesis suggests environmental control service model of context awareness based on ontology which is able to handle actively the exceptional situations which is credible in real greenhouse cultivating environment without intervention of human as defining relations between each environmental factors and control factors utilizing ontology which is possible to grow continually.

Also, in order to verify whether suggesting control service model of context awareness is able to handle positively the exceptional situation generated in effective environmental control service and greenhouse cultivating environment, this thesis examined and presented its result through virtual scenario.

92.2 Related Work

92.2.1 Context-Aware Control Service

The reason why people communicate efficiently is that our conversation does not only depend on exchanging language but also uses positively the information about situations which surrounds talkers (physical, emotional and historical situation). There can be serious distortion if people underutilize these informations or ignore them. Thus, people act and speak considering law, regulation and manner consciously or unconsciously and these words and action can be examples of using context.

Context-aware in computing environment means controlling environment according to situation which system is changing, personal taste and intention by identifying and understanding users and environment that surrounds users [2]. In

other words, it changes according to environment by itself and offers user-centered service. Also, same service is proved and adopted as a different form of optimal service.

For example, considering daily life of people, they perform various roles depending on who they meet or what kind of situation they are in, so they need to receive different services according to those roles. And types of service can be changed according to various surrounding environment although they play same roles.

Main subjects of a research to perform context-aware service can be defined as following; First, context modeling for how should we express and store situation information to make machine understand it; Second, context deduction about how understand context based on informations from physical environment; Third, interoperability for sharing context knowledge between information equipment which are independent and dispersed [3].

As context modeling methods, there are Key-Value Models, Markup Scheme Models, Graphical Models, Object Oriented Models and Logic Based Models, and based these context models, research for context-aware service in various area based on these context models has been actively conducting.

92.2.2 Service Based on Ontology

Ontology defines not only standard terms which describe the knowledge to specify concept and knowledge of certain domain and but also relation between terms, in addition, it includes inference rules that is able to expand those. One of the roles of domain is solving problems when different databases use different terms about same concept, and this became an important factor which allows reuse and sharing of knowledge between web based knowledge process and application programs [4].

By conceptualizing situation data about specific situations and expressing, deducing the relation, research about services based on ontology in various areas such as optimal path searching, medical service and knowledge map service has been conducting [5–7].

92.3 Context-Aware Control Service for Vertical Farm Environment

92.3.1 Structure of Context-Aware Control Service

In vertical farm consisted of various sensor networks, sensor data is collected and managed from middleware of vertical farm.

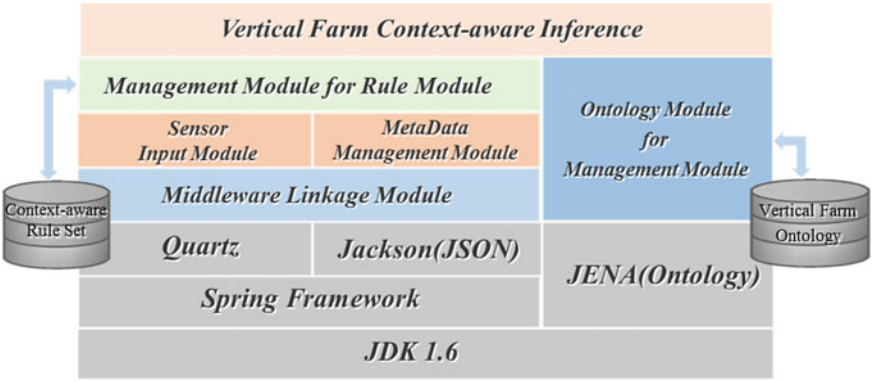


Fig. 92.1 Control service architecture based on context-aware

Table 92.1 System component

Component	Explain
Vertical farm context-aware inference	Perform deduction through Jena ontology framework
Management module for rule module	Manage rules which are used for deduction based on ontology
Sensor input module	Input sensing data to ontology model
	Apply sensing data from middleware to ontology model
Meta data management module	Manage sensor network metadata
	Apply it to ontology as required change of metadata from middleware
Middleware Linkage module	Perform system interface between middleware↔ontology systems
Ontology model management module	Load ontology file and manage raw model and inference model
Quartz	Scheduling framework is used for deducting periodically
Jackson	Manage file list in JSON form
	JSON framework is used for object serialization of files
Spring framework	Lightweight application framework manages bins which is component of system
JENA	Ontology framework is a ontology framework supporting base deduction

Ontology deduction system is included to same container with middleware of vertical farm, reference between interface modules is set up through DI (Dependency Injection) of Spring, and it performs interfaces such as metadata management, sensing data input, deduction rule management and inference result notice.

Table 92.2 Crop information

Crop	Temp	Hum	pH	Control priority		Nutrient
				Morning	Afternoon	
Tomato	15–22 °C	55–68 %	5.8–6.0	1. Temp 2. Hum	1. Hum 2. Temp	Within 4 h after sunrise

Structure of components and framework which consist sensor abnormality sensing system suggested in this thesis are same with Fig. 92.1, and in Table 92.1, it showed components of system structure map.

92.3.2 Collecting and Categorizing Information of Vertical Farm's Controlled Environment Data

Generally, every plant needs optimal growth environment according to growth stage and in order to control optimal growth environment in vertical farm environment for each crops, control factor to control growth data of crops, growth environment is needed and also definition of the relation between those are necessary. Thus, in this article, basic growth environment of 'Tomato' which is an object crop to verify through ontology deduction and control object to control environment are defined.

Followings are informations from Rural Development Administration about growth environment of tomatoes [8].

- Optimal growth temperature is lower than 22 °C during daytime and higher than 15 °C during night time.
- Evening time affects most to growth speed in growth and development.
- Temperature management is needed before noon and humidity management is needed after noon.
- So excessive ventilation during sunset has bad effect to growth of crops and weakens rapidly its power, it needs to reduce stress of crops with moderate ventilation.
- Proper moisture for farmland is between 55–68 %.
- Nutrient solution should be produced within 4 h after sunrise.
- PH level rises during growth period so PH level of irrigation should be between 5.2–5.5 and PH level of farmland should be between 5.8–6.0.

To apply growth environment of tomato to ontology model like above, we defined as Table 92.2.

Table 92.3 shows control object to create optimal growth environment in vertical farm environment.

Table 92.3 Control target

Ventilation control	These are for ventilation control inside of vertical farm and automatic control objects are such as skylight, side window and extractor fan
Temperature control	These are for temperature control inside of vertical farm and automatic control objects are such as celling, side windows, extractor fan, hot air heater, hot-water boiler and Fan and Pad
Humidity control	These are for humidity control inside of vertical farm and automatic control objects are such as side windows, extractor fan, dehumidifier and humidifier
Irrigation control	These are for humidity and moisture/PH/EC level control inside of vertical farm and automatic control objects are such as drip irrigation and nutrient solution supplying device
CO_2 control	These are for adjusting CO_2 inside of vertical farm and automatic control objects are such as CO_2 control system, side windows and extractor fan
Light control	These are for environment control for quantity of light inside of vertical farm and automatic control objects are blackout curtain and artificial light control device

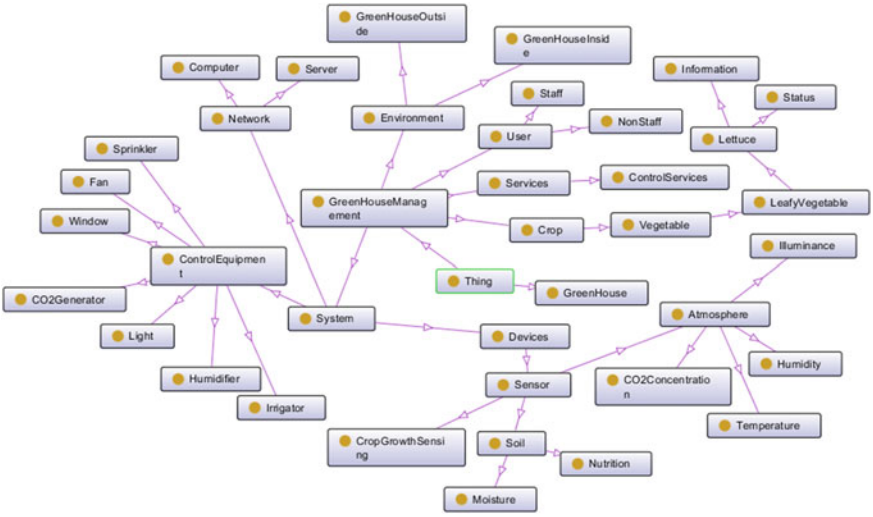


Fig. 92.2 Ontology model for provide context-aware control service

92.3.3 Ontology Model for Automatic Growth Environment Control of Crops

Service model suggested in this thesis uses ontology model to control automatically the growth environment of crops in greenhouse. In other words, suggesting ontology model is used as a communal data to minimize interference of human for

Fig. 92.3 Ontology model for auto control

context-aware control service in greenhouse environment and to collaborate naturally between computing resources. Figure 92.2 is the ontology model to provide context-aware control service suggested in this thesis.

As you can see in Fig. 92.3, the top class of ontology model for automatic control in vertical farm environment is "UrbanizedVerticalFarm" and it was classified into "CropGrowthKnowledge", "VerticalFarm" and "VerticalFarmEnvironment".

VerticalFarmEnvironment class is a class that defines environmental factors which can affect to growth environment of plants inside and outside of vertical farm, and it also defines necessary facilities and systems which are needed for environment control.

CropGrowthKnowledge is the class that defines knowledge about growth of crops in vertical farm and it includes optimal growth environment and disease and insect pest.

92.3.4 Control Service Process

Figure 92.4 shows control service process suggested in this thesis. Spring container generates Bin which take in charge of interface from middleware to ontology deduction system and it prepares that interface is available through method call each other by performing DI between mutual interface. And then, if middleware receives sensing value, it delivers those values to ontology deduction system and sends appropriate services which is proper to the situation according to the rule of deduction engine through middleware.

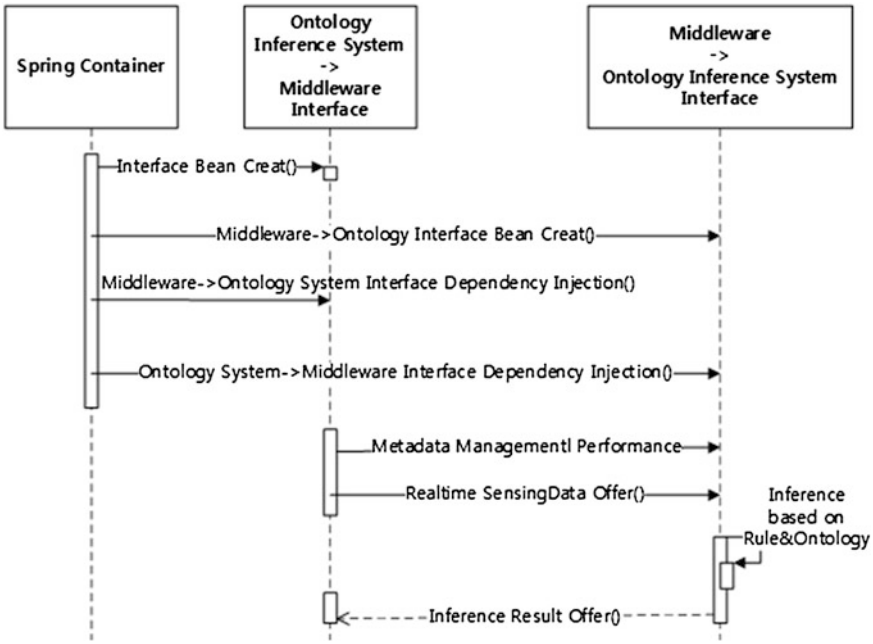


Fig. 92.4 Control service process

92.3.5 Inference Rule

In order to provide intelligent control to vertical farm, rules that is able to expand the relation and class between each object of ontology model should be needed. This thesis mentions only part of deduction rules to check abnormal condition of sensors among deduction rules as following;

- Assume the abnormal condition of sensors by checking differences between sensors which are installed in same area.
- Regard that sensor node which isn't reported the sensing value for certain period has an obstacle.
- Check the range of observed value and regard that if it is out of range, it is in abnormal condition.

Figure 92.5 shows part of sources which realized deduction rules like above by using RDF language.

```
@prefix pfssn: <http://etri.re.kr/NET/pfssn#>.
@prefix ssn: <http://purl.oclc.org/NET/ssnx/ssn#>.
.
.
[pfssnrule1: (?node1 rdf:type pfssn:PFTempSensor) .
        (?node2 rdf:type pfssn:PFTempSensor).
                (?node1 ssn:onPlatform ?platform).
                (?node2 ssn:onPlatform ?platform).
                (?observation1 ssn:observedBy ?node1).
                (?observation2 ssn:observedBy ?node2).
                (?observation1 ssn:observationResult ?output1).
                (?observation2 ssn:observationResult ?output2).
                (?output1 ssn:hasValue ?value1).
                (?output2 ssn:hasValue ?value2).
                (?value1 pfssn:hasQuantityValue ?float1).
                (?value2 pfssn:hasQuantityValue ?float2).
                difference(?float1, ?float2, ?diff).
                greaterThan(?diff, 5) .
                -> (?node1 pfssn:status '1'^^xsd:integer)   (?node2 pfssn:status '1'^^xsd:integer).
].
```

Fig. 92.5 Sensor's abnormal status inference based on RDF

92.4 Conclusion

This thesis suggested Context-aware control service model based on ontology to make up for weaknesses which is hard to handle positively without intervention of human for dependency and particular situation of existing systems that simply set and control value of optimal growth environment in order to provide optimal environment for growth of crops in greenhouse.

Suggested Context-aware control service is easy to expand and correct as it defines factors considered to control greenhouse environment based on ontology, and it also can handle actively in customized prescription and unexpected situation according to automatic control and condition of crops and system without intervention of human.

Professional knowledge is needed for crop cultivation, in the future, acceleration of context-aware control system is needed to make possible for unified administration across overall crops management through applying rules which is equipped with regular diversity based on professional knowledge and mapping of ontology in environment area.

Acknowledgments This research was supported by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the CITRC (Convergence Information Technology Research Center) support program (NIPA-2013-H0401-13-2008) supervised by the NIPA (National IT Industry Promotion Agency).

References

1. Korea Food Security Research Foundation-Causes and solutions of the global food. http://www.foodsecurity.or.kr/bbs/view.php?bbs_id=qnaa07&doc_num=57
2. Kim S-H (2011) A study of ontology-based context modeling in the area of u-convention. Korea Inf Processing Soc 28(3):123–139
3. Chen H, Finin T, Joshi A (2004) Semantic web in the context broker architecture. IEEE international conference on PERCOM04, Washington, IDEEE, Computer Society, pp 277–286
4. Mcllraith S (2002) Semantic enabled Web services. XML-Web services ONE conference, vol 6
5. Ma YB, Kim JK, Lee JS (2012) Modeling and simulation of ontology-based path finding in war-game simulation. Korea Soc Simul J 21(1):9–17
6. Rural Development Administration—Agricultural Technology Comprehensive Information. http://www.rda.go.kr/board/board.do?mode=html&prgId=arg_cropfarmskillEntry#ac_con_view
7. Lee B-M, Kim J, Kim JH, Lee Y, Kang U-G (2011) A customized exercise service model based on the context-awareness in u-Health service. J Korea Inf Technol 9(2):141–152
8. Kim J, Park C, Jung J, Lee H, Jung H, Min K, Kim H (2003) The design and implementation of knowledge map service system based on ontology. Korea Inf Sci Soc 30(1):527–529 Spring Workshop Presentation File
9. Strange T, Linhoff-Popien C (2004) A context modeling survey. The sixth international conference on ubiquitous computing, UbiComp

Chapter 93
Implementation of an Integrated Log Analysis System Through Statistics-Based Prediction Techniques

Kwangman Ko

Abstract Integrated log analysis systems, which could collect, store and analyze a large volume of log and big data in real time by analyzing firewall logs, continue to expand their applications to a variety of fields such as abnormal network behavior detection, use pattern analysis with web log analysis, fraudulent order analysis and detection for internet shopping malls, inside information leakage analysis and detection. This paper presents a result of designing and implementing an prediction engine applying statistics-based log analysis(regression analysis, time-series analysis, cluster analysis and discriminant analysis etc.) technologies, which could overcome problems of trying to implement with GNUR, mathematical and statistical libraries, for finding preemptive action through concentrated guard during an expected security accident time period by analyzing and predicting security-related infra logs.

Keywords Integrated log analysis · Statistical prediction analysis · Security · Big-data analysis

93.1 Introduction

As the computer communication environment is advanced at a rapid pace and its performance is improved, a diversity of security-related issues has been presented as a serious problem. In particular, integrated log analysis systems, which could

K. Ko (✉)
School of Computer Information Engineering, Sangji University, Wonju, South Korea
e-mail: kkman@sangji.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_93, © Springer Science+Business Media Dordrecht 2014

collect, store and analyze a large volume of log and big data in real time by analyzing firewall logs, continue to expand their applications to a variety of fields such as abnormal network behavior detection, use pattern analysis with Web log analysis, fraudulent order analysis and detection for Internet shopping malls, inside information leakage analysis and detection [1, 2].

The log analysis products are divided into domestic simple security management ones [3], foreign-made log analysis ones and the existing ESM solution ones to supply at home and abroad, and it is expected that the ESM, which could not store raw logs, would be losing its existence value in the market as the laws and regulations requiring to store raw logs are increased [4].

In addition, it is considered that domestic simple security management products, which could store raw logs but their correlation analysis functions are insufficient, could not also get out of low-price centered markets. Domestic log-related market is divided into high-priced integrated log analysis one and low-priced simple log management one, for the former, domestic and foreign-made products are competing, and for the latter, domestic companies are also facing a competitive situation.

Anymon Plus [3], an integrated log analysis product, overwhelms its competitors' products in PoC thanks to the world's top real-time big data collection, storage and analysis performance (processing 40,000 events per second), and is emerging as the best product in the market. In addition, Anymon Plus is an ESM product storing raw logs, and is growing to a product that could conquer both log analysis and ESM market at the same time because it provides dynamic analysis functions as excellent as foreign-made products.

This paper presents a result of designing and implementing an prediction engine applying statistics-based log analysis (regression analysis, time-series analysis, cluster analysis and discriminant analysis [5] etc.) technologies, which could overcome problems of trying to implement [6] with GNU R [7, 8], mathematical and statistical libraries, for finding preemptive action through concentrated guard during an expected security accident time period by analyzing and predicting security-related infra logs. This could help actively use to develop solutions which could preemptively respond to security accidents and threats and be introduced general-purposely by small and medium sized businesses.

This paper is organized as follows. Section 93.2 introduces the design of a system model forming the basis of a statistics-based integrated log analysis system to be developed in this paper and a statistics-based prediction technique to be implemented. Section 93.3 describes how to implement detailed component modules' specific functions and core prediction algorithms of the integrated log system, and then presents the result of an experiment. Section 93.4 draws a conclusion and explains about the future study.

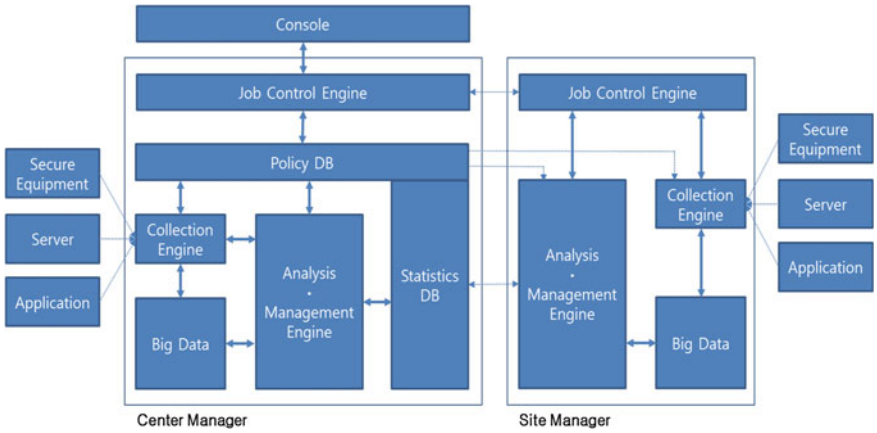


Fig. 93.1 General integrate-log analysis system

93.2 Based Works

93.2.1 Integrated Log Analysis System

The existing typical integrated log analysis system is mainly composed of a center manager and a site manager as Fig. 93.1.

The collection engines of center and site managers gather various logs from security equipment, servers and applications that are clients' logs to be collected. For this case, agents are loaded, and the site manager has a structure capable of distributed collection, storage and analysis because it could be expanded in parallel unlimitedly.

93.2.2 Design of the Integrated Log Analysis System

The statistics-based integrated log analysis system, which is ultimately designed and implemented in this study, has a structure as Fig. 93.2, in which how to develop statistical prediction engines is practically classified into four main methods.

The first method to develop the statistical prediction engine is an attempt to connect or share interfaces with the existing products to interoperate with GNU R, which has an advantage of using a statistically and completely proven tool, but involves inefficiency of real-time processing due to a problem of interworking and operation speed. In addition, it excessively occupies system resources such as CPU and main memory, and there are insufficient interfaces for source data and output, so it involves difficulties of development.

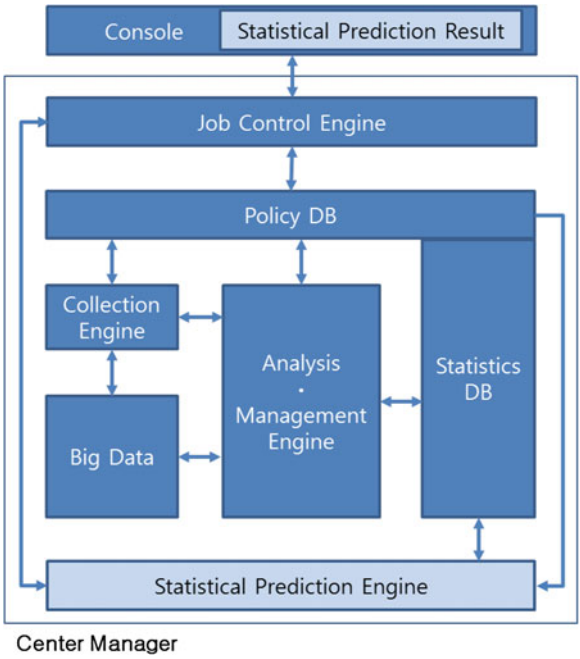


Fig. 93.2 Overview of statistic-based prediction engine system

The second method is an attempt using mathematical and statistical libraries, which has an advantage of generality in addition to developing the prediction engine.

However, it should accept an expensive price policy for each server, and a difficulty of implementing additional logic is presented as a disadvantage. The third method is an attempt to purely develop a new statistical prediction engine, which could secure all the sources related to the implementation and it has an advantage of optimized modification considering the characteristics such as operating systems and development languages as necessary. However, there is a disadvantage of requiring considerable resources and time needed for development. Finally, an attempt to develop it based on a mathematical standard library could minimize verification by using a standardized library and the standard library could be substituted for parallel process etc. In addition, it has a feature that could be concentrated in developing the prediction engine to minimize the development period. However, it should be solved the consideration for the license issue that may be encountered when selecting an open-source library and the problem of expense that may be arisen when using additional libraries.

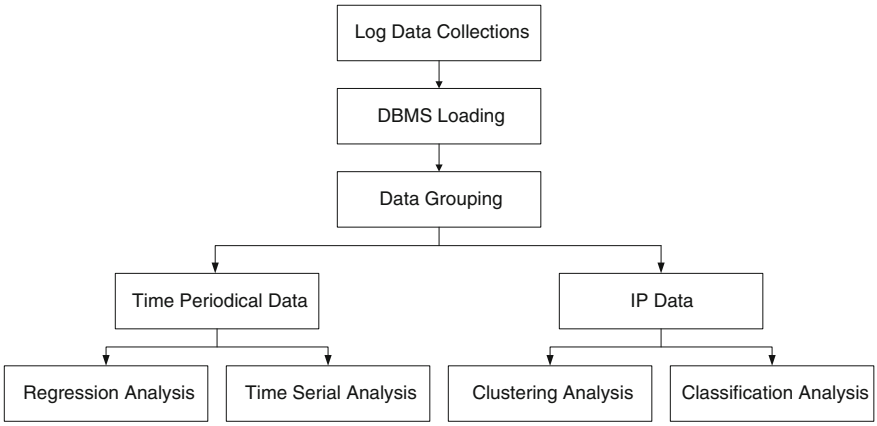


Fig. 93.3 Flow of statistic-based prediction model generation

93.3 Statistics-Based Prediction Engine System

93.3.1 Generation of the Prediction Model Based on Statistics

In order to periodically analyze a large volume of the collected log data to generate a model for analyzing and predicting abnormal data, it was applied the analysis flow as Fig. 93.3.

For the basic log data, it was applied the whole data in which the level that could determine changes of data by the period is at least three weeks, and it was targeted at the log data for each specific IP and the specific time zone logs in order to decide whether or not to be abnormal data.

93.3.2 Integrated Log Analysis System Modules

This paper developed a mathematical standard library, in which modules could be self-replaced gradually based on open-source libraries, to design the statistics-based prediction engine as Fig. 93.4 and to implement respective modules.

- Data import module: Import data with xls, csv and text formats into the internal memory
- Data editing module: Remove nulls of the data imported, convert characters into numbers, and set additional variables
- Basic stat module: Compute basic statistics for the data processed (average, maximum, minimum, standard deviation and variance etc.)

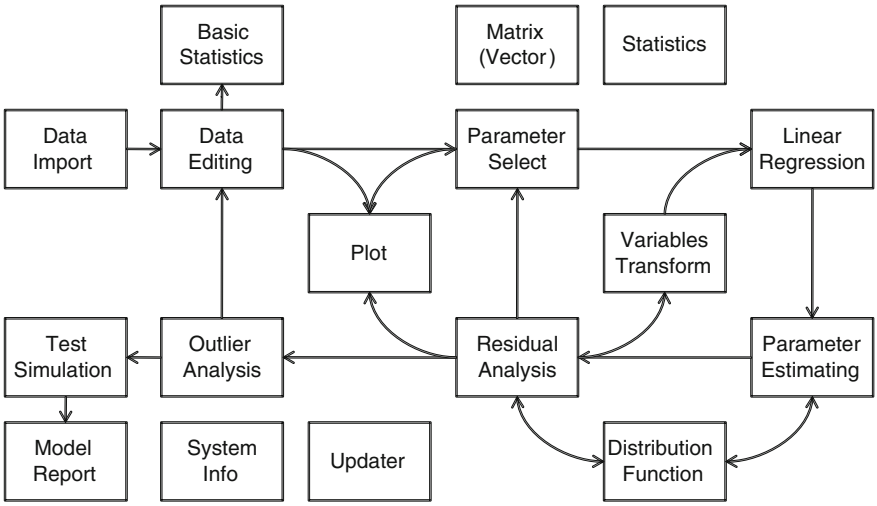


Fig. 93.4 Flow of statistic-based prediction model generation

- Matrix module: Compute inverse matrix required in the regression analysis process
- Statistics module: Compute basic statistics (use in parameter estimating)
- Parameter select module: Select variables to be statistically analyzed
- Linear regression module: Compute linear regression analysis
- Parameter estimating module: Compute statistics such as ANOVA and R^2 to evaluate the regression analysis results
- Variables transform module: Transform linear variables into various functions (x − > 1/x, log x, x ^ n etc.)
- Residual analysis module: Understand whether there is cross correlation (normal, equal variance or independence) or not between the parameters obtained
- Plot module: Output graphs for various statistics
- Distribution function module: Compute distribution functions needed when finding ANOVA etc.
- Outlier analysis module: Extract/remove outlier data through various algorithms
- Test simulation module: Use some of input data as verification data to test whether or not to predict well
- Model report module: Output statistical analysis results as a format such as xls, doc, html or pdf
- System info module: Understand a system specification to compute matrix more efficiently
- Updater module: An automatic update module of program itself.

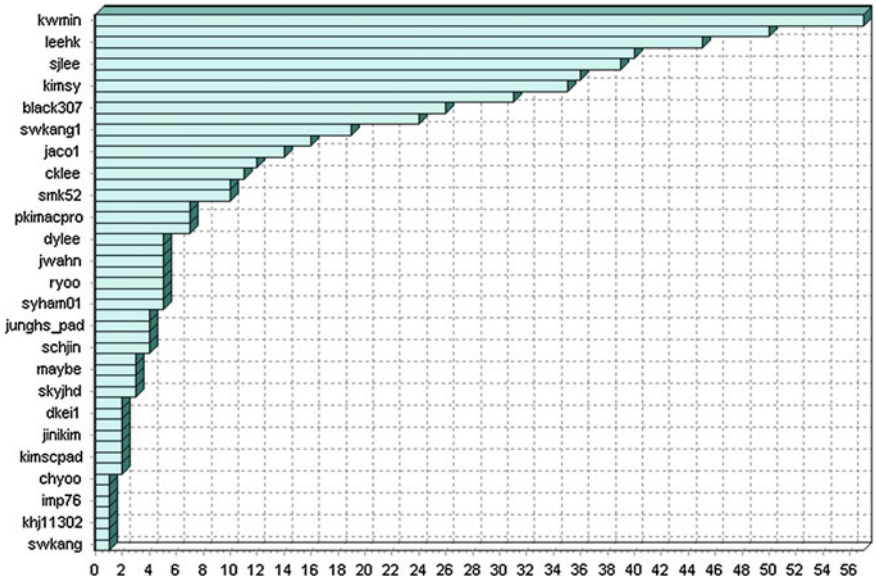


Fig. 93.5 Example of simulated log enevts

93.3.3 Experimental Results and Analysis

In order to verify execution of the integrated log analysis system implemented in this paper and its prediction model, it was examined through an actual field application and simulation. Figure 93.5 is an example of average log event inputs of applying the integrated log analysis system developed in this study to the real log data of simulated Authentication sever to measure real and expected log events 2013-12-01 to 2013-12-31. And, Fig. 93.6 shows a result of prediction with regression analysis and Fig. 93.7 represents a result of prediction with time-series analysis.

In addition, it uses the regression analysis information to provide the supporting data for future prediction and budget investment, the prediction information through data and graphs collected for each equipment or period, the information that could be used as base data for measures of equipment extension and load balancing, and the threshold reaching time based on the result of threshold setup and regression analysis, so it has an effect that could reasonably adjust the time to introduce the budget.

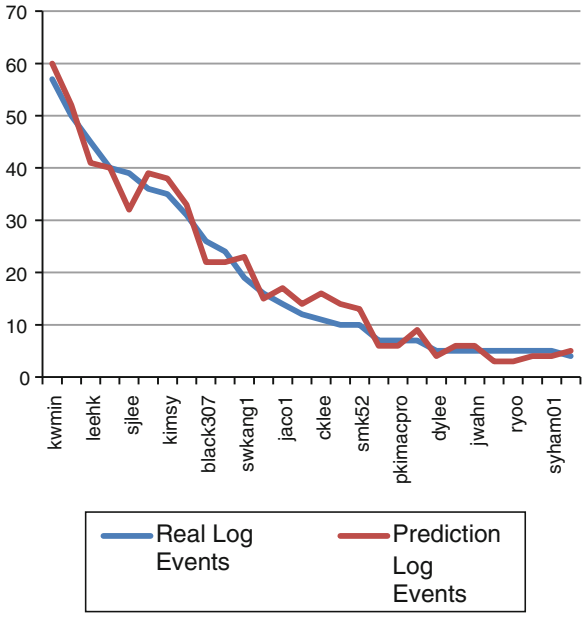


Fig. 93.6 Results of prediction with regression analysis

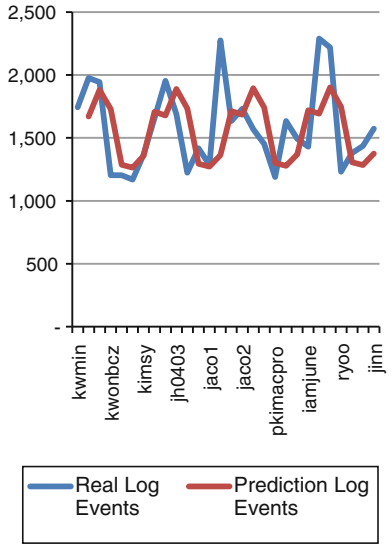


Fig. 93.7 Results of prediction with time-series analysis

93.4 Conclusion

The integrated log analysis system, which could collect, store and analyze a large volume of log and big data in real time by analyzing firewall logs, continue to expand their applications to a variety of fields such as abnormal network behavior detection, use pattern analysis with Web log analysis, fraudulent order analysis and detection for Internet shopping malls, inside information leakage analysis and detection. Domestic log-related market is divided into high-priced integrated log analysis one and low-priced simple log management one, for the former, domestic and foreign-made products are competing, and for the latter, domestic companies are also facing a competitive situation.

In order to analyze and predict security-related infra logs to find preemptive action through concentrated guard during an expected security accident time period, it was preceded the attempt to implement with GNU R, mathematical and statistical libraries. The statistics-based integrated log analysis system implemented in this paper uses time-series analysis information to strengthen the real-time monitoring function and to provide time-series prediction information for real-time CPU/memory/disk usage, and uses the regression analysis information to provide prediction information through data and graphs collected for each equipment or period to give information that could be used as base data for measures of equipment extension and load balancing. Currently, in order to improve the real-time performance of prediction engine and the accuracy of prediction, a study to apply various statistics-based prediction techniques is in progress, and a study is planned to increase reliability through feedback for applications to various fields.

References

1. Chen W-Y, Wang J (2009) Building a cloud computing system for intrusion dection system. CLOUD SLAM
2. Zikopoulos P, Eaton C (2011) Understanding big data: analytics for enterprise class hadoop and streaming data. McGraw-Hill, Osborne Media
3. UNET System, http://www.unet.kr
4. Lee J (2013) Design and analysis of efficient log unification schemes for mobile database systems. Korea University
5. Lewicki P, Hill T Statistics methods and applications. http://www.statsoft.com/v12
6. Lee H (2010) Design and implementation of web attack detection system based on integrated web audit data. Journal of KIISC 11(6)
7. Paul Teeter, R Cookbook, Insight Press, 2012
8. GNU R Progect, http://www.r-project.org/

Chapter 94
The Comparison of Pricing Schemes for Cloud Services

Se-Hak Chun, Byong-Sam Choi, Young-Woong Ko and Soon-Hak Hwang

Abstract This paper investigates subscription pricing, pay-per-use pricing, and two-part tariff from perspectives of a provider's profit, consumer surplus, and social welfare, respectively. In particular, this paper extends previous studies on pricing schemes to two-part tariff in addition to subscription pricing and pay-per-use pricing. This paper shows that two-part tariff is more efficient than the other two pricing from a provider's, consumers' and social point of view.

Keywords Cloud computing · Service models · Pricing scheme · Two-part tariff · Subscription pricing · Pay-per-use pricing

S.-H. Chun
Department of Business Administration, School of Business and Technology, Seoul National University of Science and Technology, 172 Gongreung 2-dong, Nowon-gu, Seoul 139-743, South Korea
e-mail: shchun@seoultech.ac.kr

B.-S. Choi
Industry and Strategy Department, Samsung Economic Research Institute, 1321-15, Seocho 2-dong, Seocho-gu, Seoul 137-955, South Korea
e-mail: byongsam.choi@samsung.com

Y.-W. Ko
Department of Computer Engineering, Hallym University, 1 Hallym Daehak Gil, Chuncheon, Kangwon-Do, South Korea
e-mail: yuko@hallym.ac.kr

S.-H. Hwang (✉)
Department of Liberal Arts, School of Humanity and Social Science, Seoul National University of Science and Technology, 172 Gongreung 2-dong, Nowon-gu, Seoul 139-743, South Korea
e-mail: riccardo@seoultech.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_94, © Springer Science+Business Media Dordrecht 2014

94.1 Introduction

Cloud service can be regarded as services in utility computing, where computing services are delivered in a manner similar to traditional utilities such as water, electricity, gas, and telephony [1]. So, as in other utility services, pricing schemes from a provider's standpoint plays a pivotal role in market expansion [2, 3]. So far, studies on the managerial aspect of cloud computing is far from sufficient [4]. Thus, right pricing schemes for cloud services is a critical challenge because user demand, provider's profit, and social welfare vary with the service models and pricing schemes [2, 3].

This paper investigates subscription pricing, pay-per-use pricing, and two-part tariff from perspectives of a provider's profit, consumer surplus, and social welfare, respectively. In particular, this paper extends previous studies of Chun and Choi [4] and Song [5] and to two-part tariff price scheme. The result shows that two-part tariff is more efficient than the other two pricing models such as subscription and pay-per-use pricing from a provider's, consumers' and social point of view.

This paper is organized as follows: Sect. 94.2 discusses service models and price schemes in cloud computing. Section 94.3 investigates right pricing schemes for cloud services using a theoretical analysis and discusses implications. Section 94.4 presents conclusions.

94.2 Cloud Service Pricing

As more providers are starting to offer cloud services, how to determine the right price for users is now becoming increasingly critical for cloud providers. This is because pricing is able to regulate the supply and demand of computing services and thus affects both providers (who supply the services) and users (who demand the services) respectively [2].

There are many different potential pricing policies available to cloud content providers. In subscription pricing, users pay on a recurring basis to access software as an online service [2]. The customer subscribes to use a preselected combination of service units for a fixed price and a longer time frame, usually monthly or yearly [6]. In pay-per-use pricing, users only have to pay for what they use. While pay-per-use pricing could achieve the social efficiency, it is not easily acceptable for users who want to control their budgets [3].

94.3 The Model

This paper investigates subscription pricing, pay-per-use pricing, and two-part tariff from perspectives of a provider's profit, consumer surplus, and social welfare, respectively.

Our study differentiates itself from the previous ones as follows: First, we analyze two-part tariff in addition to subscription pricing and pay-per-use pricing. Second, we consider the cost structure of a provider as Chun and Choi [5]s model. The marginal cost of providing cloud services might affect the choice of pricing schemes by a provider.

For simplicity, we assume that there is a monopoly provider of cloud service. The cost of providing the service has both fixed and variable cost elements and is expressed as $cq + f$, where c and f are the marginal cost and fixed, respectively, and q is the quantity of service consumption. We can assume that $f = 0$ without loss of generality.

The provider chooses the pricing scheme $T(q) = A + pq$, where A is the lump-sum price and p is the unit price. The pricing scheme is subscription pricing when $p = 0$ and $A > 0$, pay-per-use pricing when $p > 0$ and $A = 0$, and two-part tariff when $p > 0$ and $A > 0$. The provider profit can be expressed as

$$\pi = (AN + pQ) - cQ,$$

where N is the number of consumers purchasing service and Q is the total quantity of service consumption.

Let us turn to the demand side. We assume that Consumer u has the net utility or surplus $S_u(q) = V_u(q) - T(q)$ when she purchases q unit of service and 0 does not buy at all, where $V_u(q)$ is the total value and $T(q)$ is the payment. The parameter u represents the type of a consumer. We assume that u is uniformly distributed along [0, U] with unit height, i.e., u ∈ [0, U]. So, the total number of consumers is U.

For simplicity of analysis, we assume that $V_u(q)$ has the specific form

$$V_u(q) = \left\{u^2 - (u - q)^2\right\}/2.$$

where $V_u(0) = 0$, $V_u'(q) > 0$, and $V_u''(q) < 0$. By maximizing her utility, Consumer u shows a downward sloping linear demand.

$$\begin{aligned} &\max_q S_u = \max_q \left[\left\{u^2 - (u - q)^2\right\}/2 - (A + pq)\right], \text{ and} \\ &q(p) = u - p \end{aligned} \tag{94.1}$$

where u is interpreted as the maximum willingness to pay for the service. Note that the consumption of service occurs only when $u \geq p$.

By inserting (94.1) into the expression of $S_u(q)$, we have the surplus of Consumer u.

$$S_u(p, A) = (u - p)^2/2 - A \tag{94.2}$$

94.3.1 Subscription Pricing

Under subscription pricing, price does not depend on the amount of consumption, so $p = 0$ and $A > 0$. From (94.1) and (94.2), it follows that

$$q = u \quad \text{and} \quad S_u = u^2/2 - A.$$

A consumer purchases cloud service when her surplus is nonnegative.

$$u^2/2 - A \geq 0, \text{ or } \quad u \geq (2A)^{1/2}.$$

Since u is uniformly distributed along [0, U], the number of consumers who purchase service is

$$N = U - (2A)^{1/2}.$$

The total quantity of service can be obtained by summing up the quantity of service consumption across all the consumers.

$$Q = \int_{(2A)^{1/2}}^{U} u du = U^2/2 - A$$

and the profit of the provider is given as follows.

$$\begin{aligned} \pi &= AN - cQ \\ &= A\left\{U - (2A)^{1/2}\right\} - c(U^2/2 - A). \end{aligned}$$

From the first order condition ($d\pi/dA = 0$), we have the optimal subscription price as follows.

$$A^S = 2(U + c)^2/9. \tag{94.3}$$

Note that superscript S is used to denote 'subscription pricing'. The marginal consumer who is indifferent from buying and not buying is

$$u^S = (2A^S)^{1/2} = 2(U + c)/3.$$

Using (94.3), the number of consumers, the quantity of service use, the profit of the provider, and the consumer surplus at the equilibrium can be easily calculated.

$$N^S = (U - 2c)/3,$$
$$Q^S = (5U + 2c)(U - 2c)/18, \text{ and}$$
$$\pi^S = (4U + c)(U - 2c)^2/54.$$

The consumer surplus and the social welfare can be obtained accordingly.

$$CS^S = \int_{2(U+c)/3}^{U} S_u du = (7U + 4c)(U - 2c)^2/162,$$

and

$$\begin{aligned} SW^S &= CS^S + \pi^S \\ &= (19U + 7c)(U - 2c)^2/162. \end{aligned}$$

The equilibrium exists when the maximum willingness to pay is sufficiently larger than the marginal cost of providing service, or

$$U \geq 2c. \tag{94.4}$$

We assume that (94.4) holds throughout this paper.

94.3.2 Pay Per Use Pricing

Next, turn to pay per use pricing. Since $p > 0$ and $A = 0$, we have

$$q = u - p \quad \text{and} \quad S_u = (u - p)^2/2.$$

The purchase occurs when $q = u - p \geq 0$, or $u \geq p$. The number of consumers purchasing service, the total quantity of service consumption, and the profit of the provider can be calculated similarly to the case of subscription pricing.

$$N = U - p,$$
$$Q = \int_p^U (u - p)du = (U - p)^2/2, \text{ and}$$
$$\begin{aligned} \pi &= pQ - cQ \\ &= (p - c)\left\{(U - p)^2/2\right\}. \end{aligned}$$

The first order condition ($d\pi/dp = 0$) generates the optimal pay per use price.

$$p^P = (U + 2c)/3.$$

Superscript P is used to denote 'pay-per-use pricing'. The marginal consumer who is indifferent from buying and not buying is

$$u^P = p^P = (U + 2c)/3.$$

From this, the number of consumers purchasing service, the quantity of service consumption, the profit of the provider, the consumer surplus, and the social welfare at the equilibrium can be derived as follows.

$$\begin{aligned} N^P &= 2(U-c)/3, \\ Q^P &= 2(U-c)^2/9, \\ \pi^P &= 2(U-c)^3/27, \\ CS^P &= \int_{(U+2c)/3}^{U} S_u du = 4(U-c)^3/81, \text{ and} \\ SW^P &= 10(U-c)^3/81. \end{aligned}$$

94.3.3 Two-Part Tariff

In case of two-part tariff, it holds that $p > 0$ and $A > 0$. It follows that

$$q = u - p \quad \text{and} \quad S_u = (u-p)^2/2 - A.$$

The condition that a consumer purchase service is given by

$$(u-p)^2/2 - A \geq 0, \text{ or } \quad u \geq p + (2A)^{1/2}.$$

The number of consumers purchasing service, the total quantity of service consumption, and the profit of the provider are given as follows.

$$\begin{aligned} N &= U - p - (2A)^{1/2}, \\ Q &= \int_{p+(2A)^{1/2}}^{U} (u-p)du = (U-p)^2/2 - A, \text{ and} \\ \pi &= (AN + pQ) - cQ \\ &= A\left\{U - p - (2A)^{1/2}\right\} + (p-c)\left\{(U-p)^2/2 - A\right\}. \end{aligned}$$

By simultaneously solving the first order conditions ($d\pi/dA = 0$, $d\pi/dp = 0$), we have the optimal two-part tariff.

$$A^T = 2(U-c)^2/25, \text{ and } \quad p^T = (U + 4c)/5,$$

To denote 'two-part tariff', we use superscript T. The marginal consumer who is indifferent from buying and not buying is

$$u^T = p^T + (2A^T)^{1/2} = (3U + 2c)/5.$$

From this, the number of consumers, the quantity of service use, the profit of the provider, the consumer surplus, and the social welfare at the equilibrium can be calculated as follows:

$$\begin{aligned} N^T &= 2(U-c)/5, \\ Q^T &= 6(U-c)^2/25, \\ \pi^T &= 2(U-c)^3/25, \\ CS^P &= \int_{(3U+2c)/5}^{U} S_u du = 6(U-c)^3/125, \text{ and} \\ SW^T &= 16(U-c)^3/125. \end{aligned}$$

94.3.4 Comparisons

Let us examine which pricing scheme maximizes the provider's profit. By comparing π^i's (i = S, P, T) we obtain the following proposition.

Proposition 1 *The profit of the monopoly provider is maximum under two-part tariff and minimum (under)subscription pricing, that is,*

$$\pi^S \leq \pi^P < \pi^T.$$

The equality holds only when the marginal cost of providing service is zero.

Proof From (94.4),

$$\begin{aligned} \pi^P - \pi^S &= c(3U^2 - 8c^2)/54 \geq 0, \text{ and} \\ \pi^T - \pi^P &= 4(U-c)^3/675 > 0. \end{aligned}$$

It holds that $\pi^P = \pi^S$ when $c = 0$. □

The merit of subscription pricing from provider's perspective is to receive much revenue from the consumers with low willingness to pay. On the contrary, the provider can obtain a large amount of revenue from the consumers with high willingness to pay under pay-per-use pricing. As shown in the fact that $\pi^P = \pi^S$ when $c = 0$, the provider can make equal revenues from either pricing by optimizing lump-sum and unit prices. However, the quantity of consumption is too large under subscription pricing compared to the other pricing schemes

($Q^P < Q^T < Q^S$). When the marginal cost of providing service is non-negligible, the profit from subscription pricing is least among all the pricing schemes. Two-part tariff balances the revenue from high-end consumers and that from low-end consumers and maximizes the profit of the provider.

Let us turn to the consumer surplus and social welfare. The calculation is straightforward, so we propose the following proposition without a proof.

Proposition 2 *Consumer surplus is highest under pay-per-use pricing and lowest under subscription pricing, that is,*

$$CS^S < CS^T < CS^P.$$

Low-end consumers prefer pay-per-use pricing to subscription pricing because the former better reflect the actual amount of her consumption. On the other hand, high-end consumers prefer subscription pricing to pay-per-use pricing because the former makes her payment constant regardless of the actual amount of her consumption. The preference for two-part tariff lies in between.

As a result, the marginal consumer who is indifferent from buying and not buying under pay-per-use pricing (u^P) does not purchase service under two-part tariff. Similarly, the marginal consumer who is indifferent from buying and not buying under two-part tariff (u^T) does not purchase service under subscription pricing. It follows that $u^P < u^T < u^S$ and $N^S < N^T < N^P$. Under pay-per-use pricing, the largest size of consumers is served and the consumer surplus is highest. Under subscription pricing, the consumers with low willingness to pay cannot purchase the service due to the high level of initial payment (A^S).

Taking the above two propositions together into account, we propose the following proposition. The proof is straightforward and can be omitted.

94.3.5 Discussion and Implications

Our predictions are that the cloud service models with sizable marginal costs such as IaaS and PaaS tend to be coupled with pay-per-use pricing, and the others with subscription pricing. Pay-per-use pricing is general among IaaS and PaaS and subscription pricing is often used SaaS offerings. The cost structure of service providers cannot fully explain the reality, but is one of the important determinants of pricing schemes.

Despite its economic performance, two-part tariff is not common in the cloud computing industry. One explanation would be the complexity of implementation. However, we expect that two-part tariff would become more popular among service providers as the market competition get fiercer and the profit margin grows narrower.

94.4 Conclusion

In this paper, we investigates subscription pricing, pay-per-use pricing, and two-part tariff from perspectives of a provider's profit, consumer surplus, and social welfare, respectively. This paper shows that two-part tariff is more efficient than the other two pricing from a provider's, consumers' and social point of view.

Our main results are first, the profit of the monopoly provider is maximum under two-part tariff and minimum (under) subscription pricing. Second, consumer surplus is highest under pay-per-use pricing and lowest under subscription pricing.

Acknowledgments This work was supported by Seoul National University of Science and Technology.

References

1. Buyya R, Yeo CS, Venugopal S, Broberg J, Brandic I (2009) Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility. Future Gener Comput Syst 25(6):599–616
2. Yeo CS, Venugopal S, Chu X, Buyya R (2010) Autonomic metered pricing for a utility computing service. Future Gener Comput Syst 26:1368–1380
3. Li C-F (2011) Cloud computing system management under flat rate pricing. J Netw Syst Manage 19(3):305–318
4. Chun S-H, Choi B-S (2013) Service models and pricing schemes for cloud computing. Cluster Computing (In Press)
5. Song JD (2006) An applied study of DRM to distribution and price structure of digital contents. J Global Sch Mark Sci 16(1) (In Korean)
6. Blau B, Stößer J (2009) Business models in the service World, IT Pro March/April 2009, pp 36–41

Chapter 95
Platform Design for Co-creation

Soon Goo Hong, Hyung Rim Choi, Hyun Jong Kim, Chang-hyun Park and Hyun Mi Lee

Abstract Following the advancement of IT and changes in consumer consciousness, enterprises are actively inducing consumer participation for value creation. Accordingly, co-creation, in which enterprises cooperate with diverse stakeholders including consumers to jointly create values, is emerging as a new paradigm of enterprise management. For enterprises to utilize co-creation systematically, platforms through which diverse stakeholders' opinions can be collected and reflected are required. Therefore, in this study, the requirements and functions of co-creation platforms were derived through case studies and related structures, and components of platforms were defined. In addition, a co-creation platform was designed based on the operation scenario through which co-creation occurs in the processes of enterprises' new product development, planning, and production. As one of the pioneering studies in platform for co-creation, this research contributes to both theory and practice.

Keywords Co-creation · Platform design

95.1 Introduction

Thanks to the advancement of IT, consumers can easily obtain diverse kinds of information on products and services in real time through the Internet and SNS (Social Network Service), among others, utilizing IT devices such as smartphones and tablet PCs. Consumers use such information for decision making in purchasing products and services. In addition, as consumers share and spread experiences with the purchase and use of products and services, they affect other consumers' purchases.

S. G. Hong (✉) · H. R. Choi · H. J. Kim · C. Park · H. M. Lee
Dong-A University, Busan, South Korea
e-mail: shong@dau.ac.kr

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_95, © Springer Science+Business Media Dordrecht 2014

Because consumers have changed from being passive to being active along with the advancement of IT, enterprises have realized the importance of consumers and are now actively using their ideas and opinions. Enterprises have moved beyond the level of simple consumer management and are involving the consumers in value creation by allowing them to participate actively in all processes of value change, such as product planning, tests, development, and marketing. Co-creation, which is cooperation between enterprises and consumers to create new values, was first introduced in The Future of Competition by Prahalad and Ramaswamy [1], and has become a new paradigm of enterprise management.

The effects of co-creation can be maximized through IT. In the age of Web 2.0, people have become capable of the realization of opening, sharing, and cooperation through the Internet because IT has been developed. In the relationship between enterprises and consumers, information is exchanged more actively through the Internet than ever before. However, enterprises are not yet actively utilizing IT technologies for realizing co-creation. Although enterprises are attempting communications with customers in diverse methods such as through their websites and SNS, the methods are operated independently by individual enterprises. Therefore, overlapping IT investments for the utilization of co-creation are made, and also the investments are not systematic.

The purpose of this study is to develop a platform for realizing co-creation. Co-creation platforms should be able to use the internal information systems that are already being used by enterprises, IT, and external media such as SNS. In addition, co-creation platforms should be designed so that external media conveniently used by consumers, such as SNS and enterprises' information systems, can be interlocked with each other.

95.2 Requirements and Functions of Co-creation Platform Design

95.2.1 Requirements of Co-creation Platforms

To design platforms for realizing co-creation, the requirements and functions of these platforms should be derived. Co-creation platforms should connect enterprises and consumers with each other and provide cyberspaces where common values and necessary functions can be created. Thus, previous studies and cases were investigated to derive the requirements of these platforms.

The four cases in Table 95.1 are representative cases of co-creation where enterprises had consumers actively share their opinions, and the enterprises succeeded in the production of their products. These cases are appropriate for analyzing the requirements of co-creation platforms.

Starbucks and Vitamin water used SNS to enable consumers to register their ideas and product designs firsthand. Quirky enabled the manufacture and sale of

Table 95.1 Co-creation cases where IT was used

Enterprise	Channels	Utilization of co-creation	Characteristics	Requirements
Starbucks	SNS	Path for communication with consumers	Collect consumer opinions through diverse SNS channels and reflect the opinions on service	Collection and analysis of data from the outside, such as from SNS
Vitamin water	Facebook	Collection of tastes and ideas through application	Collect consumer information using diverse media and applications	Expandability and application usability
Quirky	Home page	Industrialization and commercialization of consumers' ideas	Provide platforms that connect between idea providers and users	Ease and reliability of idea registration and adoption
4 Foods	Home page	Consumers design, register, and sell products	Provide information on the present situation of product registration and sales and revenues to consumers	Consumer participation and value creation/ distribution

new products and products improved from existing ones based on ideas presented by several people.

For enterprises and consumers to be able to propose ideas through various media, the enterprises should be able to comprehensively process various kinds of information generated by diverse media. In addition, various functions of co-creation should be capable of analyzing information and supporting cooperative activities using IT technologies and resources external to platforms. Therefore, requirements of co-creation platforms may include information sharing, consumer participation, enterprises' data collection and analysis, value realization and distribution, and system expandability.

To derive co-creation platforms' functions, the implementation of enterprises' value creation activities through co-creation platforms was examined through the scenario of Enterprise A.

Enterprise A intends to actively utilize consumers' participation in the stages of new product development, production, and sales. To this end, Enterprise A analyzes consumers' opinions and complaints in the stage of product planning. It makes diverse stakeholders such as customers and experts participate in all processes of value chains, such as idea collection and selection for production,

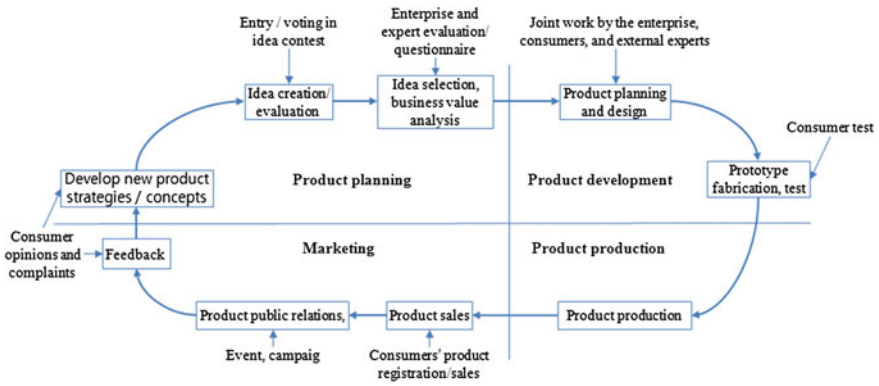


Fig. 95.1 Co-creation in the processes of new product development

planning and design for industrialization, prototype tests, marketing, and feedback. In these processes, co-creation platforms are utilized so that the new product development processes are repeated, as shown in Fig. 95.1.

In enterprises' new product development and production, co-creation makes the following actions to be processed through enterprises' internal information systems; consumers' presentation of opinions and complaints, invitation of ideas of the public, joint work of enterprises, consumers, and external experts, consumer tests, consumers' product sales, and consumers' participation in events the platform is in charge of connection between enterprises and consumers, and collection and transmission of information.

95.2.2 Functions of Co-creation Platforms

The functions of co-creation platforms based on the aforementioned scenario of Enterprise A can be divided into primary functions and secondary functions, as shown in Table 95.2. The primary functions are information collection, analysis, and propagation for co-creation; the secondary functions are platform process management functions. Among the primary functions, information collection refers to collecting information from external media and existing information systems in addition to the information generated by the platform. Information analysis is analyzing and processing collected information from inside and outside the platform. Information propagation is transmitting the information generated by the platform to consumers through external media. The secondary functions include managing the operations and processes of individual modules in the platform, checking the results of information processing through the manager, and final review and decision making.

Co-creation platforms can be illustrated by conceptual diagrams, as shown in Fig. 95.2.

Table 95.2 Functions of co-creation platforms

Division		Detailed functions	
Primary functions	Information collection	Information collection through external media	• Entry, voting, and opinion (consumers, experts) collection through external media • Collect the responses to questionnaire surveys through external media • Receive applications for participation in events through external media
		Information collection within the platform	• Collect consumer opinions and complaints prepared in the platform
		Information collection from existing systems	• Collect consumer opinions stored in the DB(database) of the existing information system (customer management system, ERP, etc.)
	Information analysis	Inside the platform	• Process the information collected by individual service modules • Generate the results of event participation and registration • Select ideas through voting information • Store the results of processing in/out of the platform
		Utilize existing systems	• Analyze and store collected information utilizing existing systems • Perform joint works utilizing cloud services
	Information propagation	Propagate information through external media	• Propagate event notices through external media • Propagate event contents through external media • Propagate event results through external media
		Deliver messages	• Deliver the details of events in service modules to the manager • Deliver the results of entries and registration to consumers
Secondary functions		Process management	• Creation and management of service modules • Management of processing processes between service modules and information transmission • Management of work information and data on consumer participation and joint work services
		Delivery to the manager	• Review of ideas and decision making • Select participants in the development process from among applicants for event participation and give them access rights

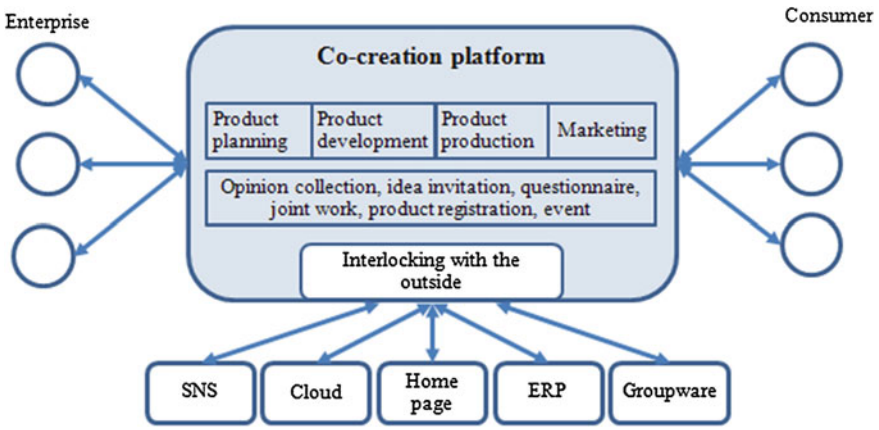


Fig. 95.2 A conceptual diagram of a co-creation platform

95.3 Design of Co-creation Platforms

95.3.1 Co-creation Platforms' Structure and Components

Co-creation platforms should be constructed as follows:

First, co-creation platforms should have functions to collect information for interlocking with external media and to propagate event information to consumers. They should be capable of utilizing collected information using existing systems. Consumers' opinions and complaints prepared outside of enterprises are delivered to the relevant enterprises according to items that correspond to the enterprises and their products. Those enterprises store the information and analyze the contents using their information systems. In this way, the contents of events occurring in enterprises can be quickly and widely delivered to consumers through external media such as SNS.

Second, co-creation platforms basically deliver their writings to enterprises and consumers, collect enterprises' writings on SNS through interlocking with external SNS, deliver the writings to enterprises while analyzing the writings to collect writings about enterprises and products and analyze and store the contents. In addition, co-creation platforms should enable enterprises to deliver answers to consumers' opinions and complaints. These functions may be expanded and used for idea collection, events, and questionnaires, and additional functions needed by enterprises may be added.

A block diagram of a co-creation platform is shown in Fig. 95.3. A co-creation platform is composed of a service module, service creation and process management modules, a legacy system management module, and an Open API (Application Programming Interface). The functions of individual modules are shown in Table 95.3.

The co-creation service module provides functions necessary for enterprises to handle business through co-creation as services. For instance, these services include providing screens necessary to conduct questionnaire surveys, sending

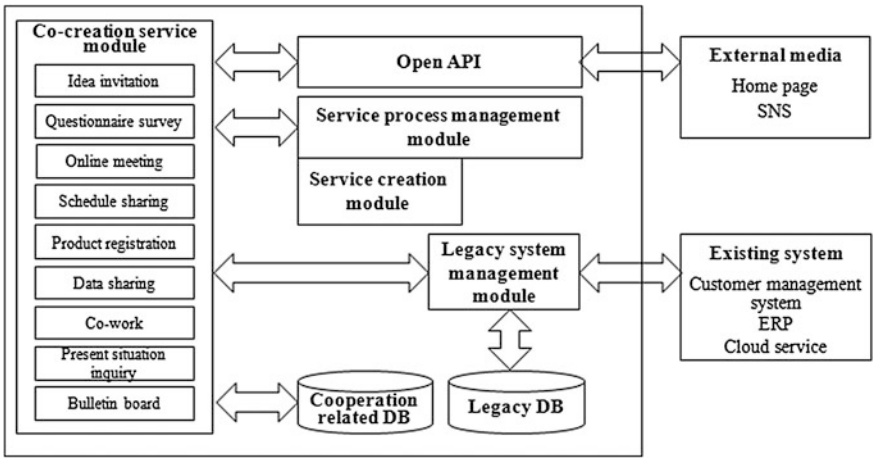


Fig. 95.3 A block diagram of a co-creation platform

Table 95.3 Platform components

Module	Functions
Co-creation service module	Co-creation related services. Individual service modules are defined on input data and output data and linked with other services
Service creation module	Service creation and service life cycle management
Service process management module	To implement work processes through co-creation, this module links individual services and integrates and stores the data generated in individual services
Legacy system management module	Manages existing information systems and process works such as statistics and analysis
Cooperation-related DB	Manages data (big data, cloud data, etc.) generated by the service process management module
Open API	Provides interfaces for functions (external services) needed by service modules

questionnaires, receiving responses, statistical processing, collecting consumer opinions and ideas, preparing questionnaires, and joint work. Individual services can be organized with diverse languages and the forms of data that can be generated through services are also different.

The service creation module manages activated co-creation services. This module uses Web Service technology to create and find services. For use of co-creation services, this module makes lists of services (WSDL: Web Service Definition Language) and registers the lists in directories such as UDDI (Universal Description, Discovery and Integration) to make service searches and use easy.

The service process management module defines the order of works for linkages with individual co-creation services and delivers data between individual services to support work processes through co-creation. Modules for supporting co-creation are written in different languages and create diverse forms of data. To organically link these modules with each other, services necessary for co-creation are created

through the service creation module and changes in work processes are supported using information such as data links between individual services and orders.

The platform is interlocked with home pages, SNS, and cloud computing services through the open API. In addition, the platform collects data prepared outside through the open API, classifies the data through data collection agents, and delivers the data to processes and modules for the relevant functions.

95.3.2 Platform Operation Scenario

To design the platform based on the functions, composition, and modules of the platform defined earlier, a platform operation scenario was composed to define the links and data flows of modules for platform operation. Among the processes for product planning, development, production, and marketing, the operation scenario was prepared, centering on product planning and development processes in which co-creation mainly occurs.

95.3.2.1 Product Planning

In the new product planning stage, consumers' opinions and complaints are reviewed, ideas are provided by consumers, and effects on consumers and markets are grasped through questionnaires to plan new products.

Product planning operation scenario

1. The manager creates a new product planning project through the service process management module and sets up a data collection period.
2. The process management module creates and manages services necessary for product planning.
   - 2.1 The manager selects invitation of ideas from the public and questionnaires on service functions from among the data collection methods (i.e., opinion collection, invitation of ideas from the public, questionnaire survey).
   - 2.2 When these services have been selected, the service process module receives the locations of services necessary to it from the directory service using the WSDL of the service creation module and creates and executes a public idea invitation module and a questionnaire survey service module using SOAPs (Simple Object Access Protocol).
   - 2.3 The public idea invitation module and questionnaire survey service module send the ideas and the results of aggregation of questionnaire surveys periodically generated through e-mails and post the same on SNS.
3. During data collection periods, individual service modules periodically monitor bulletin boards and SNS and store the collected information in the cooperation related DB.

3.1 The public idea invitation module monitors external media to collect information (entries, dates of posting, types of ideas, titles, contents of ideas, URL).

4. When the data collection period has expired, the process management module terminates the existing public idea invitation service and questionnaire survey service.
5. The public idea invitation results (idea voting ranks) and questionnaire survey results are generated and delivered to the manager.

95.3.2.2 Product Development

In the new product development stage, the ideas selected in the product planning stage are used to design products through joint work with consumer participation, and prototypes are tested.

Product development operation scenario

1. Through the service process management module, the manager creates a new product development project and sets up an application period and a working period.
2. The process management module creates and manages services necessary for product development.
   2.1 The manager selects joint work and prototype test service functions from among the product development methods.
   2.2 When services have been selected, the service process module receives the locations of services necessary to it from the directory service using the WSDL of the service creation module and creates and executes joint work, prototype test service modules, and questionnaire survey service using SOAPs (Simple Object Access Protocol).
   2.3 The joint work module and prototype test service module post applicant recruitment information on bulletin boards and SNS to deliver the information to consumers.
3. During the application period, individual service modules monitor external media (e-mails, bulletin boards, SNS, etc.) and store the collected applicant information in the cooperation related DB.
   3.1 The joint work service collects applicant information (applicants, area of application, contents of application) through external media and stores it in the cooperation-related DB.
   3.2 The prototype test service collects applicant information (applicants, area of application, contents of application) through external media and stores it in the cooperation-related DB.
4. When the application period has expired, individual service modules generate applicant results and deliver them to the manager.

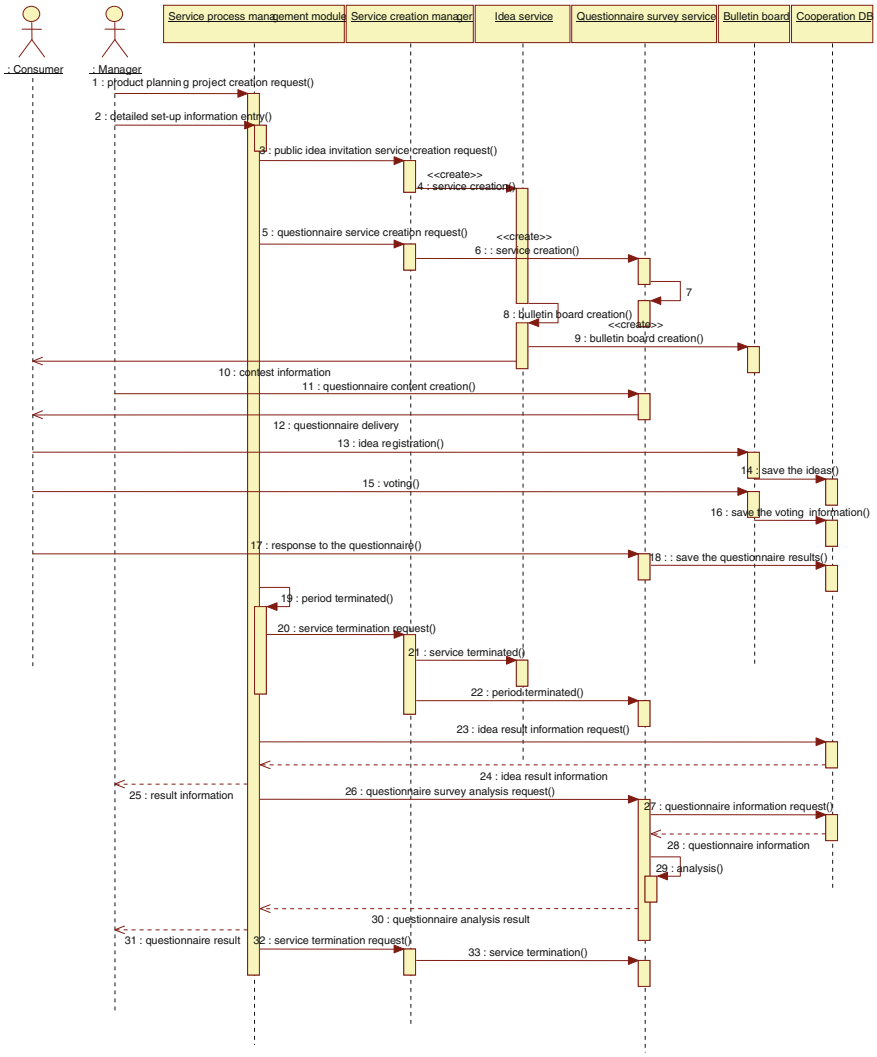


Fig. 95.4 Sequence diagram

5. When the manager has selected participants from among the applicants, these applicants are notified of their acceptance, and joint work and prototype tests are conducted.
   5.1 The joint work service provides schedules; enables data sharing; gives the right to access the joint work system; delivers messages containing the schedules, data, and work details to individual participants; and stores the details in the cooperation-related DB.
   5.2 The prototype test service collects the prototype test results prepared by the participants and stores the results in the cooperation-related DB.

6. When the work period has expired, the process management module terminates the existing joint work and prototype test service modules.
7. Joint work results (products, project, work period, participants, work details) and prototype test results (tester, test date, test contents) are generated and delivered to the manager.

95.3.3 Sequence Diagram

A sequence diagram of the product planning operation scenario is shown in Fig. 95.4. The manager creates the service modules, and consumers participate by sharing ideas and answering questionnaires. When the period has expired, the results are delivered to the manager, and the services are terminated.

95.4 Conclusion

In this study, a platform for effective and systematic support for co-creation between enterprises and consumers was designed. The requirements and functions of the platform were organized through investigations of diverse co-creation related cases. In addition, module linkages and data flows necessary for the co-creation platform to operate through connections between modules were analyzed to define the structure and components of co-creation. Then, centering on the product planning and development processes, platform operation and information flows were expressed as an operation scenario and a sequence diagram.

Although enterprises use many channels including home pages and SNS to collect consumer opinions and ideas, these activities are very inefficient because such activities are manually performed. However, if the activities are performed through co-creation platforms, enterprises can automatically collect consumer opinions and ideas and can eventually induce consumer participation quickly and extensively through diverse external media. In addition, through the services provided by the platforms, enterprises can realize co-creation more efficiently for a series of activities ranging from new product planning to marketing.

This study is a pioneering study of co-creation platform that can prompt future research in this field. In practice, the findings of this study provide a reference for managers considering or implementing co-creation platforms.

A limitation of the study is that, although a basic design for co-creation platforms was created, the designed platform was not implemented. Therefore, later studies should implement the platform and evaluate its use in comparison with the use of existing IT.

Acknowledgments This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2012S1A3A2033394).

References

1. Prahalad CK, Ramaswamy V (2004) The future of competition. Harvard Business Review Press, Boston
2. Kim JH, Hong JH (2011) Co-evolution. Hans Contents
3. Gong YI (2013) Social product development platform, Quirky. KISDI Commun Policy 25(14)
4. Ha BC (2001) A inter-industrial comparative analysis of new product idea generation approach. Master's Thesis, Seoul University
5. http://www.starbucks.com, 2013
6. http://www.facebook.com/vitaminwater, 2013
7. http://www.quirky.com, 2013
8. http://www.4food.com, 2013

Chapter 96
Towards the Implementation of an Ubiquitous Car Park Availability Detection System

Ian K. T. Tan, Poo Kuan Hoong, Chee Ken Hong and Low Zhi Wen

Abstract Wasted time spent on searching for available car park bays can be reduced through cost effective implementation of vacant car park bay detector that communicates to interested drivers through social media. Our proposed solution uses low cost and low power requirement Raspberry Pi board with the OpenCV library that communicates to the drivers through the Twitter micro-blogging eco-system. Preliminary tests indicate that the project can achieve detection of up to 99 % accuracy during hours of interest. The proposed solution is for open space car parks where the hours of interest are during day light.

Keywords Open source · Image processing · OpenCV · Raspberry pi · Twitter · Car park

96.1 Introduction

Searching of vacant car park bays can be time consuming for many drivers, especially in crowded car parks without any integrated car park management system. Many malls have installed costly car park management system to assist drivers to locate suitable vacant bays through the implementation of digital signages coupled with sensors at each parking bay to detect availability. However, in open space car parks such as those found in recreational parks or stadiums, implementing these integrated car park management system would be unjustifiable. They are prohibitively expensive as most of these car parks would not generate any or sufficient revenue to justify the implementation and maintenance of such systems.

I. K. T. Tan (✉) · P. K. Hoong · C. K. Hong · L. Z. Wen
Faculty of Computing and Informatics, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia
e-mail: ian@mmu.edu.my
URL: http://pesona.mmu.edu.my/iantankt/

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_96, © Springer Science+Business Media Dordrecht 2014

Furthermore, the system would not be located in a secure closed environment and will be in open spaces that are subject to varying weather conditions.

To alleviate this issue, the solution has to have a low financial entry point. Some of the major cost in installing the integrated car park management system would include the wiring required for all the sensors, individual sensors at each parking bay and the digital signages to inform drivers on the car park bay availability. However, the required system need not be 100 % accurate as it is deemed as a public service to drivers due to the fact that there would generally be none or minimal car park charges.

The objectives of the required solution are,

- To minimize the need for costly sensors and the wiring of the sensors,
- To utilize alternative methods to inform drivers in order to eliminate the need for digital signages, and
- To ensure high accuracy during daylight as these systems are typically only crowded during daylight, for example recreational parks.

96.2 Solution Overview

To achieve the objective of minimizing sensors and the amount of wiring required, the solution is designed with using image processing technology on low cost hardware. The project opted for Raspberry Pi [1] running GNU Linux, using an entry level webcam and utilizing the OpenCV image processing libraries [2].

As for alternative replacement for digital signages to guide drivers, the solution will utilize the Twitter micro-blogging medium (http://www.twitter.com/). This is a chosen medium as many existing road traffic information systems have already started using it as a preferred communication medium with road users. Furthermore, the smartphone market penetration is continuously growing and this will lead to it being a widely available channel.

96.2.1 Image Processing to Eliminate Expensive Sensor Implementation

Image processing techniques to determine car park spaces is not a new idea. Several researches [3–7] have not only proposed using this technique but have also proposed enhancements to address variations in weather conditions and better accuracy for night time.

Bong et al. [3] pre-process a reference empty car park and through their image processing techniques, they were able to obtain accuracy of around 90 % even in heavy rain conditions.

Huang et al. [4, 6] proposed a Bayesian inference method that uses a multi-faceted approach to a 3-dimensional image analysis to determine car park bay

availability. It proved to work well during daylight and acceptable accuracy when there is little light source. Their method would be taxing for a low powered device that our project planned to use. Jermsurawong et al. [5] uses dynamic statistical method to ensure that the thresholds used is dependent on the available light in their attempt to determine car park bay availability throughout the day (and night).

Work done by Yusnita et al. [7] uses image segmentation techniques that requires a brown spot to be painted on the car park bays and uses a 7 segment display to indicate the number of car park bays available. This has drawbacks in that the brown spot will fade over time under harsh weather conditions and there is a cost to the limited 7 segment display.

However, the intended solution has to be small in physical size and sufficiently low cost for it to be easily deployed anywhere. Low cost would also imply that the system will have low compute capability. With this, the project builds upon the ideas and work previous completed by [3] and [8] in using background subtraction techniques which has a very low compute requirements.

Background Subtraction Background subtraction is an elementary image processing technique which attempts to extract foreground information against a reference image, called the background image. It is generally used to extract the presence of objects in the image, in our case, the objects are vehicles.

96.2.2 Social Media for Communication Channel

Part of the car park information system is the need to communicate with drivers. This is generally done using digital signages that attempt to not only provide car park availability information but can also be used to guide the drivers to the vacant car park bays. With the emergence of autonomous and connected cars, these information can be relayed directly to the car infotainment system. However, as this has yet to be prevalent, we looked at the Twitter micro-blogging eco-system. The prevalence of smart phone has ensured that drivers are likely to be a smart phone user. This is inferred from the fact that smart phones take up a high percentage of the mobile phone market and are generally on the higher end market. If one can afford a car, then it is also likely that they would be a smart phone owner.

The solution will tweet the number of car park available on a specific twitter user account where those who are interested can “follow” that twitter account. This eliminates the need for digital signages to be installed at the car parks. However, a drawback of this approach is that it will lack the capability to guide the driver to the vacant car park space. Although a very accurate guide can be implemented, it would require the driver to turn on their Global Positioning System (GPS) on their smart phones and refer to their phones regularly whilst driving. Eventually, we envisaged that the connected autonomous vehicles will have this integrated into their vehicles infotainment system and text-to-speech capabilities to read out the directions accurately to the drivers.

96.3 Implementation

This project will utilize low cost hardware components, open source software and related libraries where applicable.

96.3.1 Hardware and Software Tools

The hardware and software tools used in this project is listed in Table 96.1. The total hardware cost is approximately USD80 (depending on shipping and fluctuation in currency exchange) and the software tools used are Open Source.

96.3.2 Processing Flow

The proposed solution has the following sequence.

- Image Capturing: Capture and store image.
- Image Pre-processing: Convert RGB to Grayscale, conduct Background Subtraction, Smooth Filtering, Thresholding, Dilation Filtering and Binary Image Output, as illustrated in Fig. 96.1.
- Car park availability detection.
- Tweet car park availability with a Twitpic image for verification (this is only used for our verification on the accuracy of the detection).

Image Capturing: Capture and Store Image GTK+ UVC Viewer is used in conjunction with the Logitech C110 web camera to capture images of the car park area every 5 min. Other than the reference image, the captured image will overwrite previously captured image in order to ensure that the onboard storage on the SD Card will not be depleted. The command used is;

```
guvcview −no_display −r 1 −c 10 −m 1 −exit_on_close
```

Pre-processing To minimize colour conversion errors in the images, the reference and captured images are converted from RGB to gray scale. Differences will then be determined by subtracting the gray scale captured image with the corresponding reference image. Figure 96.2 provides two illustrations of background subtraction. The left most image is the reference image, the middle image is the foreground image and the resultant image is on the right.

The resultant gray scale image will then be processed to remove extraneous data or noise by smoothing the image using a blurring technique. A 7 pixel by 7 pixel matrix kernel is used for this purpose. This effectively removed small changes (noise) from the image.

Table 96.1 List of hardware and software tools used

Hardware	Software
Raspberri Pi model B	Raspbian Wheezy (GNU Linux)
Logitech C110 webcam	GNU G++ compiler
Sandisk 8 GB SD card	OpenCV 2.4.4 libcurl, twitcurl and twitpic library GTK+ UVCViewer

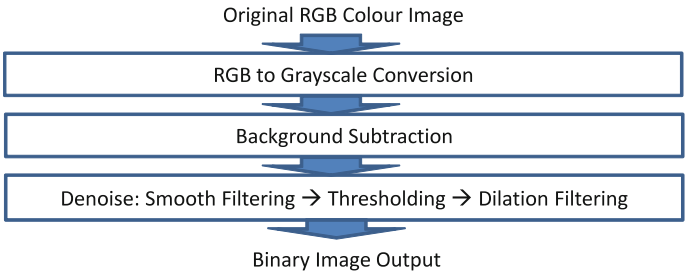


Fig. 96.1 Pre-processing flow

Fig. 96.2 Example of background subtraction

To produce a binary image (an image with just black or white pixels) to reduce the data complexity, binary thresholding is done. This is expressed as in Eq. 96.1.

$$result(x,y) = \begin{cases} 255, & \text{if } input(x,y) > tValue \\ 0, & \text{otherwise} \end{cases} \tag{96.1}$$

where *result*(x, y) is resulting binary image, *input*(x, y) is input gray scale image, and *tValue* is threshold value.

Fig. 96.3 Denoise: smooth filters, thresholding and dilation filtering

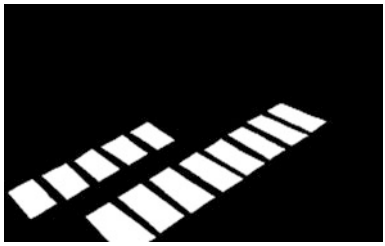


Fig. 96.4 Pre-prepared mask for the 13 car park bays

For our project, we selected the threshold value of 30 resulting in a binary image. In order to make the binary image more prominent, a dilation process is applied. This effectively accentuate the objects (white pixels) detected. A sample final output of this stage is as illustrated in Fig. 96.3.

Image Detection Instead of detection objects, this project uses occupancy detection. A binary mask is used for each car park bay in question and this is used as the reference image. The image will represent each individual car park bay, also known as the region of interest (ROI). The project was tested using 13 car park bays. The pre-prepared mask is as shown in Fig. 96.4.

The mask will then be applied to the processed binary image to ignore any unwanted region of the image. The remaining ROI is then extracted out and car park occupancy is determined by computing the sum of white (object) pixels. A car park bay is deemed occupied if the sum is 70 % or greater.

Twit Status and Twitpic Image Once the number of car park vacancy is computed, the results is tweeted using the twitpic Application Programming Interface (API). The project also post the resultant image on a corresponding twitpic website to facilitate the accuracy verification of the project. The result is as shown on Fig. 96.5.

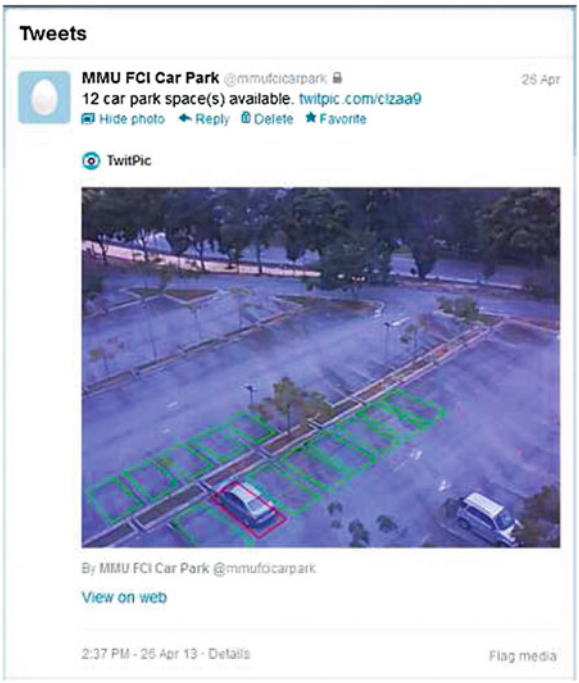


Fig. 96.5 Final tweet message

96.4 Result Analysis

Using the background subtraction technique to determine vacant car park bays has several limitations. Some of the limitations that we encountered are:

- The reference image used must represent an empty car park.
- The location and angle of the web camera used has to be positioned directly above. Furthermore, to capture a large area, the web camera has to be placed very high as the web camera does not have a sufficiently wide lens. Furthermore, as the web camera does not capture high resolution pictures, it limits the number of car parks it can include.
- As with any image processing techniques, the amount of light available affects the image capture. Time of day, weather conditions, shadows from nearby objects and even strong winds swaying trees affect the image capture.

Since the aim of this project is for open space car parks located in areas such as recreational parks, the time of day would be limited to day light hours, good weather conditions and open spaces with little tall neighbouring objects to cast shadows.

We conducted our accuracy measurements within the following conditions:

- During day light hours, in this case 8 a.m. to 6 p.m.
- Web camera is placed on the fourth floor of a building.

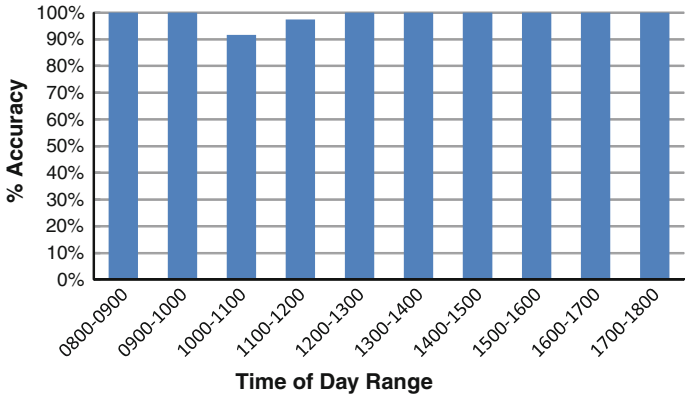


Fig. 96.6 Average accuracy throughout day light hours

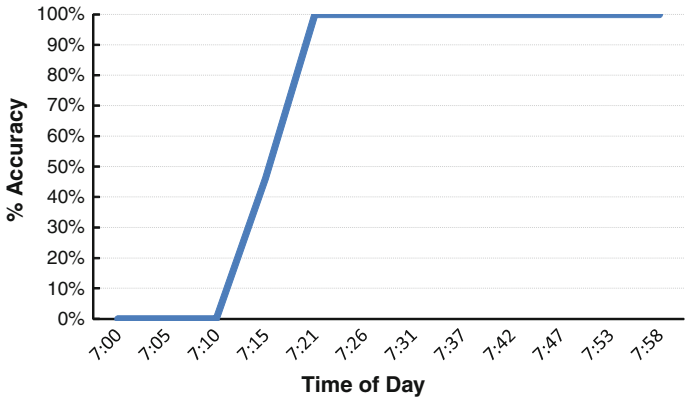


Fig. 96.7 Accuracy between 7:00 a.m. and 8:00 a.m

- We monitor 13 car park bays.
- Measurements are only taken during good weather conditions.

For every hour, we computed 12 instances (every 5 min) and through accuracy verification using the twitpic images, we took an average accuracy rate.

Figure 96.6 shows that the system is able to perform well. The average accuracy for the whole day (between 8 a.m. and 6 p.m.) exceeds 99 %. There was a slight dip in accuracy before and around mid-day and we account this error to the shadow from the neighbouring building. Further analysis was conducted for the accuracy during dawn (7 a.m. to 8 a.m.) and dusk (6 p.m. to 7 p.m.) hours and found that the accuracy experience sudden changes at specific times and is not a gradual deterioration. Figures 96.7 and 96.8 illustrate the accuracy at these times respectively.

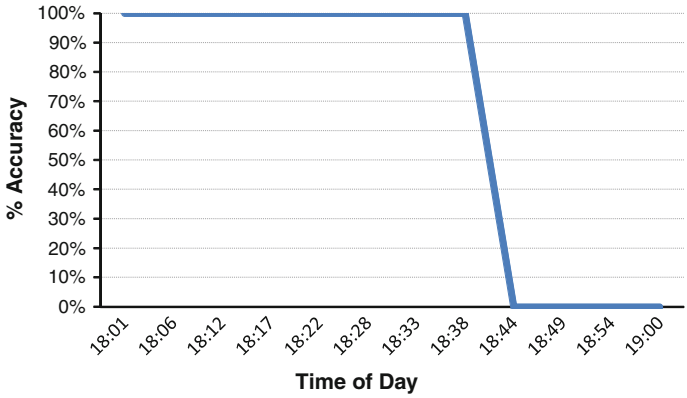


Fig. 96.8 Accuracy between 6:00 p.m. and 7:00 p.m

96.5 Conclusion

The proposed solution has met the low cost requirements with accuracy level of 99 % during day light hours. The implementation as tested still requires cable connection for power and network. Furthermore, the system was located indoors on the fourth floor of a building overlooking the car park area and hence is not weather-proofed.

Future work such as weather-proof housing, solar power [9] and wireless fidelity (Wi-Fi) connectivity will complete the hardware solution. Further improvements on the accuracy during dawn and dusk can be achieved through using appropriate reference images. The system can store reference images for the various hours of the day in order to increase the accuracy rates. This would be similar to the work done by [5] on using different threshold to extract the image differences.

References

1. Computer Laboratory: Raspberry Pi (Online). Available http://www.cl.cam.ac.uk/projects/raspberrypi/
2. Bradski G (2000) The OpenCV Library. Dr. Dobbs J Softw Tools
3. Bong D, Ting K, Lai K (2008) Integrated approach in the design of car park occupancy information system (COINS). IAENG Int J Comput Sci 35(1):1–8
4. Huang C–C, Wang S-J (2010) A hierarchical bayesian generation framework for vacant parking space detection. IEEE Trans Circ Syst Video Technol 20(12):1770–1785
5. Jermsurawong J, Ahsan M, Haidar A, Dong H, Mavridis N (2012) Car parking vacancy detection and its application in 24-hour statistical analysis. In: Proceedings of 10th international conference on frontiers of information technology (FIT), pp 84–90

6. Huang C-C, Dai Y-S, Wang S-J (2012) A surface-based vacant space detection for an intelligent parking lot. In: Proceedings of 12th international conference on ITS telecommunications (ITST), pp 284–288
7. Yusnita R, Norbaya F, Basharuddin N (2012) Intelligent parking space detection system based on image processing. Int J Innov Manage Technol 3(3):232–235
8. Funck S, Mohler N, Oertel W (2004) Determining car-park occupancy from single images. In: IEEE intelligent vehicles symposium, pp. 325–328
9. Hayward D (2012) How to make a raspberry Pi solarpowered FTP server, 7 December 2012 (Online). Available http://reviews.cnet.co.uk/desktops/how-to-make-araspberry-pi-solar-powered-ftp-server-50009923

Chapter 97
Improvised Classification Model for Cloud Based Authentication Using Keystroke Dynamics

T. Senthil Kumar, Abhijit Suresh and Aadarsh Karumathil

Abstract The etymology of communication is the transmission of data. Data has to be transmitted through different devices, network topologies and geographic locations. The strength of communication has tripled with the advent of cloud technologies providing high scalability and storage on demand. The need for cloud security is increasing in an alarming rate and using biometric techniques over traditional password based alternative has proved to be efficient. A behavioral biometric such as keystroke dynamics can be used to strengthen existing security techniques effectively. Due to the semi-autonomous nature of the typing behavior of an individual it is difficult to validate the identity of the user. This paper proposes a model to validate the identity of the user which acclimatizes to tolerance across multiple devices and provides a robust three dimensional model for classification. As an additional layer of security the model is transformed after every login to prevent professional intruders from predicting the acceptance region.

Keywords Authentication · Cloud · Classification model · Keystroke dynamics · Biometric entropy

T. Senthil Kumar · A. Suresh · A. Karumathil (✉)
Department of Computer Science Engineering, Amrita School of Engineering, Ettimadai, Coimbatore, India
e-mail: aadarshadk@hotmail.com

T. Senthil Kumar
e-mail: drtsk111@gmail.com

A. Suresh
e-mail: abhijit5893@hotmail.com

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7_97, © Springer Science+Business Media Dordrecht 2014

97.1 Introduction

Entropy is the quantitative measure of randomness. To determine the best entropy method, time complexity involved in the computation, randomness of each source, error rate, accuracy and cost involved in the system has to be calculated. Entropy methods can be generally classified as biometric, hardware and software entropy methods. Since hardware and software methods are not scalable across multiple devices or sustainable enough to handle the pace of technology, behavioral biometric entropy methods are considered.

Biometric entropy deals with the generation of seed value depending on the measurement and statistical analysis of human body characteristics such as fingerprints, eye retinas and irises, facial patterns and hand measurements for authentication purposes. Keystroke dynamics proves to be cost efficient and scalable since the day to day components of a system (i.e. keyboard) are considered for classification. The concept of keystroke dynamics has prevailed from the age of legacy systems and is expected to sustain into the near future.

97.2 Related Works

There has been many researches related to keyboard dynamics. Rudrapal et al. [1] proposed an algorithm for internal attacker detection by analyzing user keystroke credential. Chang et al. [2] used rhythm click characteristics from non-keyboard devices for user authentication. Bergadano et al. [3] proposed an access control system through keyboard dynamics. Monrose et al. [4] proposed a biometric authentication technique by identifying the habitual rhythm patterns. Everitt et al. [5] proposed an online biometric verification system for use over the internet by combining a typing style test and a mouse based signature test. Boechat et al. [6] presents a biometric authentication through identifying a person's habitual rhythm to type in conventional keyboard. Kekre et al. [7], proposed an approach where the keystroke timing sequence are calculated and normalized then this information is used for generating normalized probability distribution of dynamic passwords using relative entropy and Euclidian distance. Joyce and Gupta [8] used unique digital signature in the form of a stream of latency periods between keystrokes to verify the identity of the user. Johansen [9] researched about keyboard dynamics on a device with a touchscreen. Song et al. [10] studied on the timing analysis and timing attacks on SSH and showed that it is possible to identify the user based on the keystroke timing and the security risks associated with it. Gaines et al. [11] describes a set of experiments performed to investigate whether people can be identified by the way they type. Their finding implied that touch typists appear to have a typing signature and that this method of distinguishing typists might prove the basis for a computer authentication system.

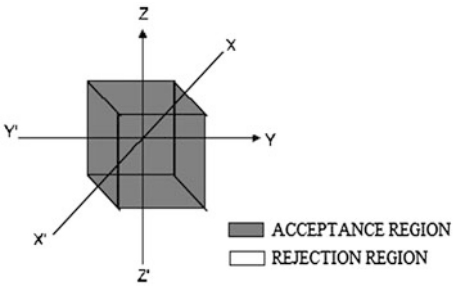


Fig. 97.1 Sample classification model showing the accepted region and rejection region

97.3 Method

Classification model is a predictive model that has a categorical, ordinal, or binary target. The model discussed in this paper is a binary model which returns 1 if the user is authorized and 0 otherwise. To classify the user based on the parameters available for measurement, there is a need for a classification model that is optimal as well as efficient. It should also be able to satisfy the following requirements: Support for different types of keyboards, optimal tolerance and the Error rate should be minimum. The parameters which are used in this paper include: Dwell time: (hold and release time), Pressure, Tolerance rate. The three parameters that has been considered for classification is modeled and visualized in a three dimensional space. The dwell time and pressure is represented using X and Y axis respectively and tolerance in the Z axis. A separate three dimensional space with a model will be assigned to each user. To find the optimal shape for classification, a novel shape fitting algorithm is applied by analyzing existing two and three dimensional shapes. In general basic 2-d shapes like square, rectangle and triangle and 3-d shapes like cube, cuboid and sphere satisfy one or more requirement but not all. To arrive at a shape that satisfies all the requirement, there is a need to combine a shape or two or limit it to a particular variety. Hence the required shape is evolved starting from monogon (Fig. 97.1).

97.3.1 Evolution of Shape

See (Tables 97.1 and 97.2).

97.3.2 Calculation

Based on our evolutional analysis the required shape is constructed step by step. This data is processed and stored in the cloud.

Table 97.1 Two dimensional shapes

Name	Properties	Comments
Henagon (or monogon)	Edges and vertices: 1 Schläfli symbol {1} Coxeter diagram: Dual polygon: Selfdual	Since pressure, dwell time for different type of keyboard is different, a shape surrounding a single point will be not possible
Triangle (or trigon) 3	Edges and vertices: 3 Schläfli symbol {3} (for equilateral) Internal angle (degrees): 60° (for equilateral)	When represented in 3 dimensional space the shape is uniformly expanding. Based on the parameters we are considering the relationship is not proportional and hence uniform expansion is not allowed
Quadrilateral (or tetragon) 4	Edges and vertices: 4 Schläfli symbol {4} (for square) Internal angle (degrees): 90° (for square)	The quadrilateral may be a good shape if we consider approximately four different (key, value) pairs but considering factors like user mood the shape should allow certain tolerance in such a way that it is optimal
Pentagon 5	Edges and vertices: 5 Schläfli symbol {5} Coxeter diagram: Symmetry group:Dihedral (D5), order 2 × 5 Internal angle (degrees): 108° Dual polygon: self Properties: convex, cyclic, equilateral, isogonal, isotoxal	Same as tetragon
Hexagon 6	Edges and vertices: 6 Schläfli symbol {6} Coxeter diagram: Symmetry group: D6, order 2 × 6 Internal angle (degrees): 120° Dual polygon: self Properties: convex, cyclic, equilateral, isogonal, isotoxal	Hexagon is similar to tetragon but considering the base area of the shape, if there area is high the acceptance rate is higher than the threshold which will not provide a suitable security
Heptagon 7	Edges and vertices 7 Schläfli symbol {7} Coxeter diagram: Symmetry group: Dihedral (D7), order 2 × 7 Internal angle: 128.571° Dual polygon: self Properties: convex, cyclic, equilateral, isogonal, isotoxal	Heptagon with 7 sides is irregular and unless if the points does not lie close to each other it will not satisfy our requirement

Table 97.2 Three dimensional shapes

Names	Properties	Comments
Square based pyramid	Faces: 4 triangles 1 square Edges: 8 Vertices: 5 Vertex configuration: 4(32.4) (34) Symmetry group: C4v, [4], (*44) Rotation group: C4, $[4]^+$, (44) Dual polygon: self Properties: convex	Square based pyramid can be plotted with the help of the parameters and can satisfy our requirement but the formation is rare and so considering square properties for base may produce unwanted results
Pentagonal bipyramid	Faces: 10 triangles Edges: 15 Vertices: 7 Vertex configuration: 4(32.4) (34) Symmetry group: D_{5h}, [5,2], (*225), order 20 Rotation group: D_5, $[5,2]^+$, (225), order 10 Dual polygon: pentagonal prism Properties: convex, face-transitive, (deltahedron)	Pentagonal bi-pyramid is expected to suit the required shape and will support up to 5 different type of keyboards. But some of its properties cannot be followed and hence cannot be reproduced completely

Cartesian coordinate system: A three dimensional Cartesian coordinate system, with origin O (0, 0, 0) and axis line X, Y and Z where X represents the dwell time, Y represents the pressure (for multiple keyboards) and Z is the tolerance factor respectively. The training data (dwell time, pressure) to be plotted is explicitly obtained from the user during registration. The user can add support for multiple type of keyboards by providing the respective training data.

Finding the Base Area: The values obtained from the user is plotted in a 2-d plane to form a polygon and its base area is calculated.

$$\text{Base Area (A)} = \left| \frac{(x_1y_1 - y_1x_2) + (x_2y_3 - y_2x_3) \cdots + (x_ny_1 - y_nx_1)}{2} \right| \quad (97.1)$$

where x_1 is the x coordinate of vertex 1 and y_n is the y coordinate of the nth vertex etc. and $3 < n < 5$. Notice that the in the last term, the expression wraps around back to the first vertex again.

Finding Centroid of Base: The centroid of a non-self-intersecting closed polygon defined by n vertices (x_0, y_0), (x_1, y_1),…, $(x_n - 1, y_n - 1)$ is the point (C_x, C_y), where

$$C_x = \frac{1}{6A} \sum_{i=0}^{n-1} (x_i + x_{i+1})(x_iy_{i+1} - x_{i+1}y_i) \quad (97.2)$$

$$C_y = \frac{1}{6A} \sum_{i=0}^{n-1} (y_i + y_{i+1})(x_iy_{i+1} - x_{i+1}y_i) \quad (97.3)$$

In these formulas, the vertices are assumed to be numbered in order of their occurrence along the polygon's perimeter, and the vertex (x_n, y_n) is assumed to be the same as (x_0, y_0). Note that if the points are numbered in clockwise order the area A, computed as above, will have a negative sign; but the centroid coordinates will be correct even in this case.

Height: The two dimensional base is extended into three dimensions by drawing perpendicular lines (w.r.t XY plane) with height h with the centroid as one of its vertices. The height is 2.54 % of base area where 2.54 is the crossover error rate in keystroke dynamics [12].

$$\text{Height (h)} = 0.0254 * \text{Base_Area} \tag{97.4}$$

Representing Model and setting the tolerance level: The 3-d model obtained as a result of extension should be represented as equations. The model is represented as collection of planes and the corresponding plane equations are represented with the help of three point A, B, C per plane. The Cartesian equation of a plane is $a \cdot x + b \cdot y + c \cdot z + d = 0$, where (a, b, c) is the vector that is normal to the plane.

Three points (A, B, C) can define two distinct vectors AB and AC.

$$\overrightarrow{AB} = (x_B - x_A)\hat{I} + (y_B - y_A)\hat{J} + (z_B - z_A)\hat{K} \tag{97.5}$$

$$\overrightarrow{AC} = (x_C - x_A)\hat{I} + (y_C - y_A)\hat{J} + (z_C - z_A)\hat{K} \tag{97.6}$$

Cross product of two vectors is calculated to obtain the normal vector. One point is substituted into the Cartesian equation to solve for d.

$$\overrightarrow{AB} \times \overrightarrow{AC} = \text{NORMAL VECTOR } (a, b, c) \tag{97.7}$$

$$-d = a \cdot x + b \cdot y + c \cdot z \tag{97.8}$$

This set of equations obtained as a result is stored in the cloud for classification. The area enclosed by the different planes form the acceptance region. The tolerance level is the value of z where the dwell time (x) and pressure (y) obtained from the user would be used to classify. Initially it is set to 0.

Classification: The dwell time (x) and the pressure (y) during the next login is taken as input. z is the current tolerance value. Classification is done by calculating the position of the given point (x, y, z) with respect to the plane equations obtained in the previous step.

$$\left.\begin{array}{l} a \cdot x + b \cdot y + c \cdot z + d > 0 \\ a \cdot x + b \cdot y + c \cdot z + d < 0 \end{array}\right\} \tag{97.9}$$

If the classification is successful and if the x value is more than the previous x value the tolerance level is incremented by 0.1 else it is decremented by 0.1. If the classification is not successful the tolerance level remains same. In case of the point (x, y, z) failing to classify for a particular tolerance level z, the next adjacent level from either side will be considered for classification but the tolerance level will not be updated.

Shifting Plane Equation after Login

$$\left.\begin{aligned} X_{new} &= X_{old} + k \\ Y_{new} &= Y_{old} + k \\ Z_{new} &= Z_{old} + k \end{aligned}\right\} \tag{97.10}$$

Where k is the average time spent and is calculated as the ratio of total time spent to the number of logins. The corresponding equation are updated.

$$k = \frac{\text{Total_Time_Spent} \quad (\text{min})}{\text{Number_of_Logins}} \tag{97.11}$$

97.4 Results and Discussion

Based on the different shapes analyzed, there is a need to model a shape that should satisfy the requirements. Different type of keyboards have different dwell time and pressure values for the same user. This paper considers dwell time as the base and form an n-sided polygon $3 < n < 5$. The reason for choosing a polygon over the fixed shape is because of the variation in the number of the points and the corresponding shape properties. Once the base is fixed there is a need for including a third factor of tolerance to accommodate mood variations and physical weariness. Based on the reference it has been found that the tolerance of keyboard dynamics is 2.54 %. So the shape is extended (with equal scale) to third dimension to a height of 2.54 % of the base area on both sides. The tolerance level is updated after every login based on certain criteria and the classification is done on a particular tolerant level with the dwell time and pressure as the input. Thus a unique classification model pertaining to the user's behavior is obtained as a result of this process (Fig. 97.2).

Updating the tolerant level would assist in recognizing the exact tolerant level for a particular user. This pattern can be further analyzed and can be reused to classify the nature of the mood variations of a particular user. Predicting the nature of mood variation can be utilized for implicitly invoking applications that would seek to stabilize this mood shift. This classification model can be further optimized using curve fit but obtaining the corresponding equations would be challenging. The efficiency of this system relies on the precision of the training data. Further research can be done to modify this semi-supervised system into an unsupervised learning system for better interoperability.

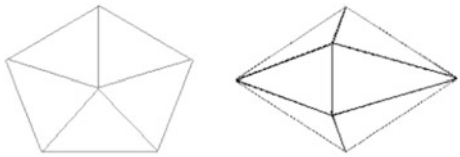


Fig. 97.2 Rough representation of the shape

97.5 Conclusion

At present the keyboard modality need to address various problems pertaining to: Support for different type of keyboards, the biometric data aging where the input provided by the user varies over time, Interoperability and Robustness. This paper presents a new three dimensional model for classification based on the parameters involved to address these current challenges. Our approach allocates a unique space for each user and a model mapped to classify the data based on input every time the user logins. This model leverages the existing models due to:

- Dynamic model positioning approach proves difficult to determine the acceptance region.
- Tolerance region scaling has proved optimal to address the biometric data aging and user mood factors.
- Dynamic base provides the user the option to configure the model for different type of keyboards.
- Compact shape of the model increases the accuracy of classification. Our analysis suggests that our technique is simple, stable and viable to practice. Further research includes dynamically altering the shape of the model based on feeding back minor error rates into the system.

References

1. Rudrapal D et al (2013) Internal attacker detection by analyzing user keystroke credential. Lecture Notes on Software Engineering 1(1)
2. Chang T-Y et al (2011) User authentication using rhythm click characteristics for non-keyboard devices. In: Proceedings of the 2011 international conference on Asia agriculture and animal (IPCBEE)
3. Bergadano F, Gunetti D, Picardi C (2002) User authentication through keystroke dynamics. ACM Trans Inf Syst Secur (TISSEC) 5(4):367–397
4. Monrose F, Rubin A (2000) Keystroke dynamics as a biometric for authentication. Future Gener Comput Syst 16(4):351–359
5. Everitt RAJ, McOwan PW (2003) Java-based internet biometric authentication system. IEEE Trans Pattern Anal Mach Intell 25(9):1166–1172
6. Boechat GC, Ferreira JC, Carvalho EC (2006) Using the keystrokes dynamic for systems of personal security. Trans Eng Comput Technol. Enformatika 18(1):200–205
7. Kekre HB et al (2011) Keystroke dynamic analysis using relative entropy & timing sequence Euclidian distance. In: Proceedings of the international conference and workshop on emerging trends in technology, ACM

8. Joyce R, Gupta G (1990) Identity authentication based on keystroke latencies. Commun ACM 33(2):168–176
9. Johansen UA, Keystroke dynamics on a device with touch screen
10. Song DX, Wagner D, Tian X (2001) Timing analysis of keystrokes and timing attacks on SSH. In: USENIX security symposium, vol 2001
11. Gaines R, Press S, Lisowski W, Shapiro N (1980) Authentication by keystroke timing. Rand report R-256NSF, Rand Corporation
12. Kumar ST, Priyanka US, Abinaya K (2013) Article: towards an approach for improved security in wireless networks. In: IJCA proceedings on international conference on innovation in communication, information and computing 2013 ICICIC 2013(1):23–27, March 2013. Published by Foundation of Computer Science, New York, USA
13. Adler A, Youmaran R, Loyka S (2005) Information content of biometric features. In: Proceedings of biometrics consortium conference
14. Stefan A (2000) The base-rate fallacy and the difficulty of intrusion detection. ACM Trans Inf Syst Secur (TISSEC) 3(3):186205
15. Ilonen J (2003) Keystroke dynamics. In: Advanced topics in information processing–lecture (2003), 03–04
16. Ashbourn J (2000) Biometrics: advanced identity verification. Springer, Berlin
17. Kanade S et al (2009) Application of biometrics to obtain high entropy cryptographic keys. In: Proceedings of World Academy of Science: Engineering and Technology, vol 52 (2009)
18. Adler A, Youmaran R, Loyka S (2009) Towards a measure of biometric feature information. Pattern Anal Appl 12(3):261–270
19. Uludag U et al (2004) Biometric cryptosystems: issues and challenges. Proc IEEE 92(6):948960
20. Li Y et al (2011) Study on the BeiHang keystroke dynamics database. In: IEEE international joint conference on biometrics (IJCB)
21. Teh PS, Yue S, Teoh ABJ (2012) Feature fusion approach on keystroke dynamics efficiency enhancement. Int J CyberSecurity Digital Forensics (IJCSDF) 1(1):20–31
22. Banerjee SP, Woodard DL (2012) Biometric authentication and identification using keystroke dynamics: a survey. J Pattern Recogn Res 7:116–139
23. Shanmugapriya D, Padmavathi G (2009) A survey of biometric keystroke dynamics: approaches, security and challenges. arXiv preprint arXiv:0910.0817
24. Mandujano S, Soto R (2004) Deterring password sharing: user authentication via fuzzy c-means clustering applied to keystroke biometric data. In: Proceedings of the fifth mexican international conference in computer science, 2004. ENC 2004, IEEE
25. Yu E, Cho S (2004) Keystroke dynamics identity verification—its problems and practical solutions. Comput Secur 23(5):428–440
26. Ashbourn J (2000) The distinction between authentication and identification. Paper available at the Avanti biometric reference site. homepage.ntlworld.com/avanti
27. Axelsson S (2000) Intrusion detection systems: a survey and taxonomy, vol 99. Technical report
28. Roberts C (2006) Biometric technologies fingerprints. Publisher name and location are missing, pp 1–23
29. Jain AK, Flynn P, Ross AA (2007) Handbook of biometrics. Springer, USA. ISBN ISBN-13: 978-0-387-71040-2
30. Bergadano F, Gunetti D, Picardi C (2002) User authentication through keystroke dynamics. ACM Trans Inf Syst Secur (TISSEC) 5(4):367–397

Author Index

A
Ahn, Kyung Jin, 305
Almuairfi, Sadiq, 11
Alnumay, Waleed S., 1
Aritsugi, Masayoshi, 169

B
Bae, NamJin, 669, 833
Biçakçi, Mert, 65
Bong, Wonsun, 453, 505
Bu, Heehyung, 37
Bui, Tuan-Anh, 211
Burgstaller, Bernd, 719
Byun, Yun Sang, 515

C
Çakir, Mehmet Ulaş, 65
Chae, Seung-Hoon, 677
Chang, Beom-Hwan, 535
Chang, Ku-Young, 347
Chatterjee, Pushpita, 1
Chen, Guey-Shya, 337
Chen, Hsing-Lung, 119
Chen, Huaming, 321
Chen, Ming-Sze, 119
Chilamkurti, Naveen, 11
Cho, Hong, 763
Cho, Kyungeun, 597, 661, 731
Cho, Seoungjae, 597, 661, 731
Cho, YongYun, 669, 833
Choi, Byong-Sam, 853
Choi, Hyung Rim, 863
Choi, Jung-Won, 103, 111
Choi, Kyu-Jin, 447
Choi, Mi-Jung, 575
Choi, Young-Hwan, 261
Chun, Se-Hak, 853
Chung, Kidong, 297
Chung, Tai-Myoung, 535, 553
Chung, Yongwha, 677

E
Elovici, Yuval, 439
Eun, Kim Gyoung, 707

F
Fang, Yung-Chin, 127

G
Gao, Yuxiang, 127
Ghosh, Uttam, 1
Goswami, Kalyan, 271
Guri, Mordechai, 439

H
Ham, Hyo-Sik, 575
Han, Youngshin, 43
Hang, Su, 719
Heo, Gyunyoung, 87
Hong, Bonghee, 203
Hong, Chee Ken, 875
Hong, Dowon, 347
Hong, Gwang-Soo, 821
Hong, Sangjin, 799
Hong, Soon Goo, 863
Hong, Youn-Sik, 75
Hoong, Poo Kuan, 875
Hoseini, Zaniar, 289
Hu, Jun, 321
Hu, Shu-Hua, 119

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7, © Springer Science+Business Media Dordrecht 2014

Huang, Jiun-Long, 49
Huang, Yi-Hung, 329
Huh, Eui-Nam, 211
Hung, Jason C., 321, 329
Hung, Pham Phuoc, 211
Hur, Junbeom, 229
Hwang, Eenjun, 261
Hwang, Kwang-il, 495
Hwang, Soon-Hak, 853
Hwang, Won-Joo, 21

I
Im, Subin, 615

J
Jang, Si-Hwan, 57
Jang, Uijin, 473, 565
Jang, WonAng, 247
Jang, Won-Chul, 177
Jang, Yu-Jong, 643
Jeong, Chisoon, 719
Jeong, Hohyeon, 739
Jeong, Hwa-Young, 313
Jeong, In-Kyu, 153
Jeong, Jaehoon (Paul), 739
Jeong, Kanghun, 423
Jeong, Young-Sik, 313, 467
Jhang, Kyoung-Sun, 525
Jho, Nam-Su, 347
Jin, Daxing, 597
Jo, Woo-Seung, 447
Joo, Sang-Hyun, 401, 57
Jung, ByungOk, 587
Jung, Daeyong, 483
Jung, Euihyun, 605
Jung, Jaehee, 305, 785
Jung, Jai-Jin, 313
Jung, Woo-Tak, 95

K
Kang, Dong-Ho, 525, 535
Kang, Min-A, 447
Kang, Myeongsu, 135, 153, 161, 177, 193, 393
Kang, Won Min, 313
Karumathil, Aadarsh, 885
Kim, Byeong-Woo, 763
Kim, Byoung-Koo, 525, 535
Kim, Byung Wook, 29
Kim, Byung-Gyu, 271, 289, 29, 821
Kim, Byun-Gon, 103, 111
Kim, Cheol-Hong, 153, 193
Kim, Cheonshik, 697
Kim, Daehoon, 261
Kim, Daekeun, 297
Kim, Dong-Hoon, 75
Kim, Hwa Jong, 357
Kim, Hwan-Hee, 575
Kim, Hyun Jong, 863
Kim, Hyunjoo, 553
Kim, Hyun-jung, 383
Kim, Hyun-Woo, 467
Kim, Ikkyun, 553
Kim, Jaehyo, 615
Kim, Jaekwon, 43
Kim, Jeong Ah, 409
Kim, Jeong Nyeo, 279
Kim, Jinuk, 615
Kim, Jonghwa, 37
Kim, Jong-Myon, 135, 153, 161, 177, 193, 393
Kim, Juyoung, 605
Kim, Kwan-Woong, 103, 111
Kim, Kyungsup, 95
Kim, Myung Ho, 365
Kim, Myung-Sup, 575
Kim, Sanghoon, 423, 687
Kim, Sungho, 37
Kim, Sung-Soo, 57
Kim, Taeyoung, 43
Kim, Yong Cheol, 453, 505
Kim, Yong-Kab, 103, 111
Kim, Young-Hyuk, 221
Kitasuka, Teruaki, 169
Ko, Kwangman, 843
Ko, Young-Woong, 853
Koh, Byoung-Soo, 587
Koo, Dongyoung, 185, 229
Koo, Insoo, 145
Krasner, Anatoly, 439
Kwak, Jin, 515, 643
Kwak, KyungHun, 669, 833
Kweon, JuWan, 247
Kwon, Gihwon, 415, 431
Kwon, Jihye, 615
Kwon, Ki-Ryong, 21
Kwon, Ohmin, 185
Kwon, Ryoungkwo, 431

L
Lee, Chang-hun, 383
Lee, ChangYong, 247

Lee, DoHoon, 247
Lee, Eung-Joo, 21
Lee, Eunseok, 739
Lee, GiYeol, 809
Lee, Goo Yeon, 357
Lee, Gwanghyung, 253
Lee, Hae-Yeoun, 401
Lee, Heung Ki, 305, 785
Lee, HwaMin, 809
Lee, Hyun Mi, 863
Lee, Jae Dong, 313
Lee, Jae-Gwang, 221
Lee, Jae-Kwang, 221
Lee, Jae-Pil, 221
Lee, Jeesoo, 687
Lee, Ji-Won, 401
Lee, Jong-Hyeok, 271
Lee, Jongsik, 43
Lee, Jung-Ryun, 773
Lee, Jung-Won, 739
Lee, Ki Young, 75
Lee, Kye-Shin, 289
Lee, Kyu-Chul, 447
Lee, Kyungmin, 203
Lee, Min-Feng, 337
Lee, Sang-Young, 461
Lee, Seongyoon, 615
Lee, Seung-Hwan, 751
Lee, Suk-Hwan, 21
Lee, Yong, 357
Lim, Hyungmin, 473
Lim, Il-Kown, 221
Lim, JongBeom, 483
Lim, Joonhyung, 605
Lin, Kuan-Cheng, 329
Lin, Yung-Tso, 329
Liu, Hongzhe, 731
Lu, Yuqian, 237

M
Mafole, Prosper, 169
Manabe, Yuki, 169
Moon, Chang Hyun, 313
Moon, Hae-Min, 677
Moon, Hyeonjoon, 423
Moon, Nammee, 799

N
Na, Jung-Chan, 525, 535
Nam, Sung-wook, 495
Namgung, Hyun, 221
Nasridinov, Aziz, 653

O
Oh, Jung-Min, 799

P
Pan, Sung Bum, 677
Park, Chang-hyun, 863
Park, DooSoon, 809
Park, JinGwan, 247
Park, Jong Hyuk, 313
Park, Keon-Jun, 103, 111
Park, Minoh, 43
Park, Namje, 545
Park, PyungKoo, 373
Park, Sachoun, 415
Park, Sang-Hyun, 773
Park, Su-Wan, 279
Park, Yong-Hun, 135
Park, Young-Ho, 653
Piao, Yongxu, 203

Q
Qi, Ji, 321

R
Rew, Jehyeok, 261
Rho, Seungmin, 261
Ryong, Beak, 821
Ryou, JaeCheol, 373, 95
Ryu, HoYong, 373

S
Senthil Kumar, T., 885
Seo, Junsang, 161, 193
Seo, SangHo, 587
Seok, Seung-Joon, 751
Shao, Qun, 237
Shin, Dongil, 253
Shin, Dongkyoo, 253
Shin, JinSeop, 373
Shin, Jinsoo, 87
Shin, Kwangmu, 297
Shin, Yongjoo, 185
Shon, Dongkoo, 161
Singh, Chirpreet, 237
Son, Hanseong, 87
Son, Yun-Hee, 447
Song, Bok-Hee, 633
Ssin, Seungyoub, 661
Stap, Cindy, 127
Suh, Changjin, 473, 565

Sung, Yunsick, 597
Suresh, Abhijit, 885

T
Tan, Ian K. T., 875

U
Uddin, Jia, 153
Um, Kyhyun, 597, 661, 731

V
Veeraraghavan, Prakash, 11
Vu-Van, Hiep, 145

W
Wang, Jen-Chieh, 337
Wang, Jun-Zhe, 49
Wang, Xuan, 321
Wen, Low Zhi, 875
Weon, Il-Yong, 383
Whang, Sun-Min, 633
Wong, Kok-Seng, 365
Woo, Kim Byeung, 707

X
Xi, Yulong, 731
Xu, Xun, 237

Y
Yang, Ching-Nung, 697
Yang, Lei, 321
Yang, Seung-Hun, 821
Yang, Zong-Hua, 49
Ye, Xinfeng, 237
Yi, Gangman, 305, 785
Yoo, SeongMin, 373
Yoon, Hyunsoo, 185, 229
Yoon, Sangeun, 653
Youn, Taek-Young, 347
Yu, Heonchang, 483
Yun, Han-Kyung, 633

Z
Zhang, Jie, 357
Zhou, Qingguo, 321
Zhou, Rui, 321

Subject Index

A
1-hop neighbor nodes, 774, 776, 778, 779, 780
2D-3D conversion, 402, 406
3.20 Cyber Incident, 554
3D face representation, 423. *See also* Face representation
  average face model, 426
  database, 428
  ICIA algorithm, 427–428
  samples, 426, 427
  scans, 424, 425
3-way chaining authentication, 76, 77, 84
4 × 4 network, 60*t*
5 × 5 network, 59, 61*f*
5 M pixel CMOS Camera, 688
64 bit block encryption, 227
7 × 7 network, 59–60, 61*f*
8/4 Hamming code, 25
802.1aq Shortest Path Bridging, 566
9 × 11 network, 59, 61, 62*f*
Abnormal behavior detection, 554, 556
  cyber-targeted attack response system, 556–557
  large-scale data analysis engine, 557–560
  using MapReduce, 560, 561*f*
Absolute temperature (AT), 250
Accelerator pedal position, 710, 712, 713, 714*f*
Access control model, 517, 520
  access control module, 520
  applicability, 522, 523*t*
  application of policy, 522, 523*t*
  components of, 521*t*
  context controller, 521
  context information, 517–518
  data transmittance, 520
  flexible assignment, 522, 523*t*
  immediate response, 522, 523*t*
  information transmittance, 520
  policy management system, 521
  problem analysis, 518–520
  review process, 522
  role management system, 521–522
  user data, 520
Access control techniques, 522
Access point (AP), 174, 441, 444, 453, 454, 507
  3 × 3 access points, 471
  Wi-Fi access points, 506, 513
Access Points Routers, 444
ACI 318 Code 1995, 816–817, 819
ACK storm, 2, 6
Active Contour Models (ACM), 679
Active pixel (AP), 424
Active set top box
  ActiveSTB, 308
  MP4 versus H.264 scalable streaming, 307–308
  scalable multimedia, 309–310
  scalable stream, 309*t*, 310*t*
  simulation environment, 309*f*
Active Shape Model (ASM), 680
ActiveSTB, 307, 308
  NS-2 simulator, 308
Activity Based Access Control (ABAC), 517, 519
Ad hoc TCP (ATCP), 6, 7
Adaboost and Haar-like algorithm, 384–385, 388
Adaboosting, 769, 790*t*
Adaptive media streaming, 306
Adaptive navigation system. *See* Self-Adaptive navigation system (SeNS)
Adaptive normalized cross correlation (ANCC), 298–299
Adaptive smoothing (AS), 454, 455, 458, 510, 511*f*
  original noisy signal, 456*f*
Adaptive streaming scheme, 306

J. J. (Jong Hyuk) Park et al. (eds.), *Frontier and Innovation in Future Computing and Communications*, Lecture Notes in Electrical Engineering 301,
DOI: 10.1007/978-94-017-8798-7,

© Springer Science+Business Media Dordrecht 2014

Adaptive support weight (ASW), 298
Adhesion force, 688, 690, 692, 695
Adhesion techniques, 688
Ad-hoc routing. *See* VANET (Vehicular Ad-hoc Network)
Advanced Driver Assistance System (ADAS), 763, 764, 771
Advanced emergency braking system (AEBS), 764, 768, 771
  braking force of, 768*t*
  command, 769*f*
  safe distance secured by, 770*f*
AEEC (Airlines Electronic Engineering Committee), 416
AES-256, 225
  data encryption conversion code, 226*f*
Affective computing, 338
  emotion classification, 339, 340*t*
  emotions, 339
  fuzzy cognitive mapping, 340
  Thayer's emotion model, 339–340
Ahead-of-time (AOT) decompression, 720
Air and fuel ration (AFR), 249
AIS, 545
Alcohol consumption and skin aging, 262, 264–265, 266
Alphanumeric passwords, 12
Amazon EC2, 483
  resource types, 489*t*
Android devices, 580
Android malware, 577*f*
Android Market, 474
Android phone, 254
Android systems, 314
  operating systems, 315
Android vaccine, 479–480
  in AV-Test, 480*t*
Angular misalignment fault (AMF), 179
ANSI/ISA-99 international standard, 536
Anymon Plus, 844
AODV (Ad hoc On-demand Distance Vector), 773, 774
App Store, 133, 474
  App for administrative works, 474
  App verification system of, 478
  illegal copy prevention system of, 478–479
  illegal copy prevention technology by, 479*t*
  verification items by, 479*t*
Application layer access control block, 540
Application Programming Interface (API), 434, 673
  App Store, 478–479
  data publishers, 448
  OS kernel, 417
  pull-based, 547
  push-based, 547
  resource information, 579
  twitpic, 880
  for Wi-Fi Direct, 444
Application-specific integrated circuits (ASICs), 136
APT (Advanced Persistent Attack), 554, 556, 557, 562
Arden syntax, 409
Area defuzzification method, 149
ARINC 653 Standard, 416
  ARINC 653 APEX (Application/Executive), 416–417
  core module component relationships, 417
ARPA, 545
Array control unit (ACU), 142, 195, 196
Artificial intelligence (AI), 238, 598, 599, 662
  domain, 338
Artificial Neural Network (ANN), 634, 635
Attitude Heading Reference System (AHRS), 616
  specification of, 618*t*
Attraction force generators, 688
Audio classification, 393
  GA operators' generation, 397–398
  GA-based FCM for, 396–397
  initialization, 397
  result, 399*t*
Audio features extraction, 394–396
  checking termination criterion, 398
  feature vector configuration, 396
  features distribution, 395*f*
  low root-mean-square ratio, 394
  spectrum spread, 394
  zero-crossing rate, 395–396
Audit mechanism, 186
  delegation of, 188
  for outsourced data, 188–189
Augmented Reality (AR), 467–468
Authentication signal, 38–39
Authentication technology. *See also* Log authentication system
  introduction to, 590
  PKI authentication service, 591
AuthentiGraph, 13
Auto-configuration, MRA transform, 680–681
AutoFACT, 787
Automatic annotation processing, 786

Automobile detection algorithm
for blind spot detection, 830, 831
cascade-type classifier, 826*f*
data training process, 825
image according vehicle distance, 826*f*
pre-processing image classification, 828*f*
using proposed feature training, 827*f*
shadow area extraction method, 828*f*
verification process, 825–826
Automobile identification algorithm, 822
Automotive simulation models (ASMs), 708
Autonomous drive path planning, 732
Awase-E system, 13
Azimuth, 766
CS_{ego} transformation by, 767*f*

B
Back propagation (BP), 91, 104, 105
of BN, 90, 92
Background subtraction, 162–163, 877
accuracy measurements, 881, 882*f*, 883*f*
in car park management, limitations, 881
example of, 879*f*
Backoff-free fragment retransmission (BFFR), 172–173
energy consumption, 175*f*
energy efficiency, 174*f*
with no contention overhead, 174*f*
Backpacking, 322, 324
Backpropagation neural network (BPNN) model, 140, 340–341, 341*f*
mapping of, 141
process, 342
Basal metabolic rate (BMR) estimation, 264
Base area, 889
Base station, 76
Basic routing protocols, 773
Basic safety message (BSM), 765
Basic stat module, 847
Basic Station Set (BSS), 359
Bayesian network (BN), 88, 92, 93
Bayesian probability-based imitation learning, 600
Beacon scheduling, 496–497
completion time, 501–502
slot collision avoidance, 497*f*
Beacon-enabled mode, 495
Bearing fault (BF), 179
Behavior execution, 661–662
basic task generation, 663–664
inoperability of server and 2D camera, 664–666
server protocol for task execution, 666–667
Behavior frequency table by status, 600, 601*t*
Behavior-based detection, 576, 577
Bellman-Ford (BF) algorithm, 154
connected routers with different link weights, 156*f*
CPU/GPU implementation of, 158*f*
distance vector routing algorithm, 155
GPU-based, 156–158
parallel implementation of, 158*f*
performance evaluation, 158–159
using CPU, 157*f*
using GPU, 157*f*
Big data, 554
abnormal behavior detection based on, 556
analysis, 844, 851
security research using, 555
Bilinear-pairing, 3
Biometric authentication, 18, 383, 886
Biometric entropy, 886
Biometric system, 12
Biomimetic approaches, 688
Biosignal based interface
comparison with existing pointing input device, 623, 628–630
interface performance evaluation with EMG intensity, 621, 623
movement path, 622*f*
moving distance, 625*f*
position manipulation performance of interface, 620–621
position-control performance model, 622*f*
repeated modified model, 622*f*, 626*f*
Bi-stable Processor (BP), 88
BIT (Biology Information Technology), 22
Bit error rates (BERs), 2, 26, 31, 170, 777
in mutated codon sequence, 26*f*
Bits-per-pixel (bpp), 705
Blast2GO, 787
Blind spot, 764, 822
areas, 823*f*, 830
Block-level deduplication, 230
Bluetooth, 248, 475, 476*t*
messages, 444
tethering, 741
BMS-WebView-1, 55, 56*f*
Bowed shaft fault (BSF), 179
Bridge priority masking, 566
cyclic priority matrix, 569–571
and priority matrix, 566–568
problem definition, 569
Bring-your-own-device (BOYD), 128
Broken rotor bar fault (BRF), 179
Bucket boundaries, 348
Bucket index ordering (BIO) problem, 349

Bucket index ordering (BIO) problem (*cont.*)
bucket boundary determination, 352
counter measure, 353
cyclic query, 354, 354*f*
in existing bucketing schemes, 352–353
random query, 353
in range query, 350–351
Bucket-based database management techniques, 348
Bucket-based index, 348
adversarial model, 349
bucket construction and data query, 348–349, 351*f*
random index assignment, 351*t*
Building recognition method, 505
range of search space, 506*f*
Bytecode compression methods, 720
decompression of, 720

C

C/C++ language, 759
Caroo pro, 248
CarSim, 764
Cartesian coordinate system, 46, 765, 889
Cartesian equation, 890
Cat swarm optimization (CSO), 330, 331–333
seeking mode, 331–332
tracing mode, 332–333
CA-TRBAC, 519
CEB-FIP Model Code 1990, 814–815
Central processing unit (CPU)
-based implementation and GPU, 159*t*
hardware specification of, 157*t*
over GPU, speedups of, 160*f*
Centroid of base, 889
Challenge-Response protocol, 186
Channel bandwidth, 775
Chest CT image, lung segmentation of, 679
histogram of, 682*f*
image segmentation using MRA Level-set, 681*f*
volume data of, 683, 685
Choregraphe, 662
CIE-rg chromaticity space, 120–121, 121*f*
CINDER (Cyber-INsiDER), 555
CKAN (Comprehensive Knowledge Archive Network), 448
functional structure of, 449*f*
semantic web layer cake, 449*f*
Class meta-model element, 462–463
graphical representation of, 463*f*
Classification model, 887
cross product of vectors, 890
shape, rough representation, 892*f*
shifting plane equation, 891
three dimensional shapes, 889*t*
tolerance level, 890, 891
two dimensional shapes, 888*t*
Classification, 330, 890
emotion classification, 339
GA-based FCM for audio classification, 396–397
Client and devices, 439, 440*f*
broadcasting capabilities, 441
validator devices, 441
Climbing robots, 687
Clinical decision supporting (CDS) system, 409
system architecture for, 412
in u-health portal, 411
Cloud, 887, 890
Cloud Computing (CC), 44, 48, 128, 186, 212, 323, 483, 854
cloud customers (CCs), 212
cloud data center, 516
cloud manufacturing, 238
Cloud Provider layer, 213
cloud security, 374
customer layer, 213
implementation and analysis, 218–219
problem formulation, 215–218
resource reallocation in, 45–46
security assessment, 380–381
system architecture, 213–215, 214*f*
Cloud service pricing, 854
monopoly provider, 855
Cloud Service Providers (CSPs), 186, 374, 375, 381
storage service, 186
Cloud services, 322, 323, 374, 854
Cloud storage technique, 186, 374
bilinear map, 187
cloud service provider (CSP), 231
computation cost, 190*t*
computation overheads, 233, 234*t*
construction, 188–189
data uploader, 231
deduplication over encrypted data (*see* Deduplication over encrypted data)
definition, 187–188
notation, 190*t*
performance evaluation, 190–191
security analysis, 189–190
storage overheads, 233, 234*t*
tag-generation time, 191*f*
updater, 231
verification time, 191*t*

Cloud-based approach, tool selection
automatic query generation, 241–242
implementation, 243–244
resource virtualization, 240–241
suitable tool, selecting, 242–243
tool matching scheme, 239–240
Cluster estimation. *See* Subtractive clustering method
Cluster formation, 80–81
Cluster-head (CH) node, 80
Clusters, 81*f*
CMOS 0.35 μm technologies, 290
CMOS 65 nm technologies, 290
Co-creation, 864
cases, 865*t*
Co-creation platforms
block diagram of, 869*f*
components of, 869*t*
conceptual diagram of, 868*f*
design, 864–866
functions of, 866, 867*t*
product development, 871–873
product planning, 870–871
sequence diagram, 872*f*, 873
structure and components, 868–870
Coding DNA sequence (CDS) watermarking, 22
goals of requirements, 23
Coding sequence watermarking, 22
Coding tree block (CTB), 272
Coding unit (CU), 272
Codon numerical mapping, 24f
Cognitive radio (CR) technology, 145
system model, 146–147
Cognitive radio sensor network (CRSN), 146
Cohen-Daubechies-Feauveau (CDF) biorthogonal wavelets, 25
Coincidence Processor (CP), 88
Cold chain management, 203
FCQ for (*see* Floating continuous query (FCQ))
RFID tag application in, 204*f*
sensor tag application in, 204*f*
Collective services, fabricated data to, 441
Collision, 733
Color pixel recognition, 162
Color segmentation, 119, 163–164
Command and Control (C and C) server, 607
Command filter, 541
Compact index (CI), 58
Compute Unified Device Architecture (CUDA), 154–155
C code, main function of, 159*f*
Computer communication environment, 843
Computer image analysis, 678
Computer numerical control (CNC) machine tools, 241
Computer Reservations Systems (CRS), 322
Computer vision, 800
-based fire detection, 162
problem, 505
Computer-Aided Process Planning (CAPP) systems, 238
Computerized Tomography (CT), 678
Congestion avoidance, 5
Congestion window (CWND), 3, 8
Congestion window limit (CWL), 3, 8
Constant bit rate (CBR), 174
Constructive sparse positive realization method, 98–100, 101
Content providers (CPs), 587
Contention Aware Scheduling (CAS) algorithm, 216
Contention free period (CFP), 496
Context information, 517–518
aggregator, 518
interpreter, 518
widget, 518
Context-aware control service, 834–835
abnormal status inference, 841*f*
collecting and categorizing information, 837–838
context deduction, 835
context modeling, 835
control service process, 839, 840*f*
control target, 838*t*
crop information, 837*t*
inference rule, 840
ontology model for, 838*f*
structure of, 835–837
system components, 836*t*
vertical farm environment, 836*f*
Context-aware GUI based logistics service, 670
Continuous active mode (CAM), 170
Continuous query, 203, 204
processing performance, 208
Control systems, 88, 96, 222, 225, 536, 537
blocking unauthorized access to, 540, 541
categorization of network attacks against, 538*f*
Control theory, 119
Conventional hardware routers, 154
Converged network adapter (CNA), 131
Cooperative safety system, 764
Co-privacy, 367
Coprocessor, 95

Correlation analysis, 554, 555, 556, 558, 560, 844
CPU branch target buffers (BTBs), 721
Crack width calculation program, 817*f*
Crack width formulation, 814
  ACI 318 Code 1995, 816–817
  calculation, comparison of, 818*f*
  CEB-FIP Model Code 1990, 814–815
  Eurocode 2, 816
  structural performance verification of JSCE, 816
CRAE (Co-Relational Analysis Engine), 557–558
Crisp rules, 104, 108
CSA (Cloud Security Alliance), 374
CSMA/CD protocol, 358
CSO-SVM, 330
CU splitting termination (CST), 272–274, 278
Cued-recall based technique, 13
Current time stamp-based travel time, 748*t*
Customer relationship management (CRM), 127
Cutting tools ontology, 238, 239, 241, 244
Cyber genome programs, 555
Cyber security, 87
Cyber security evaluation index (CSEI), 89
Cyber security risk model, 88
  for power plant RPS, 91*f*
Cyber threats, 528
  Application Protocol attacks, 529, 530*t*
  flow authentication, 532–533
  flow whitelist, 533*f*
  IP packet filtering, 529–530
  network intrusion detection, 530
  network protocol attacks, 528, 529*t*
  system authentication, 531, 532*f*
Cyber-attacks, 526, 537
  in SCADA networks, 528, 531, 534
Cyber-targeted attack, 554
  response system, 556–557
  structure of, 557*f*

D

Dalvik Machine, 584
Data anonymization, 366
Data collection, 365
Data editing module, 847
Data hiding, 697
  512 × 512 grayscale cover images, 704*f*
  issues with, 697–698
  matrix encoding, 698
Data import module, 847
Data integrity, 186, 191
Data mining techniques, 606, 607
Data sharing, 447–448
  and LOD module, 451*f*
  module definition, 450
  platform architecture, 450, 451*f*
  platform design, 449
Data transfer module, 124
Database management system (DBMS), 347
Daubechies 4-tap filter, 165
DC coefficient, 39
DC motor, 692
DCF inter frame space (DIFS), 171
DCT coefficient, 38
DE2-70 board, 124, 125
Decentralized workflow scheduling systems, 485
Decoder, block diagram of, 40*f*
Decoding procedure, 703
Deduplication over encrypted data, 231–233
  audit of outsourced data, 232–233
  initial upload, 231–232, 232*f*
  retrieval of outsourced data, 233
  subsequent upload, 232
  system setup, 231
  update of outsourced data, 232
Deduplication technique, 230
  bilinear Diffie-Hellman (BDH) assumption, 230
  bilinear map, 230
Deep packet inspection (DPI), 537
Default access control filter, 542*f*
Default filter, 540
Delay time in routing process, 783, 783*t*
Denial of Service (DoS) attack, 89
Denoising scheme, 148–150
Depth-image-based rendering (DIBR), 401, 402
Description logic (DL) queries, 241, 242
  generation procedure, 242
Detect accident (DA), 250
Deterministic and Synchronous Multi-channel Extension (DSME), 496–497
  enhanced DSME design, 499–500, 501*f*, 502*f*
  first experiment, 498–499
  revised algorithm, validity of, 501–502
  2nd revision, 500
DIBR video watermarking, 402
  embedding, 402
  fidelity test, 405
  robustness test, 406
  spatial embedding, 403–404, 403*f*
  temporal embedding, 404
  visual quality results, 406*t*

watermark extraction, 405
Digital forensic
investigation procedure, 645*f*
of SNS real time data, 644
Digital like design concept, 290
Digital optical processor, 96
Digital rights management (DRM), 37
three SD videos from, 41*f*
Digital signal processors (DSPs), 136
Digital video frame obfuscation method, 39–41
decoder, 40*f*
encoder, 40*f*
human visual system, spatial frequency response of, 40*f*
SD videos, 41*f*
Dijkstra algorithm, 325
Direct attached storage (DAS), 128
Directed Acyclic Graph (DAG), 215, 484, 787
sample, 216*f*
Discontinuity preserving smoothing (DPS), 454, 510
average error in localization, 512*f*
IDW-interpolated map, 511
and path loss model, 455*f*
reconstructed radio map, 511
Voronoi-interpolated map, 511
Discrete Cosine Transform (DCT), 638
Discrete phase-type (DPH) distribution, 99, 100
Discrete-state discrete-time Markov chain, 99
Discretionary access control (DAC) scheme, 517, 519, 523*t*
Distance evaluation technique (DET), 180, 181, 182
Distributed coordination function (DCF), 171, 173
Distributed Data Centers (DDCs), 378, 379, 380
Distribution function module, 848
DNA copyright protection, 22, 23
DNA crypt algorithm, 22, 25
DNA cryptography, 22
DNA sequence watermarking, 23–25
DNA steganography, 22
DNA Watermarking, 22
DNP3, 526, 530, 536, 537, 544
DO-178B certification, 416
DOM parser, 546, 548, 549
Domain Name System (DNS), 605
cache, 607
Korea National DNS, 607, 608*f*
log analysis, 609, 611, 612
logs, 606, 607, 608, 610, 612
query, 606
servers, 606, 612
Domestic Android App Stores, 478
Domestic simple security management, 844
DOM-parser-based EtriIVEF library, 546
DoS (Denial-of-Service) attack, 89, 528, 529*t*, 537
Drag gestures, 733, 734*f*, 735, 736*f*
Draw Sequential Direction (DSD), 468
direction cognition of, 469*f*
diverse locking systems, 469*t*
Draw-A-Secret technique, 13
Drivetrain, 711*f*, 716
Driving behavior, 248, 252
DRLES speed function, 684
DRSLE-based method, 680
DSDV (Destination-Sequenced Distance-Vector), 775
dSPACE, 708
DSR (Dynamic Source Routing), 773, 774, 775
Duqu, 536
DWT (discrete wave transform)
-based watermarking, 22
coefficients, 22, 23, 24, 26
on multi-core architecture, 141*f*
Dynamic analysis, 576, 844
-based detection, 577
using CarSim, 764
Dynamic datacenter, 128
Dynamic neural fields (DNF), 598
Dynamic random topology, 6–7
Dynamic time warping (DTW), 262
extrinsic aging factors and skin texture age, 268*f*
in measuring skin aging, 262, 266–267

E

E2BOX's EBIMU-9DOFV2 product, 617
Eagle mGuard, 537
Earliest Finish Time (EFT), 217, 218
Earliest Start Time (EST), 217
Early fire detection, 161
Early warning systems, 66, 70–72, 72*f*
ECDIS, 545
Eco-driving, 248, 252, 740
Edge movement, 694*f*. *See also* Robots
EEPR algorithm, 778
advanced, 778–779
examples, 779*f*, 780*f*
performance metrics, 782–783
simulation setup, 781–782
Eigenfaces, 388, 389, 390*f*

Electronic Stability Program (ESP), 710
E-mail, 468
Embedded image processing, 691–692
EMG sensor, 618
  abductor digiti minimi, 619*f*
  graphic input using, 621*f*
  and quasitension, 619*f*, 627*f*
EMG-based interface, 616
Emotion classification, 339, 340*t*
Emotions, 339
Empirical mode decomposition (EMD), 178, 698
e-Navigation, 545
Encapsulation, 588–589
Encryption application model, 222
Encryption key, 374, 377, 378, 379, 380, 381, 382
  key management, 375, 376
Energy detection, 146
Energy-efficient Probabilistic Routing (EEPR), 775, 784
  advanced algorithm, 778–781
  algorithm (*see* EEPR algorithm)
  metrics in, 776–777
  packet forwarding process, 776
  simulation, 781–783
Energy-Enterprise Security Management (E-ESM), 222, 223*f*
Engine volume efficiency (Engine VE), 249
Enterprise class solution, 128
Enterprise computing, 127–129
  App store challenges, 133–134
  components of, 129
  converged infrastructure, 128, 130, 131, 134
  deployment, 128, 131, 133, 134
  enterprise architecture, 129*f*
  Ethernet speed, 132*f*
  external storage trends, 132–133
  memory trend, 130*f*
  multiple roles/processes, 134
  network trends, 133
  resources allocation challenges, 133, 134
  RFS count, 132*f*
  segregated networking, 128, 131
  server trends, 129–132
  SSD capacity, 131, 132*f*
Enterprise security management (ESM), 222–223, 844
Enterprises' information systems, 864
Entity-relationship (E-R) diagram, 656, 658*f*
EntriVEF Pull Parser, 546
Entropy, 179, 181, 182, 316, 680, 886
  coding, 272
  password entropy, 19
Environment control, 834
Environmental pollution and skin aging, 262
EON guideline, 409
Equal-cost multi-path (ECMP) technology, 565, 566
Equal-cost multi-tree (ECMT), 566
Erbium-doped fiber amplifiers (EDFAs), 97
Error compensation, 802*f*, 804, 807
  gaze data processing, 805*f*
  gaze time and monitoring, 806*f*
Error detection, 390, 822
Esper/Storm, 557
Ethernet, 128, 131, 536, 692
  and ECMP, 566
  routing, 566
  speed, 132*f*
e-Tourism, 323
ETX (Expected Transmission Count) metric, 775, 776–777
  forwarding probability via, 779*f*
Euclidean distance, 46, 105, 454, 508, 886
Euler angle, 616, 617
Eurocode 2, 814, 816
Event mechanism model, 734*f*
Evil Twrin, 222
Exclusive region (ER), 30
  -based coverage of receiver, 32*f*
  -based scheduling, 31–32
  scheduling scheme, 30
Explicit Congestion Notification (ECN), 3, 4
Exploiting the modification direction (EMD), 698
EXPRESS-based data models, 239
  EXPRESS standard, shortcomings, 240
Extended inter frame space (EIFS), 171
Extension name, 464
Extracellular matrix (ECM), 262
Extrinsic (external) aging factors, 262
Eye tracking, 801, 804, 807
  errors, 808

F

Face labeling algorithm (FLA)
  candidate face region, 386
  continuous labeling between frames, 386–387
  labeling accuracy, 389–390, 390*f*
  procedure, 387–388
  results, 389*f*, 389*t*
Face region detection and face learning
  Adaboost and Haar-Like Algorithm, 384–385

linear discriminant analysis (LDA), 385
local binary pattern histogram (LBPH), 385, 388
principle component analysis (PCA), 385
Face representation, 423, 424
eigenspace, 427
eigenvalue, 427
eigenvectors, 427
ICIA algorithm, 427–428
and Max Planck Institute (MPI) results, 428*t*
optical and SIFT flow, comparison, 425*f*, 425*t*
shape information, 426
texture correspondence, 424
vertices correspondence and model generation, 424–427
Facial animation, 634
Fast Fourier transform
algorithm, 636
analysis, for sentence, 639*f*
Fast PU decision method, 272, 274–276
block motion complexity (BMC), 274
and HM reference software 10.0, 276*t*
rate-distortion (RD) curves, 277*f*
weight factor for, 275*f*
Fault classification, of induction motor, 177, 178
data acquisition, 178–179
data conversion, 179
feature extraction, by GLCM, 179–180
using SVR, 180–181
Fault signatures, 181, 182*f*
FFriction, 690
FGravity, 690
Fiber-optic signal processors, 97, 98, 100
Fiber-optical signal coprocessors, 96
Fibre Channel, 128
File-level deduplication, 230
Fingerprint Wi-Fi Radio Map, 453–454
average localization error, 458*f*
interpolation, 454–455
map resolution with RSS, 456–457
Fire detection algorithm, 139–140
color segmentation, 140*f*
four-stage algorithm, 140*f*
moving regions detection, 140*f*
2-level discrete wavelet transform, 141*f*
Fire detection method, 135
background subtraction, 162–163
classification, 165–166
color segmentation, 163–164, 164*f*
color variations using spatial wavelet analysis, 164–165
false negatives (FN), 166
MATLAB 2012b, 166
percentage of FN (PTN), 166
percentage of TP (PTP), 166
true positives (TP), 166
vision based, 136
Fire wall, 536
Fire-containing movies, normalized wavelet energies with, 165*f*
Firewall logs, 844, 851
Fisherfaces, 388, 389*t*, 390*f*
Fitt's law, 616, 620, 622*f*
Fixations, 803
Flame, 536
Floating continuous query (FCQ), 205–206, 209
decomposition, 206–207
performance evaluation, 207–208
processing flow, 207*f*
Flow filter, 541
Flow whitelist, 531, 533*f*
Force intensity, performance intensity, 624, 626–628
Ford-Fulkerson Algorithm (FFA), 745
Foreign-made log analysis, 844
Formants frequency, 633, 636, 638, 639, 640
FPGA, 120, 124, 125
Fragment retransmission, 172–173
FTraction, 690
Fuel efficiency model, 248
system flow diagram, 248–251, 249*f*
Fuel gauge level, 251*f*
Functional annotation, 786
data features, 787, 789*f*
non-binary form, 788
Saccharomyces cerevisiae, 787, 788
Functional gene annotation, 786
Fuzzy c-means, 194, 394
Fuzzy cognitive map (FCM), 338, 340, 341*f*
detecting emotion colors, 342*f*
Fuzzy interference rule, 148
Fuzzy logic, 44, 46
Fuzzy matching method, 635
Fuzzy neural networks (FNNs), 104
Fuzzy relation equation-based approach, 111
Fuzzy system
for spectrum sensing (*see* Spectrum sensing, fuzzy for)
system model, 146
for wavelet denoising, 150

Fuzzy variable, 47*f*
Fuzzy-based resource reallocation scheduling model (FRRSM), 44
architecture of, 46*f*
rule bases, 47*t*
Fvacuum, 690

G
GA-HNN3, 59
Gamma value correction, 298
Gateway authentication, 81*f*, 82
Gaussian distributed random variable, 747
Gaussian noise, 31
Gaussian random variable, 146
Gaze behavior analysis system, 800, 801
basic gaze data and monitoring results, 805, 806*f*, 807
data compression patterns, 807, 807*f*
experiment environment, 804–805
gaze time and monitoring, 806*f*, 807
object centered (*see* Object centered gaze behavior analysis system)
object-based gaze behaviors, 808
Gaze data processing process, 804
Gaze duration, 803
GBSR (Greedy Border Superiority Routing), 752, 753, 754*f*, 755
Gender and skin aging, 262
Gene annotation, 786
Gene expression, 786
Gene ontology (GO), 786, 787, 788, 795–796
dataset reconfiguration, 790
error rate applying several methods, 793*f*
error rate for each term, 794*t*
negative, 791, 792*t*
positive, 791, 792*t*
Generalized reactivity (GR1) formula, 431, 432
importance of 433
overall processes, 434*f*
General-purpose microprocessors (GPPs), 136
Generic-based back propagation neural network (GBPNN), 136
Genetic algorithms (GAs), 329
crossover, 397
mutation, 397–398
selection, 397
Geometrix FaceVision 200, 429
Georeferenced class, 463*f*
with temporality, 463*f*
Geo-spatial intelligence systems, 71
Gestures, 732
drag gesture experiment, 736*f*
kinds of, 734*f*
point gesture experiment, 735*f*
pseudo code for movement, 734*f*
recognition, 731
recognition rate analysis, 737*t*
scissors gesture experiment, 736*f*
slide gesture experiment, 736*f*
GLIF, 409
Global Distribution System (GDS), 322
Global Positioning System (GPS), 248, 441, 506, 516, 673, 674, 675, 765
-based location information, 671
-based position recognition technology, 670
for car parking information, 877
signaling channel, 440, 453
in traffic information, 741, 744
GMDSS, 545
GNSS application, 465, 466
GNU R, 844, 851
GOA (gene ontology annotation), 787
Google Android's Dalvik VM, 720
Google glass, 212, 213
GOP structure, 38
Gotcha, 787
GPSR (Greedy Perimeter Stateless Routing) protocol, 752, 753
greedy forwarding, 753
local maximum processing, 753
Graphical Models, 835
Graphical-based password techniques, 12
Graphics processing units (GPUs), 154, 194
-based Bellman-Ford (BF) algorithm, 156–158
-based software paradigm, 154
and equivalent CPU-based implementation, 159*t*
hardware specification of, 157*t*
multi-threading arrangement on, 155*f*
over CPU, speedups of, 160*f*
Gray level co-occurrence matrix (GLCM), 179–180
sensitivities of, 182*f*
Greedy forwarding, 753, 754*f*
Greenhouse cultivating environment, 834
Group of video coding experts (VCEG), 271
GUI-based situation information processing technology, 670
GUIDE, 409

GZTR (Geographical Zone Based Two-tier Routing) algorithm, 755
average hop counts, 760*t*
next zone selection algorithm, 755–757
packet loss rate, 760*f*
relay node selection algorithm, 757–758
throughput rate, 759*f*

H
H.264 Scalable Streaming, 310, 311
Haar-like feature, 825, 826*f*
Haar-Waveletes, 823
Habitat monitoring, 76
Hadoop Distributed File System (HDFS), 606
Hadoop file system (HDFS: Hadoop Distributed File System), 556
Hadoop platform, 560
Hamming + 1 scheme, 700
and Hamming + 3 scheme, 705*t*
and matrix encoding, 705*t*
Hamming + 3 scheme, 700, 701
and Hamming + 1 scheme, 705*t*
and matrix encoding, 705*t*
Hamming codes, 698–699
binary Hamming codes, 699
1-error correcting codes, 701
Hand gestures, 616
Hard c-means (HCM) clustering, 104, 105, 106, 108
Hard disk device (HDD) capacity, 131
Hard scatter partition, 112, 117
of LBG algorithm, 116*f*
Hard threshold function, 147
Hardware-software co-design scheme, 120, 125
Harris-Benedict equation, 264
HBAE (Host-Based Analysis Engine), 557
Henagon shape, 888*t*
Heptagon shape, 888*t*
Hexagon shape, 888*t*
Hidden Markov model (HMM), 598
ED decision and, 509*f*
tracking, 508, 513
tracking errors by, 510*f*
Hidden pilot sequence, 30
Hierarchical task network (HTN), 663
High efficiency video coding (HEVC), 271
High performance cluster computing (HPCC), 128
HIGh security and light weigHT (HIGHT), 227
High utility sequential (HUS)
HUS-Span, 55
HUS-Span algorithm, 50
pattern mining, 49
High-performance TI DSP, 136
High-resolution digital images, 678
HMIs (Human Machine Interfaces), 526, 536
Homomorphic encryption scheme, 366
algorithm, 368–369
performance, 370*f*
preliminary definitions, 367
Host Discovery, 528, 529*t*
Host-based detection, 577
Hounsfield Unit (HU) histogram distribution, 682
Huffman coding, 721
Huffman decoding, 720
Huffman-encoded bytecode, JIT-decompression of, 720
HUINS SoC Master 3 board, 138, 139*f*
Human visual system (HSV), 39
spatial frequency response of, 40*f*
Human-computer interaction (HCI), 801
Hybrid deduplication scheme, 230
Hybrid electric vehicles (HEVs), 708
Hybrid method, pixel value's transition
adoption of curve fitting, 300–301, 301*f*
getting transition state, 299–300, 300*f*
Hypo-monocyclic-Coxian representation, 100

I
I/F filter, 540
I^2Core System Management Application, 70
IaaS, 860
IBM data generator, 55, 56*f*
IBM QRader, 555
IBS/INS, 545
ICCP, 536
I-Concatenation operation, 51, 53
Iconic stereotypes, 462–463
stereotype set class and, 465*f*
IDC (Internet Data Center), 516
Idle, stop, and start (ISS) function, 708
IEA, 787
IEEE 802.11 MAC, 170
backoff-free fragment retransmission fragmentation, 172–173
classical fragmentation (CF) scheme, 172
classical fragmentation contention overhead, 172*f*
classical fragmentation overheads, 172*f*
fragmentation in, 171
performance evaluation, 173–175
IEEE 802.11 station (STA), 170

IEEE 802.11 wireless local area networks (WLANs), 169
IEEE802.15.4, 495
IEEE802.15.4 task group (TG4e), 496
IEEE802.15.4e DSME, 496–497
  performance evaluations, 497–498
  topology, 499*f*
If-then rules, 104
Illumination effects, in stereo images, 302*f*, 303, 303*f*
IMA (Integrated Modular Avionics) architecture, 415
Image content, 799
  information detection, 800
Image data-per-processing element (IDPE) variation, 194, 200
Image object representation method, 800–801
Image processing
  hardware tools, 878, 879*t*
  image capturing, 878
  image detection, 880
  integrated car park management system, 876–877
  pre-processing, 878–880, 879*f*
  software tools, 878, 879*t*
  twit status, 880
  twitpic image, 880
Image segmentation, 678, 680, 684, 685, 877
  using Level-set, 679–680, 681*f*
Image-based automobile detection, 822
Implicit Password Authentication System (IPAS)
  advantages, 13–14
  domains, 14
  experiment, 15–16
  sample image with implicit objects, 16*f*
  security of, 19
  survey analysis, 17–18
  survey questions, 16–17
  usability of, 18–19
  user test phase, 15
Inactive pixel (IP), 424
Incoherent domain, 98
Independent component analysis (ICA), 598
Induction motor, 177
  fault classification of, 178–179
IndusCAP-Gate (Industrial Cyber Attack Prevention Gate) system, 536
  application, 539*f*
  architecture, 539*f*
  prototype platform, 543*f*
Industrial firewall, design of
  architecture and application of IndusCAP-Gate system, 538, 540
  implementation and evaluation, 543
  multiple access control filter based blocking of unauthorized access, 540–543
Information Integration Core (I^2Core), 67, 68
  basic data flow, 71*f*
  benefits of, 73
  early warning systems, 70–72, 72*f*
  earthquake sensor, 73
  execution concept, 70
  features, 68, 69*f*
  human interaction, 71
  sample execution, 70–73
Information integration, 324
Information systems
  consumer systems, 68*t*
  data sources, 67*t*
Information systems integration, 66. *See also* Information Integration Core (I^2Core)
  standalone system, problems of, 66–67
Information technology (IT), 279
Initial contour, 684
Initiative service, 324
Integrated car park management system, 875
  denoise, 880*f*
  hardware tools, 878, 879*t*
  image processing techniques, 876–877
  implementation, 878–881
  limitations, 881
  low financial entry point, 876
  pre-prepared mask, 880*f*
  social media for communication channel, 877
  software tools, 878, 879*t*
Integrated log analysis, 847–848
  design of, 845–846
  system, 845, 845*f*
Integrated starter-alternator (ISA), 708
Intel Xeon cache, 130*f*
Intel Xeon cores, 130*f*
Intelligent driver assistance system. *See* Advanced Driver Assistance System (ADAS)
Intelligent logistics, 675
Intelligent robots, 687
Intelligent transportation system (ITS), 764
INTER mode, 272
Interconnect technology, 131
Interface and Test Processor (ITP), 88
Interior permanent magnet synchronous motor (IPMSM), 709
  controller, 710*f*
International Association of Lighthouse Authorities (IALA), 545

Internet data centers (IDCs), 222
Internet navigation, 467
Internet protocol routing, 154
Interoperability standards, 66
InterPro (IPR), 786
  correlation coefficient among, 793*t*
  original, 793*t*
  weighted IPR, 792, 793*t*, 794–795
Interpro2go, 787
InterProScan, 786
Inter-streaming multiprocessors (Inter-SMs), 194
Inter-system VTS Exchange Format (IVEF) protocol, 545
INTRA mode, 272
Intrinsic (internal) aging factors, 262
Intrusion Detection System (IDS), 222, 223, 530
Intrusion Prevention System (IPS), 222, 223
Inverse Compositional Image Alignment (ICIA), 427–428
Inverse distance weighting (IDW) interpolation, 454, 457
Inverse DWT (IDWT), 22, 25
IP/MPLS networks, 565
I-phone, 254
ISO 14649 (STEP-NC), 241, 243
ISO26262, 716
ISS, 787
IT firewall technology, 537
IT logistics convergence. *See* Smart logistics service
IVEF (Inter-VTS Exchange Format) Service, 546
iWatch, 212
IxBD (Interface-BigData) interface, 558
IxLBVA (Interface-LB to VA) interface, 558

J

Japan Society of Civil Engineering (JSCE), 816, 819
  PBD, 816, 819
  structural performance verification of, 816
Java VM (JVM), 720
Joint Collaborative Team on Video Coding (JCT-VC) group, 271
Joint substream video model (JSVM) codecs, 309
Junk DNA, 22
Just in time (JIT)
  decoder, 720, 721
  decompression, 720
  Huffman decoding, 720, 722, 728
  operation model, 128

K

k-anonymity, 366, 367
Key escrow, 588
Key management methodology
  comparison of, 589
  encapsulation, 588–589
  key escrow, 588
  trusted third party (TTP), 589
Keyboard dynamics, 886
Keystroke dynamics, 886
Key-Value Models, 835
KINECT camera, 254, 255, 257, 258
Kinect, 616, 732
Knowledge encoding method
  knowledge authoring method, 411–412
  in u-health portal, 410, 411–412
Knowledge-based identification, 314
Korea Internet & Security Agency (KISA), 606
Korea national DNS, 606
  deployment status of, 608*f*
  structure of, 607
Korean diphthongs, 637*t*
Korean single vowel
  lip shape and tongue's location of, 638*f*
  single vowel triangle, 637*f*
Korean vocal sound system, 634
  characteristics of, 637
  vowel vocal sound system, 637

L

Lagrange multipliers, 180
Laplace transform, 99
Large-scale data analysis engine, 557–558
  feature description, 560*f*
  functions of, 559*t*
  major host-based features, 559*t*
  structure of, 558*f*
Layered streaming scheme, 306
LBAE (Large-scale Batch Analysis Engine), 557
Leap, 733
Level-set, 678
  segmentation for body detection, 681*f*
  speed function, 679
Lifting based integer DWT transform, 23

Likiwid Performance tools, 726
Limit state design, 814, 819
Linde-Buzo-Gray (LBG) algorithm, 112
  hard partitioned spaces, 116*f*
  input-output spaces, 116*f*
  model outputs, 116*f*
  nonlinear process, 113, 114, 117
  rule generation, 112, 113, 114
Linear inference, 113
Linear Predictive Coding (LPC) algorithm, 634
Linear regression module, 848
Linear temporal logic (LTL), 431, 432
  decentralized setting, 433*f*
  decentralized specifications, 432–433
Linguistic modeling, 111
Linux background, 444
Linux kernel, 579
Lip sync solution, 635
  block diagram of, 636*f*
  in digital contents, 640
Local binary pattern histogram (LBPH), 385, 388
Local maximum processing, 753, 754*f*
Local memory, 195
Local overlapping bucket (LOB) scheme, 352
Localized Big Data Analysis (LBDA), 606, 608
  advantage of, 609
  for big data, 613
  C&C domain, 611
  Hadoop installation, 609
  Map/Reduce of, 611
Location area management, 57
Location area planning, 58
  solution representation, 59
Location Areas (LA), 58
Location authentication, 440–441
Location mechanism, 442
  basic scheme, 442
  registration phase, 442–443
  secure scheme, 442
  supported technology, 443–444
Locations
  area-based coupons, 440–441
  reliable detection of, 441
  with no signals, 441
Locking system for security improvement (LSSI), 468
  D-manager, 470
  DB-manager, 470
  design system of, 468, 470, 469*f*
  DSD, 468
  IDP-manager, 470
  implementation of, 470–471
  initial execution screen, 470*f*
  input time, 471*f*
  L-service, 470
  performance evaluation, 471
  screen upon receiving 10 directions, 471*f*
Locking system. *See* Locking system for security improvement (LSSI)
LOD (Linked Open Data), 448
  life cycle and tools, 450*f*
  LOD2, 448
Log authentication system
  authentication processing, 594, 595*f*
  design of, 592–593
  OSP system environment, 591
  role of, 593
  signature authentication, 595*f*
  system configuration, 593–594
  X.509 institutional authentication format, 592*f*
  X.509 user authentication format, 592*f*
Log-chromacity color, normalization of, 298
Logic Based Models, 835
Look-up tables (LUTs), 142
Low density radio map, interpolation of, 510–511
Low rate wireless personal communications (LR-WPAN), 495
Low-Cost Airline Service System, 324
  booking services, 326
  computing engine, 325–326
  real-time information transformation, 325
  task manager, 325
  web services, 325
Low-cost airlines (LCA), 322
Low-cost carrier (LCC), 322
LQS-Tree, 51
  and Utility Matrix, 52*f*
LRIT, 545
LSAE (Legacy Security event Analysis Engine), 557
Lung segmentation
  distribution of variance and area, 683*f*
  experimental results, 685*t*
  image segmentation using Level set, 679
  image segmentation using MRA Level set, 681*f*
  MRA Level-set method using auto-configuration of the initial contour, 680–684, 685*f*

M
MAC layer collisions, 4
MAC protocol data units (MPDU), 171
MAC service data unit (MSDU), 171
Machine learning algorithms, 576, 578, 584
Machining, 238
Macro modeling, 293–295
  for TVC circuit, 294*f*
  for VTC circuit, 294*f*
Magnetic devices, 688
Magnetic Resonance Imaging (MRI), 678
Maintenance and Test Processor (MTP), 88
Malware detection
  collection of resource information, 578
  features for, 579*t*
  resource feature for, 579
  resource information monitoring agent, 580–581, 581*f*
Mamdani method, 46
Manchester OWL syntax, 242
Mandatory Access Control (MAC), 517, 518
Mandatory Key Fragment, 378
MANET (Mobile Ad-Hoc Network) environment, 773
Many-core architecture, 194, 197, 198
MapReduce, 557
  abnormal behavior detection method using, 560
  handling process, 561*f*
  SINBAPT, platform specification, 561*f*
Markers, 254
  game "Beyond: two souls", 255*f*
Markup Scheme Models, 835
Mass air flow (MAF), 249
MATLAB, 764
  -based third-party program, 708
Matrix encoding, 698
  and Hamming + 1 scheme, 705*t*
  and Hamming + 3 scheme, 705*t*
Matrix metalloproteinases (MMPs), 262
Matrix module, 848
Maximum accessed domain SQL query, 606
MB-OFDM UWB systems, 30
  network throughput performance, 33*f*, 34*f*
  simulation parameters, 33*t*
  system model, 30–31
MCR (Main Control Room), 89
MCSO-SVM, 333
  classifiers, 333
  and CSO-SVM, 335*t*
  evaluating fitness, 333–334
  mutation operation, 333
  solution set design, 333
Mean-variance based method, 680
Medical imaging equipment, 678
Medium access control (MAC), 170
Mel-frequency cepstral coefficients (MFCC), 393, 634, 635
Memory, 131
Merkle tree, 230, 234
Message authentication code (MAC), 77
Message-Locked encryption, 230
MiBench embedded benchmark suite, 720
Microsoft Foundation Classes (MFC), 662
Mile per Gallon (MPG), 248–249
Min-Cut Max-Flow Theory, 745
MITM, 222
Mixture of Motor Primitives (MoMP), 662
Mobile administrative services, 474–475
  blacklist function, 475
  whitelist function, 475
Mobile application interface, 324*f*
Mobile based pre-order system, 654
Mobile cloud computing (MCC), 212
Mobile device (MD), 453, 507
  ED decision and HMM tracking, 509*f*
  moving device, tracking of, 508
  moving path of, 509*f*
  simulation results, 508–509
  tracking errors by EUD, 510*f*
  tracking errors by HMM, 510*f*
  tracking on real RSS measurements, 509
Mobile device management (MDM), 475, 480–481
  App management functions of, 481*t*
  components of, 477*t*
  configuration diagram of, 477*f*
  functional evaluation, 477–478
Mobile internet, 322
Mobile malicious code, 475
  types of, 476*t*
Mobile malware detection trend, 576–577
Mobile system
  database, 656
  database design, 656–658
  server, 656
  socket programming, 656
Mobile telephony, 96
Mobility E-EMS, 224–225
  implemented control system, 226*f*
Modbus, 526, 536 537
  command whitelist, 531, 533*f*
  commands verification, 533
  DPI Firewall, 537
  protocol, 526, 528
  TCP/IP ADU, 527*f*
Model report module, 848
Model-based design (MBD), 708

Modified cat swarm optimization method (MCSO), 333
Modified quadratic inference, 113
Moment fuel efficiency value, 248
Mono-cyclic coxian representation, 100
Monoscopic motion mask (MMM), 404
Moore's law, 130, 131
Motion activities
  motion vector (MV), 273
  prediction modes, 273*t*
Motion capture, 254
Motion detection, 162
Motion recognition, 254, 258
Motion sensor. *See* Attitude Heading Reference System (AHRS)
Motion vector, 38
Motion vectors merging (MVM) method, 272
Motor-starter-generator (MSG) technology, 708
Movement vector, 39
Movement-containing region detection (MCRD), 162–163
  MCRD, example, 163*f*
MP4 Scalable Streaming, 310, 311
MPEG encoder scheme, 39
MPEG-7, 800
M-sequence, 38–39
Multiband orthogonal frequency division multiplexing (MB-OFDM), 29
Multi-channel EMG, 616
Multi-core architecture, 136, 138, 142
  discrete wavelet transform on, 141*f*
  moving regions detection, 140*f*
  reference multi-core architecture, 137, 138
  synthesis results, 143*t*
  system parameter, 139*t*
  and TI C6146 processor, 143*t*
Multimedia content management, 393, 399. *See also* Audio classification; Audio features extraction
Multimodal face recognition. *See* Face representation
Multi-modal user interfaces, 254–255
  "Put-That-There" controlled systems, 254
Multiple automata, 432
Multiple Criteria Decision Making technique, 242
Multiple face tracking. *See* Face labeling algorithm (FLA)
Multiple-access control filter, 542*f*
  processing and policy of, 541*f*
Multiply-accumulator unit (MACC), 196
Multi-Resolution Analysis (MRA) Level-set, 678
  lung segmentation method using auto-configuration of the social contour, 680–684
Mutated codon sequences, BERs of extracted watermarks in, 26*f*
MySQL, 557

N
Nao robot, 662
Natural aging process, 262
  support vector machine (SVM), 262
Natural user eXperience (NUX), 598
Natural user interface (NUI), 598, 731
Navigation systems, 740
NBAE (Network-Based Analysis Engine), 557
Nehalem architecture, 722
Network allocation vectors (NAV), 171
Network attached storage (NAS), 128
Network congestion, 3
Network games, 468
Network interface card (NIC), 540
Network layer access control block, 540
Network lifetime, 782
Network throughput
  high transmission power, 362, 363*f*
  low transmission power, 362, 363*f*
  in interference-free environment, 359–360
  in signal interference, 360–361
Neural networks, 119
New Reno protocol, 2
Next Generation Sequencing (NGS), 786
Next zone selection, 756*f*
  destination zone, 758*f*
  transmission range, 756*f*, 758*f*
Nicotine consumption and skin aging, 262
Node energy, 775
Non-coding DNA sequence (nCDS), 22
Non-fire-containing movies, normalized wavelet energies with, 165*f*
Non-fuzzy inference systems
  consequence identification, 113–114
  performance evaluation, 114, 115*t*
  premise identification, 112–113
Non-fuzzy neural networks (NFNNs), 104
  learning algorithm, 105
  local spaces of, 106, 107*f*
  optimization, 106–108, 108*f*
  performance index, 106–107, 107*t*
  structure of, 104–105, 105*f*
Non-overlapping 7-pixel blocks, 702
Nonvolatile memory express (NVMe), 131
Normal (NO) fault, 179
NoSQL, 557

NP-hard problem, 58
NUI/NUX framework, 255
  hand gesture, 258*f*
  hand mouse framework, 256*f*
  hand mouse gesture recognition, 255, 257
  mapping to monitor, 258*f*
  mouse area comparison, 257*f*
  system architecture, 256*f*
NUI/NUX-based gaming, 598

O
OBD car doctor pro, 248
OBDdroid scan pro, 248
Object centered gaze behavior analysis system
  content object representation structure, 802–803
  gaze data analysis model in real time, 803–804
  system model structure, 801–802, 802*f*
Object detection, 688, 692, 695
  image processing for, 693*f*, 694*f*
Object Oriented Models, 835
OKF (Open Knowledge Foundation), 448
On-board diagnostic II (OBD II), 248
  parameters, 249*t*
On-demand instances, 483–484
"One computer has one function" concept, 415
One-dimensional (1D) vibration signals, 178
One-time password (OTP), 76
One-way hash function, 76, 77, 78, 83
Online server development for robot control
  interoperability of server, 664–666
  server protocol for task execution, 666–667
  task generation, 663–664
Online service providers (OSPs), 587
  architecture system of, 593*f*
OntoBLAST, 787
Ontologies, 239, 241
  model for automatic growth environment control of crops, 838–839
  service based on, 835
Open Source Computer Vision Library (OpenCV), 662
Open Source software, 878
OpenCV image processing, 876, 879*t*
OpFake, 584
Optical combiners, 97
Optical computing, 95
Optical digital filter, 96, 101
Optical fiber, 96, 98, 101
Optical networks, 96, 97
Optical processor, 97*f*
Optical splitters, 97
Optimization problems, 330
Optional fabric for backup, 128
Optional Key Fragments, 378
Original continuous query (OCQ), 203
  transformation, 205*f*
Outlier analysis module, 848
Out-of-bid event, 484
Over the Air (OTA) technology, 475
OWL technology, 323
OWL-DL ontology, 239, 240

P
PaaS, 860
Packet collection and control block, 538
Packet losses, 2
Packet Sniffing, 222
Pairing Based Cryptography (PBC) library, 190
PAMS (Post-Accident Monitoring System), 89
Pan coordinator (PC), 495
Parabolic-rectangular (*p-r*) stress-strain curve, 810, 812*f*, 817
Parallel computing, 154
Parallel misalignment fault (PMF), 179
Parameter estimating module, 848
Parameter select module, 848
Parent process ID (PPID), 647
Particle swarm optimization (PSO), 104, 106, 329
Passdoodle system, 13
Passfaces system, 13
Passwords (locking system), 468
Path modification, gesture definition for, 733–734
Path planning, 732
  with human intervention, 732
Patterns (locking system), 468
Pay-per-use pricing, 854, 855, 857–858
PDA-based logistics/delivery computer system, 671
PDU-DB Server, 225
Peak signal to noise ratio (PSNR), 41, 42, 276*t*, 277, 405, 703, 705
  numerical values comparison, 303*f*, 304
Pellet reasoned, 243
Pentagon shape, 888*t*
Pentagonal bipyramid, 889*t*
PER (Packet Error Rate), 777
Personal cloud, 374
Pfam, 787
Photo-detector, 97
Photonic computing, 95

Photonic devices, 96
Photonic filters, 97
Pig code, 610, 611*f*
PINs (locking system), 468
Pipelined microprocessors, 721
PIR SuperFamily, 787
Pixel-to-vertex map (PVM), 424, 428
PKI-based authentication sub-system, 588
PLC (Programmable Logic Controller), 89, 526, 537
Plot module, 848
Poisson distribution, 218
Poisson process, 31
Policy and alert management block, 540
Polygon-based object annotation method, 800
Position-control performance model, 622*f*, 623–624
Positive realization, 96, 98–100
Positive system, 98, 99, 101
Positron Emission Tomography (PET), 678
Potential high utility sequence (PHUS), 50
Prediction unit (PU), 272
Prediction-based travel time, 748*t*
Pre-order service, 655
PreScan, 764
Price history, 484
Pricing schemes, 854, 855, 859, 860
Principal component analysis (PCA), 598
Prints, 787
Priority matrix (P_S), 566–567, 567*t*
  bridge Id/path Id, 568*t*
  cyclic priority matrices, 569–571
  fairness evaluation, 571–572
  with 4-bit priority words, 568*t*
  problem definition, 569
Privacy-preserving data collection
  algorithm, 368–369
  performance, 370*f*
  preliminary definitions, 367
Private LAN, 128
Probability Based Improved Broadcasting algorithm, 775
Probability density function (PDF), 32
Probe packets, 776
Process ID (PID), 647
Processing elements (PEs), 136
  RTL schematic, 4 × 4 PE array, 142*f*
  RTL schematic, single PE, 142*f*
PRODIGY, 409
Prodom, 787
Projected-Database Utility (PDU), 50, 54–55
PROMELA model, 416
Prosite, 787
Protein function, prediction method of, 789–790
  data reconfiguration using under-sampling, 791*f*
  dataset reconfiguration, 790
  feature selection, 790–792
  weighted IPR, 792, 794–795
Protein functional site, 786
Prototyping system, multi-core
  HUINS SoC master 3 board, 140*f*
  multi-core embedded system, 137, 138*f*
  reference multi-core architecture, 137
  reference multi-core array, 138*f*
  synthesis results of, 143*t*
  system parameters, 138–139, 139*t*
Provable Data Possession (PDP), 186
Provider Link State Bridging, 566
PSO-SVM, 330
Public key cryptography, 77
  cluster formation, 80–81
Public key infrastructure (PKI), 588
Public LAN, 128
Pure recall-based technique, 13
Python script, 727

Q

Q-learning algorithm, 598, 604
Qplus-653 kernel, 416
  accessing mechanism, 421*f*
  inter-partition communication of, 417–419
  message transmission in, 419*f*
  non-error-propagation in, 419–420
  qplus_ports, 420
  qplus_queue, 420
q-subsequence, 50
Quadratic inference, 113
Quadrilateral shape, 888*t*
Quantitative sequence database (q-SDB), 50, 53*t*
Quasi-identifier, 368

R

R analyzing tool, 607
Radar, 822
Radial basis function (RBF), 166, 181, 331
Radio Map, 454
  interpolation, 506
Raspberry pi, 876
Rational Rose software, 464

Reactive routing protocol, 774
Reactor protection system (RPS)
architecture, 88
of power plants, 90–92
of research reactor, 90
Real radio signal propagation, 507
Real time lip synchronization, 635
Real time memory analysis
delivering evidences and confirmation, 646
digital evidence securement and collection, 646
investigation and analysis, 647
investigation preparation, 645–646
memory dump process, 648
memory imagification for process, 648
process list extraction, 647
real time data analysis, 648–650
real time data investigation procedure, 647
response, 646
SNS investigation procedure by, 644–645
web browser memory use list, 648, 649*f*
Real time resource reallocation for better resource utilization (RTRRBRU), 44–45
Real time vowels recognition, 635
Real-time data analysis, 557
Real-time operating systems (RTOS), 416
ARINC 653 (*see* ARINC 653 Standard)
Qplus-653 kernel (*see* Qplus-653 kernel)
Recall-based graphical techniques, 12
Received signal strength (RSS), 453, 507
map resolution with, 456–457
Recognition- based graphical techniques, 12
Reduced Sequence Utility (RSU), 50, 53–54
Reference many-core processor, 195, 196*f*
Register-transfer level (RTL), 198
Regulatory Guide 5.71, 1.152, 88
Relay node selection, 756*f*
Reno protocol, 2
Re-routing algorithm, 746
Residual analysis module, 848
Residual energy, 774, 776
variance of, 782–783, 783*t*
Resource reallocation algorithm, 45–46
Respondent-defined privacy protection (RDPP), 367
RF (RandomForest) classifier, 581
machine learning algorithm, 578
performance comparison, 582*f*
performance indicator, 583*t*
RFIDs, 203
application domain, 461
tag and sensor tag application, 204*f*
RGB color mean figure, 343*f*
Roads in target area, 748*f*
Robots
adhesion system and impeller, 689*f*
cross development model, 692*f*
design and structure, 688–691
effective battery duration, 693*t*
embedded image processing system, 691–692
existing forces analysis, 691*f*
image processing and communication, 693*f*
image processing procedure, 693*f*
improvements in, 691*f*
overall exterior view of, 690*f*
payload capacity, 693*t*
position with payload, 694*f*
speed in various direction, 693*t*
Role Based Access Control (RBAC), 517, 519
Rotor unbalance fault (RUF), 179
Round-Robin scheduling model (RRSM), 46
Route error (RERR) message, 4
Route reply (RREP) message, 4
Routing algorithm, 745. *See also* Bellman-Ford distance vector routing algorithm
calculating estimated travel time, 746
searching optimal path, 745
traffic information modelling, 745
Routing success probability, 783, 783*t*
RREQ message, 775
RSA cryptosystem, 38
RSSI (Received Signal Strength Indication), 359

S

S/Key algorithm, 76
generation of set of passwords, 79*f*
password generation phase, 78
registration phase, 78
S5PV210 microprocessor, 688
Saccades, 803
Safe and Saving index (SS-index), 248, 250, 251*f*
SAGE (Standard-based Sharable Active Guideline Environment), 409
SAX parser, 549
SCADA (Supervisory control and data acquisition) systems, 525
architecture, 526
general layout, 527*f*
Modbus protocol, 526, 528
servers, 536
SCADA Firewall, 537
Scalable streaming service, 306
Scalance, 537

Scan attack, 528, 529*t*
Scatter partition, 104, 108
Scissors gesture, 733, 734*f*
S-Concatenation operation, 53
Script-based learning, 599
SDbitmap, 497
Secure and adaptive transmission control protocol (SA-TCP), 2, 3–5
  security, 5–6
  simulation, 6–7
Security
  accident time period, 844, 851
  equipment, 845
  related infra logs, 844, 851
  related issues, 843
Security authentication methods, 314
Security control system, 222, 223, 227
Security function, 468, 469*t*
  DSD, 468
Security Information and Event Management (SIEM), 223–224
Security intelligence technology, 554–555
Security parameters, 3
Security Service Provider (SSP), 378, 379, 380, 381, 382
SEED-256, 255
  data encryption conversion code, 226*f*
Self-Adaptive Navigation System (SeNS), 740
  comparison of existing systems, 743*t*
  design, 742, 744–745
  scenario, 747
  simulator, 746–747
  system requirements, 742
Self-adaptive software, 749
Self-awareness privacy, 367
  protection (*see* Privacy-preserving data collection)
Self-enforcing privacy (SEP), 367
Semi-digital smart integrated circuit design, 289
  cyclic type analog-to-digital converter, 289
  successive approximation (SAR) converter, 290
  TVC (*see* Time-to-voltage converter (TVC))
  VTC (*see* Voltage-to-time converter (VTC))
Sensor nodes, 76, 82
  authentication by user, 81*f*, 82–83
Sensors, 203, 834
Sequence Weighted Utilization (SWU), 49
Service models, 854, 860
  scenario, 674–675
Service process management module, 869
Serviceability limit states, 810
Serviceability performance, 810
Session hijacking, 2
Set Top Box (STB), 307
Shadow area detection, 829*f*
Shannon theory, 30
Shortest path bridging (SPB), 566
SIEM (Security Information and Event Management), 555
SIFT descriptors, 506
SIFT flow, 424, 425
SIFT matching test, 507*f*
SIFT-based building matching, 512
Signal coverage, 358, 362
Signal interference, 358
  to network throughput, 360–361
  simulation topology, 361*t*
  in WiFi networks, 359
Signal to interference plus noise ratio (SINR), 30
Signature signal, 38–39
  generation, block diagram, 39*f*
Signature-based detection, 576
Simple Additive Weighting technique, 242
Simple vector filter, 120
Simplified inference, 113
Simulated annealing, 58
  with compact index, 60*f*
Simulated log events, 849*f*
Simulating synthesized automata, 434
  arbiter, 434, 435, 436, 437
  robots, 434, 435, 436, 437
  screenshot of, 435*f*
  setting of, 435*f*
  specifications of, 436*f*
Simulation
  design condition for, 714*t*
  driving environment for, 714*t*
  road condition for, 714*f*, 715*f*
Simulink, 764
SINBAPT (Security Intelligence techNology for Blocking APT), 5557, 58, 562
  cyber-targeted attack response system, 557*f*
  platform specification, 561*f*
Single instruction multiple data (SIMD), 194
Single vowel recognition, 638, 640*t*
Situation information model for smart logistics service, 672–674
  coordinate information, 673
Situation information-based intelligent logistics service model, 671–672
SK Communications hacking, 554
Skin aging
  and alcohol consumption, 265

and sleeping, 265
Skin hereditary traits
calories and exercise intensity, 266*t*
data for normalization, 264–265
evaluating, 266–267
extrinsic aging factors, 263–264
skin texture, 263
statistical estimation of, 267
Skin texture ages, 262
Skin type and skin aging, 262
SKIP mode, 272
Sliding gesture, 733, 734*f*
Slow start threshold (*ssthreshold*) value, 4
SMART, 787
Smart devices, 314, 467, 468, 470
authentication processes, 318
face recognition, 315
false acceptance rate (FAR), 318
false rejection rate (FRR), 318
graphical password system, 316
key generation mechanism, 316, 317*f*
key space, 316–318
keystroke dynamics-based user authentication (KDA), 316
locking systems, 469*f*
password, 315
pattern lock, 315
PIN, 315
security threats in user authentication in, 314–315
Smart integrated circuits. *See* Semi-digital smart integrated circuit design
Smart logistics service, 670
situation information model for, 672–673, 673*f*
structure chart, 672*f*
Smart phone, 467, 575
SmartRack device, 222
SMO, 769, 790*t*, 794*t*
SMTP (email), 66
Social network service (SNS), 575, 643, 644, 863, 864
digital forensic of, 644
identification of, 645
information collection, 646
information related to conversation of, 649
process of acquiring, 645
searching digital evidence, 651*t*
use in web browser, 650*f*
user networking information, 649
user, timeline information of, 650
Social Networking Service (SNS), 468
Soft threshold function, 147
Solar power, 883
SOX act, 128
Speaker dependent recognition, 640
Spectrum sensing, fuzzy for
membership functions of fuzzy sets, 148*f*
proposed scheme, 147–148, 148*f*
ROC curves of, 150*f*, 151*f*
simulation results, 150–151
wavelet denoising, 147
Speech recognition, 634, 635, 638, 640
Speed-limit signs
adjacent pixel, labeling, 123*f*
boundary information extraction, 122–123
boundary smoothing, 121–122
detection and recognition of, 120–121
noise elimination, 121–122
performance evaluation, 124–125
red-colored ring, boundary recognition, 122, 123*f*
speed recognition, 124
system design, 124, 125*f*
SPIN model checker, 416
Splunk application, 555
Spot instances, 483, 484
Spot prices, 484
Square based pyramid, 889*t*
SRDP, 3
Standard for Exchange of Product data (STEP), 238
Stateful packet inspection (SPI), 537
Static analysis, 576
Statistical prediction engine, 845, 846
Statistic-based prediction engine system, 846*f*
generation, 847, 847*f*
integrated log analysis, 847–848
prediction with regression analysis, 850*f*
prediction with time-series analysis, 850*f*
Statistics module, 848
Statistics-based log analysis, 844
Steel stress calculation, algorithm for
computation of steel stress, 811–814
crack width formulation, 814–817
nonlinear material characteristics of concrete, 810–811
Steganography, 22, 698, 705, 706
Stereo camera, 298
Stereo image recording (SIR), 401, 402
Stereo images, 302*f*, 303*f*
PSNR numerical values' comparison, 303*f*
Stereo matching, 297
Stereoscopic watermarking, 402
Stereotype creator, 463–464, 464*f*
configuration file, 464, 465*f*
update registry information, 464
ST-FeatCAPP, 238

Storage area network (SAN), 128, 131
Storage bounded continuous queries (SBCQ), 204–205, 206
  update algorithm, 208*f*
Strain and bond stress distribution of tension members, 815*f*
Streaming video, intermediate node for, 307
Strength limit states, 810
Stresses in steel, 810
String topology, 7
Structural limit states, 809–810
Stuxnet, 536
Subscription pricing, 854, 855, 856–857
Sub-task table, 599, 600*t*, 603*t*
  sub-tasks by virtual agent, 602*f*
Subtractive clustering method, 194–195
  evaluation metrics, 198, 199*t*
  execution time of, 199–200
  MR images, 199*f*
  parallel implementation of, 196–198
  parameters for, 196*t*
  PE configurations, 199*t*
  PE configurations for MR images, 200*f*
  process computing first potential value, 197*f*
  system utilization, 198, 200*f*
Suction-based adhesion system, 689
Sudden acceleration (SA), 250
Sugeno-Takagi models, 112
Superframe duration (SD) index table, 496
  selection criteria, 498
Superframe, 496
Support vector machines (SVMs), 165, 262, 330–331, 790*t*, 823
  data classification using, 578*f*
  malware detection via, 577–578
  in measuring skin aging, 262
  optimal hyperplane, 331*f*
  RF (RandomForest) machine learning algorithm, 578
Support vector regression (SVR), 180–181
SWISS-PROT, 787, 788, 796
SYN flooding, 2
Synopsys design compiler, 198
System info module, 848
System on Chip (SoC) integration, 129
System whitelist, 531

T
T Map, 741
Tablet PC, 254, 314, 315
Tahoe protocol, 2
Taint analysis, 576, 577. *See also* Dynamic analysis-based detection
Task scheduling, 212, 213, 216, 219
Task success probability table, 600*t*, 601
  through Q-learning, 602
Task-Role Based Access Control (T-RBAC), 517, 520
TCP New Reno, 3, 4
TCP/IP session analysis, 537
Test simulation module, 848
Texture patterns, 179
Thayer's emotion model, 339–340, 340*f*, 341
Thick clients, 211
Thin clients, 211
Thin-thick clients, 212, 213, 217, 219
Third-party auditor (TPA), 186
Threshold cryptography technique, 3
Tick Tock, 129
TIGRFams, 787
Time division multiple access (TDMA) scheduling, 33
Time-frequency code (TFC), 30
Time-to-voltage converter (TVC), 290
  circuit implementation of, 291*f*
  CMOS switch, 294
  error analysis of, 290–293
  error performance of, 292–293
  macro models for, 294*f*
TinyVM, 720, 721
Tolerance level, 890
TOMTOM, 741
Torque assist systems (TASs), 708–709
  complete ASM block layout for, 709*f*
  configuration of an HEV for, 709*f*
  efficiency map, 713*f*
  electrical parts, 709–710
  engine, 710–711, 711*f*
  integrated controller, 711–712, 712*f*
Torque Pro, 248
Touch screen, 468, 470, 472, 616
Tracking faces, 383
Traffic information
  generate and collect, 741
  infrastructure for generate and collect, 740–741
  interactions among components, 744*f*
  providing service, 741–742
Transfer control protocol/internet protocol (TCP/IP), 662
Transform unit (TU), 272
Transformed continuous query (TCQ), 204

Translation-rotation matrix, 766
Transmission control protocol (TCP), 2
Tree pruning algorithm, 272
Trial-and-error method, 810
Triangle shape, 888*t*
TrOASIS, 741, 742
Trusted Software Monitor (TSM) system
  authentication protocol phase, 283
  authorization protocol phase, 283–284
  authorization, 281
  efficiency, 285*t*
  home authentication server (TSMH), 286
  registration protocol phase, 283
  security, 285*t*
  security requirements, 280–281
  system symbol, 282
  ticket renewal protocol, 284
  ticket renewal protocol diagram, 282*f*
  ticketing method, 281–282
  TSMH authentication message, 286*f*
Trusted third party (TTP), 347, 589
Tweet message, 881*f*
Twit status, 880
Twitpic image, 880
Twitter micro-blogging, 876, 877
Two-dimensional PE array, 195
Two-part tariff, 854, 855, 858–859

U
Ubiquitous computing, 358
Ubiquitous wireless networks, 289, 363 *See also* Wireless fidelity (Wi-Fi)
u-health portal service, 410
  application case and evaluation, 412–413
  guideline for time-related knowledge, 410–411
  knowledge authoring method, 411–412
  model connected to CDS, 411
  system architecture for, 412
Ultimate limit states, 810
Ultrasound, 678
Ultra-wideband (UWB) devices, 29, 31
Unauthorized access, 538, 540
  multiple access control filter based blocking of, 540–542, 543, 544
Unified modeling language (UML), 461
Universal Description, Discovery and Integration (UDDI), 869
Updater module, 848
User authentication, 79–80
  analysis in smart devices, 315–316
  biometric-based identification, 314
  components comparison, 84*t*
  guessing attack, 83
  impersonation attack, 83
  replay attack, 83
  rhythm-based authentication scheme, 316–317
  in smart devices, 315
  stolen-verifier attack, 83
  transmission comparison, 84*t*
User datagram protocol (UDP) communication, 662
User's behavior analysis, 800
User-centric Security Service Model (USSM), 374, 375, 378
  data storage protocol, 378–379, 378*f*
  data usage protocol, 379–380, 380*f*
  requirements of, 375
USpan, 50
UV exposure and skin aging, 262, 266
UWB-WPAN, 31

V
Vacuum suction techniques, 688
VANET (Vehicular Ad-hoc Network), 751
  GZTR Protocol (*see* GZTR (Geographical Zone Based Two-tier Routing) algorithm)
  routing, problems of, 752*f*
Vapnik-Chervonekis dimensional space, 331
Variables transform module, 848
Vehicle, 247–248
  fuel used in, 251
  safety, 763
  speed (VSS), 249
Vehicle study research, 824*t*
  detection and tracking schemes, 824*t*
  detection and verification, 824
  false positive, 829, 830*f*
  test video environments, 830*f*
  test video results, 830*f*
  true positive, 829, 830*f*
Vehicle surroundings monitoring system
  analysis model for, 764–765
  block diagram of, 765*f*
  mechanism of, 765–768
  principles of, 766*f*
  safe distance secured by AEBS, 770*f*
  simulation scenario, 768, 769*f*
  verification of, 768–769
  warning zone and relative angle, 767*f*
  warning zone changes in simulation, 770*f*
Vehicle-to-infrastructure (V2I) communication, 754
Vehicle-to-vehicle (V2V) communication, 754

Vehicular tracking, 76
Verilog HDL, 141
Vertical farm, 835, 839
  context-aware control service, 835–837
  controlled environmental data, 837
  controlled target, 838*t*
  inference rule, 840
VESSEL segmentation in the Lung 2012 (VESSEL12), 684
Vessel traffic service (VTS), 545
  memory usage, 547–548, 549*f*
  memory usage, comparison, 548
  performance time, on Android, 548*f*
  performance time, on PC, 547, 548*f*
  pull-based parsers, 547
  push-based parsers, 547
  XML parser in, 546–547
VESSEL12 DB, 684
Video contents annotation, 800
Video processing technologies, 162
Video source, 799
Video watermarking. *See* DIBR video watermarking
Virtual agent learning processes, 599, 600*f*, 601
  learning objectives, 602
  3D models of virtual agent, 603*f*
Virtual desktop infrastructure (VDI) applications, 128
Virtual human learning processes, 599, 600*f*, 600–601
  3D models of virtual human, 603*f*
Virtual machines (VMs), 719
  codeword sequences, 724, 725*f*
  decoder mechanism, 723
  decoding time and preloaded branches, 725*f*
  indirect branches, 723
  instruction pointer (IP), 722
  and Latendresse's algorithm, 726
  and Mibench benchmark suite, 727*f*
  speedup of, 727*f*
Virtual Private Network (VPN), 222
Visual C#.Net, 817
Voice recognitions, 635–636
Voltage-to-time converter (VTC), 290
  circuit implementation of, 291*f*
  CMOS switch, 295
  error analysis of, 290–293
  error performance of, 293
  macro models for, 294*f*
Voronoi diagram, 31
Voronoi tessellation, interpolation, 454, 457
Vowel recognition processing strategy, 638–640

W
W3C standard, 240
Wall climbing intelligent mobile, 689*f*
War walking/driving/carting, 222
Watermark embedding on DWT domain, 24*f*
Wavelet decomposition (WD), 162, 167, 178
Wavelet denoising scheme, 147–148
Wavelet transform, 146, 147, 149
WAZE, 740
Wearable device, 616
Weather-proof housing, 883
Web Ontology Language (OWL), 239
Web Ontology Language service (OWL-S), 663
Web search, 575
Web Service Definition Language (WSDL), 869
Web Services, 66
Web-hard system, 593*f*
Weighted Information Fragmenting Scheme (WIFS), 374, 376
  algorithm proof, 377
  defragmentation process of, 377
  fragmentation process of, 376–377
Whitelist security technology, 476–477
  components of, 477*t*
  configuration diagram of, 477*f*
Whitelists, 531
Wireless ad hoc networks (WANETs), 2
Wireless communication technology, 834
Wireless connection, 306, 310, 311
Wireless devices, 440, 442
Wireless fidelity (Wi-Fi), 248
  -based indoor localization techniques, 453
  -based localization, 506
  connectivity, 883
  direct, 444
  networks, 358
  peer-to-peer (P2P), 444
  routers and access points, 444
  tethering, 444
Wireless Local Area Network (WLAN), 359
Wireless personal area networks (WPANs), 29
Wireless sensor networks (WSNs), 76
With Coffee application, 655

Workflow scheduling algorithm, 484, 485, 488*f*
considering task processing rate, 485
cost, comparison of, 493*f*
failure time, comparison of, 491*f*
initial workflow function, 488*f*
mapping relation of workflows and instances, 486*f*
recalculation operation of assigned task, 486*f*
recalculation point of task size, 486*f*
recalculation workflow function, 488*f*
resource types, information, 489*t*
rollback time, comparison of, 492*f*
simulation, parameters and values for, 489*f*
total execution time, comparison, 490*f*

X
XmlPullParser-based library, 546
XMPP (chat), 66
XOR operation, 39
X-Ray, 678

Y
YCbCr color space, 163, 164, 167

Z
z-axis motion mask (ZMM), 404
Zero-crossing rate, 393
Zone and conduit model, concept of, 536
Zone based routing, 752