

James J. Park
Hamid Arabnia
Hang-Bae Chang
Taeshik Shon *Editors*

Proceedings of the
International Conference on
Information Technology
Convergence and Services 2011
and Intelligent Robotics,
Automations, telecommunication
facilities, and applications 2011

ITCS & IRoA 2011

Lecture Notes in Electrical Engineering

Volume 108

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

James J. Park · Hamid Arabnia
Hang-Bae Chang · Taeshik Shon
Editors

IT Convergence and Services

ITCS 2011 & IRoA 2011

 Springer

Prof. James J. Park
SeoulTech Computer Science
and Engineering
Seoul University of Science
and Technology
Gongreung 2-dong 172
Seoul 139-743
Korea
e-mail: parkjonghyuk1@hotmail.com

Prof. Hamid Arabnia
Computer Science, GSRC 415
University of Georgia
Athens, GA 30602-7404
USA
e-mail: hra@cs.uga.edu

Prof. Hang-Bae Chang
Business Administration
Daejin University
Hogukro 1007
Pocheon-Si, Kyonggi-do 487-711
Korea
e-mail: hbchang@daejin.ac.kr

Prof. Taeshik Shon
Division of Information and
Computer Engineering
Ajou University, San 5
Suwon Gyeonggido 443-749
Korea
e-mail: Taeshik.shon@gmail.com

ISSN 1876-1100
ISBN 978-94-007-2597-3
DOI 10.1007/978-94-007-2598-0
Springer Dordrecht Heidelberg London New York

e-ISSN 1876-1119
e-ISBN 978-94-007-2598-0

Library of Congress Control Number: 2011940818

© Springer Science+Business Media B.V. 2012

No part of this work may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, microfilming, recording or otherwise, without written permission from the Publisher, with the exception of any material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work.

Printed on acid-free paper

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

Welcome Message from the General Chairs ITCS 2011

As the General Chairs of the 3rd Information Technology Convergence and Services (ITCS 2011), we have the pleasure of welcoming you to this conference and to this beautiful city, Gwangju, Korea on October 20–22, 2011.

In past twenty five years or so, IT (Information Technology) influenced and changed every aspect of our lives and our cultures. Without various IT-based applications, we would find it difficult to keep information stored securely, to process information efficiently, and to communicate information conveniently. In the future world, IT will play a very important role in convergence of computing, communication, and all other computational sciences and application and IT also will influence the future world's various areas, including science, engineering, industry, business, law, politics, culture, medicine, and so on.

Our conference is intended to foster the dissemination of state-of-the-art research in all IT convergence areas, including its models, services, and novel applications associated with their utilization. We hope our conference will be the most comprehensive conference focused on the various aspects of advances in all future IT areas and IT-based service, sciences and engineering areas.

We would like to thank all authors of this conference for their paper contributions and presentations. And we would like to sincerely appreciate the following prestigious invited speakers who kindly accepted our invitations, and helped to meet the objectives of the conference:

- Dr. Laurence T. Yang
Department of Computer Science, St. Francis Xavier University, Canada
- Dr. Hamid R. Arabnia
Department of Computer Science, The University of Georgia, USA

- Dr. Hong Shen
School of Computer Science, The University of Adelaide, Australia
- Dr. Hsiao-Hwa Chen
Department of Engineering Science, National Cheng Kung University, Taiwan

We also sincerely thank all our chairs and committees, and these are listed in the following pages. Without their hard work, the success of ITCS 2011 would not have been possible. Finally, we would like to thank the workshop organizers of IRoA 2011, ITMUE 2011, PCT 2011, SAE 2011 and Smartphone 2011, for their great contributions.

With best regards,

Looking forward to seeing you at ITCS 2011

Hamid R. Arabnia, University of Georgia, USA

Hangbae Chang, Daejin University, Korea

General Chairs

Welcome Message from the Program Chairs

ITCS 2011

We would like to extend our welcome and express our gratitude to all of the authors of submitted papers and to all of the attendees, for contributions and participations.

In ITCS 2011, the 3rd international conference has attracted 97 papers. The international character of the conference is reflected in the fact that submissions came from various countries.

The submitted abstracts and papers went through a through reviewing process. As a result, 34 articles were accepted (acceptance rate: 35%) for the ITCS 2011 proceedings published by Springer, reflecting (but not limited to) the following areas:

- Track 1. Advanced Computational Science and Applications
- Track 2. Advanced Electrical and Electronics Engineering and Technology
- Track 3. Intelligent Manufacturing Technology and Services
- Track 4. Advanced Management Information Systems and Services
- Track 5. Electronic Commerce, Business and Management
- Track 6. Intelligent Vehicular Systems and Communications
- Track 7. Bio-inspired Computing and Applications
- Track 8. Advanced IT Medical Engineering
- Track 9. Modeling and Services for Intelligent Building, Town, and City

And some papers were invited from Chairs and Committee members to be included in our ITCS 2011 proceedings.

Achieving such a high quality of proceedings would have been impressive without the huge work that was undertaken by the international Program Committee members. We take the opportunity to thank them for their great support and cooperation.

Sincerely yours,

Il-sun You, Korean Bible University, Korea

Sajid Hussain, Fisk University, USA

Bernady O. Apduhan, Kyushu Sangyo University, Japan

Zhiwen Yu, Northwestern Polytechnical University, China

Program Chairs

Conference Organization

ITCS 2011

Organizing Committee

Steering Co-Chair

James J. (Jong Hyuk) Park, Seoul National University of Science and Technology, Korea

General Chairs

Hamid R. Arabnia, University of Georgia, USA

Hangbae Chang, Daejin University, Korea

General Vice Chair

Changhoon Lee, Hanshin University, Korea

Program Chairs

Il-sun You, Korean Bible University, Korea

Sajid Hussain, Fisk University, USA

Bernady O. Apduhan, Kyushu Sangyo University, Japan

Zhiwen Yu, Northwestern Polytechnical University, China

Workshop Chairs

Naveen Chilamkurti, La Trobe University, Australia

Yang Sun Lee, Chosun University, Korea

Leomar S. Rosa Junior, Federal University of Pelotas, Brazil

International Advisory Board Committee

Mohammad S. Obaidat, Monmouth University, USA

Sang-Soo Yeo, Mokwon University, Korea

Han-Chieh Chao, National Ilan University, TAIWAN, ROC

Andrew Kusiak, The University of Iowa, USA

Publicity Chairs

Neeli Prasad, Aarhus University, Denmark

Qingsong Cai, Beijing Technology and Business University, China

Hongjoo Lee, Kyonggi University, Korea

Nakseon Seong, ETRI, Korea

David Taniar, Monash University, Australia

Sang Oh Park, Chungang University, Korea

Local Arrangement Chairs

Yang-Hoon Kim, Daejin University, Korea

Hyuk-Jun Kwon, Yonsei University, Korea

Registration and Finance Chair

Jonggu Kang, Daejin University, Korea

Program Committee**Track 1. Advanced Computational Science and Applications**

Cheong Ghil Kim, Namseoul University, Korea

Chin-Chen Chang, Chia University, Taiwan

Davy Van Deursen, Universiteit Gent, Belgium

Ghalem Belalem, University of Oran, Algeria

László Horváth, Óbuda University, Hungary

Maumita Bhattacharya, Charles Sturt University, Australia

Michael Schwarz, Universitat Kassel, Germany

MohammadReza Keyvanpour, Alzahra University, Iran

Ruck Thawonmas, Ritsumeikan University, Japan

Russel Pears, AUT University, New Zealand

Sanja Maravic Cisar, College of Subotica, Serbia

Sergio Pozo Hidalgo, University of Sevilla, Spain

Viktoria Villanyi, Florida Atlantic University, USA

Wolfgang Schreiner, Johannes Kepler University, Austria

Xiangyang Luo, Information Science and Technology Institute, China

Yih-Chuan Lin, National Formosa University, Taiwan

Yuan-Ko Huang, Kao Yuan University, Taiwan

Zheng, Edinburgh University, UK

Zhihui Du, Tsinghua University, China

Zsolt Csaba Johanyák, Kecskemét College, Hungary

Track 2. Advanced Electrical and Electronics Engineering and Technology

Eva Cheng, RMIT University, Australia

Feng Chen, Tsinghua University, China

Kilhung Lee, Seoul National University of Science & Technology, Korea

Somkait Udomhunsakul, Rajamangala University of Technology Suvarnabhumi

Xinghao Jiang, New Jersey Institute of Technology, USA

Jin Kwak, Department of Information Security Engineering, Soonchunhyang University, Korea

Deok-Gyu Lee, Electronics and Telecommunications Research Institute, Korea

Seungmin Rho, Korea University, Korea

Soon Seok Kim, Department of Computer Engineering, Halla University, Korea

Sangyup Nam, Kookje College, Korea

Hyukjun Kwan, Yonsei University, Korea

Yunjae Lee, SK C&C, Korea

Taewoo Roh, ING, Korea

Track 3. Intelligent Manufacturing Technology and Services

Gunter Saake, University of Magdeburg, Germany

Jinjun Chen, Swinburne University of Technology, Australia

Yao Chung Chang, National Taitung University, Taiwan

Yiannis Kompatsiaris, Informatics and Telematics Institute Centre for Research and Technology Hellas, Greece

Wansoo Kim, LG CNS, Korea

Byungsoo Ko, DigiCAP, Korea

Younggui Jung, Y.M-Naeultech, Korea

ChulUng Lee, Korea University, Korea

Heesuk Seo, Korea University of Technology and Education, Korea

Kae-Won Choi, SeoulTech, Korea

Track 4. Advanced Management Information Systems and Services

Bill Grosky, University of Michigan-Dearborn, USA

MarcoFurini, University of Bologna, Italy

Mudasser Wyne, National University, USA

Soocheol Lee, Korea Intellectual Property Office, Korea

Porandokht Fazelian, IT Manager at the TMU, Tehran

Tomoo Inoue, University of Tsukuba, Japan

William Grosky, University of Michigan, USA

Zhaobin Liu, Dalian Maritime University, China

Track 5. Electronic Commerce, Business and Management

Geguang Pu, East China Normal University, China

Grizalis Stefanos, University of the Aegean, Greece

Raymond Choo, Australian Institute of Criminology, Australia

Somchart Fugkeaw, Thaidigitalid, Thailand

Track 6. Intelligent Vehicular Systems and Communications

Chao-Tung Yang, Tunghai University, Taiwan

Fazle Hadi, King Saud University, Saudi Arabia

Hanáček Petr, Brno University of Technology, Czech Republic

Yuliya Ponomarchuk, Kyungpook National University, Korea

Min Choi, Department of Computer Engineering, Wonkwang University, Korea

Seung-Ho Lim, Hankuk University of Foreign Studies, Korea
Nak-Seon Seong, Electronics and Telecommunications Research Institute, Korea
Kilhung Lee, Seoul National University of Science and Technology, Korea
Hyo Hyun Choi, Department of Computer Science, Inha Technical College, Korea

Track 7. Bio-inspired Computing and Applications

Albert Zomaya, The University of Sydney, Australia
Alina Patelli, Gheorghe Asachi Technical University of Iasi, Romania
Debnath Bhattacharyya, West Bengal University of Technology, India
Lavi Ferariu, Gheorghe Asachi Technical University of Iasi, Romania
Rahim A. Abbaspour, University of Tehran, Iran
Satoshi Kurihara, Osaka University, Japan
Namsoo Chang, Sejong Cyber University, Korea
SeungTaek Ryoo, Hanshin University, Korea
Hae Young Lee, Electronics and Telecommunications Research Institute, Korea
Dong Kyoo Kim, Electronics and Telecommunications Research Institute, Korea

Track 8. Advanced IT Medical Engineering

Ajaz Hussain Mir, National Institute of Technology, India
Ovidiu Ghiba, Politehnica University of Timisoara, Romania
Ryszard Choras, EE of University of Technology & Life Sciences, Poland
Jiann-Liang Chen, National Taiwan University of Science and Technology, Taiwan
Georgios Kambourakis, University of the Aegean, Greece
Ilias Maglogiannis, University of Central Greece
Jansen Bart, Vrije Universiteit Brussel, Belgium
Wei Chen, Eindhoven University of Technology, Netherlands

Track 9. Modeling and Services for Intelligent Building, Town, and City

Chuang-Wen You, National Taiwan University, Taiwan
Laurent Gomez, SAP Labs France SAS, France
Pereira Rubem, Liverpool John Moores University, UK
Robert Meurant, The Institute of Traditional Studies, USA
Dongho Kim, Halla University, Korea
Yong-hee Lee, Halla University, Korea
Hyunsung Kim, Kyungil University, Korea

Welcome Message from the Workshop Chairs

ITCS 2011

It is a great pleasure to present the technical programs of the workshops held in conjunction with the 3rd Technology Convergence and Services (ITCS 2011), Kimdaejung Convention Center, Gwangju, Korea. The main aim of these workshops is to bring together academics, industry researchers and practitioners to discuss and share experience on completed and on-going research activities in the areas of intelligent robotics, automations, telecommunication facilities, and applications, technology and multimedia for ubiquitous environments, personal computing technologies, security and application for embedded systems, and smartphone applications and services. The workshops constitute an important extension of the main conference by providing a forum for discussions on focused areas that the main conference. We believe the forum will facilitate active discussions among researchers in information technologies.

The five selected workshops held in conjunction with ITCS 2011 are:

1. International Conference on Intelligent Robotics, Automations, telecommunication facilities, and applications (IRoA 2011)
2. International Workshop on Information Technology and Multimedia for Ubiquitous Environments (ITMUE 2011)
3. International Workshop on Personal Computing Technologies (PCT 2011)
4. International Workshop on Security and Application for Embedded systems (SAE 2011)
5. International Workshop on Smartphone Applications and Services (Smartphone 2011)

Among the five successful workshops, each workshop deals with various topics related to Information Technology. All the submitted papers have undergone rigorous review process by the technical program committee members for originality, contribution and relevance to the main themes of the conference. We have selected 39 best papers for presentation and publication in the conference proceedings.

As workshop chairs we wish to thank all the organizers of the workshops and the international technical committee members for their professional support. We would also like to express our gratitude to all Organizing Committee members of ITCS 2011. In particular, we would like to thank the ITCS 2011 General Chairs, Prof. Hamid R. Arabnia, and Hangbae Chang and General Vice Chair, Prof. Changhoon Lee and Program Chairs, Prof. Ilsun You, Sajid Hussain, Bernady O. Apduhan, and Zhiwen Yu. Last but not least, we would also like to thank the Steering Co-Chair, Prof. James J. (Jong Hyuk) Park for coordinating the entire conference event.

Naveen Chilamkurti, La Trobe University, Australia

Yang Sun Lee, Chosun University, Korea

Leomar S. Rosa Junior, Federal University of Pelotas, Brazil

Workshop Chairs

IRoA 2011 Welcome Message from Workshop Organizers ITCS 2011

It is our pleasure to welcome you The 2011 FTRA International Conference on Intelligent Robotics, Automations, telecommunication facilities, and applications (IRoA-11) held in Gwangju, Korea, October 20–22, 2011.

The 2011 FTRA International Conference on Intelligent Robotics, Automations, telecommunication facilities, and applications (IRoA-11), co-sponsored by FTRA will be held in Gwangju, Korea, October 20–22, 2011. The IRoA is a major forum for scientists, engineers, and practitioners throughout the world to present the latest research, results, ideas, developments and applications in all areas of intelligent robotics and automations. Furthermore, we expect that the IRoA-11 and its publications will be a trigger for further related research and technology improvements in this important subject. The IRoA-11 is co-sponsored by FTRA. In addition the conference is supported by KITCS.

We would like to send our sincere appreciation to all participating members who contributed directly to IRoA 2011. We would like to thank all Program Committee members for their excellent job in reviewing the submissions. We also want to thank the members of the organizing committee, all the authors and participants for their contributions to make IRoA 2011 a grand success.

James J. (Jong Hyuk) Park and Shigeki Sugano
IRoA 2011 Chairs

Workshop General Chairs

James J. (Jong Hyuk) Park, SeoulTech, Korea
Shigeki Sugano, Waseda University, Japan

General Vice Chair

Sang-Soo Yeo, Division of Computer Engineering, Mokwon University, Korea

Program Chairs

Taeshik Shon, Ajou University, Korea (Leading Chair)

Ken Chen, Tsinghua University, Beijing, China

Honghai Liu, University of Portsmouth, UK

Workshop Chairs

Sang Oh Park, Chung-Ang University, Korea

Jiming Chen, Zhejiang University, China

Publicity Chairs

Uche Wejinya, University of Arkansas, USA

Lianqing Liu, Chinese Academy of Sciences, China

Yunhui Liu, Chinese University of HK, China

Kazuhito Yokoi, AIST, Japan

Sang Oh Park, Chung-Ang University, Korea

International Advisory Committee

Marco Ceccarelli, University of Cassino, Italy

Panos J. Antsaklis, University of Notre Dame, USA

Kok-Meng Lee, Georgia Institute of Technology, USA

Tzyh-Jong Tarn, Washington University, USA

Kazuhiro Saitu, University of Michigan, USA

David Atkinson, Air Force Office of Scientific Research, USA

Local Arrangement Chairs

Yang Sun Lee, Chosun University, Korea

Registration / Finance Chair

Changhoon Lee, Hanshin University, Korea

Web and System Management Chair

Kyusuk Han, KAIST, Korea

Program Committee

Abdel AITOUICHE, Hautes Etudes d'Ingenieur, France

Abdel-Badeeh Salem, Ain Shams University, Egypt

Adil Baykasoglu, University of Gaziantep, Turkey

Ahmed Zobia, Brunel University, UK

Ajay Gopinathan, University of California, Merced, USA

Alessandra Lumini, University of Bologna, Italy

Alessandro Giua, Università di Cagliari, Italy

Alexandre Dolgui, Ecole Nationale Suprieure des Mines de Saint Etienne, Italy

Andreas C. Nearchou, University of Patras, Greece

Angel P. del Pobil, Universitat Jaume I, Spain

Angelos Amanatiadis, Democritus University of Thrace, Ksanthi, Greece

Anthony A. Maciejewski, Colorado State University, USA
Anthony Tzes, University of Patras, Greece
Anton Nijholt, University of Twente, Netherlands
Antonios Tsourdos, Cranfield University, UK
Arijit Bhattacharya, Dublin City University, Ireland
Arvin Agah, The University of Kansas, USA
Asokan Thondiyath, Indian Institute of Technology Madras
Barry Lennox, The University of Manchester, UK
Ben-Jye Chang, Chaoyang University of Technology, Taiwan
Bernard Brogliato, INRIA, France
Bernardo Wagner, University of Hannover, Germany
Carla Seatzu, University of Cagliari, Italy
Carlo Alberto Avizzano, Scuola Superiore S. Anna, Italy
Carlo Menon, Simon Fraser University, Canada
Cecilia Sik Lanyi, University of Pannonia, Hungary
Ching-Cheng Lee, Olivet University & California State University at East Bay, USA
Choon Yik Tang, University of Oklahoma, USA
Chunling Du, Division of Control & Instrumentation School of Electrical & Electronic Engineering, Singapore
Clarence de Silva, UBC, Canada
Claudio Melchiorri, University of Bologna, Italy
Daizhan Cheng, Academy of Mathematics and Systems Science, China
Dan Zhu, Iowa State University, USA
Daniel Thalmann, EPFL Vrlab, Switzerland
Denis Dochain, Université catholique de Louvain, Belgium
Dianhui Wang, La Trobe University, Australia
Djamila Ouelhadj, University of Portsmouth, UK
Dongbing Gu, University of Essex, UK
Eloisa Vargiu, University of Cagliari, Italy
Erfu Yang, University of Strathclyde, UK
Evangelos Papadopoulos, NTUA, Greece
Fang Tang, California State Polytechnic University, USA
Federica Pascucci, University of Roma Tre, Italy
Frank Allgower, University of Stuttgart, Germany
Frans Groen, University of Amsterdam, Netherlands
Frantisek Capkovic, Slovak Academy of Sciences, Slovak Republic
Fumiya Iida, Saarland University, Germany
George L. Kovacs, Hungarian Academy of Sciences, Hungary
Gerard Mckee, The University of Reading, UK
Gheorghe Lazea, Technical University of Cluj-Napoca, Romania
Giovanni Indiveri, University of Salento, Italy
Graziano Chesi, University of Hong Kong, Hong Kong
Guilherme N. DeSouza, University of Missouri-Columbia, USA
Gurvinder S Virk, Massey University, New Zealand
Hairong Qi, University of Tennessee, USA

Helder Araujo, University of Coimbra, Portugal
Helen Ryaciotaki-Boussalis, California State University, Los Angeles, USA
Hemant A. Patil, Gandhinagar, Gujarat, India
Hideyuki Sotobayashi, Aoyama Gakuin University, Japan
Hiroyasu Iwata, Waseda University, Japan
Hongbin Zha, Peking University, China
Huei-Yung Lin, National Chung Cheng University, Taiwan
Hung-Yu Wang, Kaohsiung University of Applied Sciences, Taiwan
Ichiro Sakuma, The University of Tokyo, Japan
Irene Yu-Hua Gu, Chalmers University of Technology, Sweden
Jean-Daniel Dessimoz, Western University of Applied Sciences, Switzerland
Jing-Sin Liu, Institute of Information Science, Academia Sinica, Taiwan
Jingang Yi, The State University of New Jersey, USA
Junn-Lin Wu, National Chung Hsing University, Taiwan
Jonghwa Kim, University of Augsburg, Germany
Joris De Schutter, Katholieke Universiteit Leuven, Belgium
Jose Tenreiro Machado, Institute of Engineering of Porto
José Valente de Oliveira, Universidade do Algarve, Portugal
Juan J. Flores, University of Michoacan, Mexico
Jun Ota, The University of Tokyo, Japan
Jun Takamatsu, Nara Institute of Science and Technology, Japan
Kambiz Vafai, University of California, Riverside, USA
Karsten Berns, University of Kaiserslautern, Germany
Kauko Leiviskä, University of Oulu, Finland
Lan Weiyao, Department of Automation, Xiamen University, China
Leonardo Garrido, Monterrey Tech., Mexico
Libor Preucil, Czech Technical University in Prague, CZ
Loulin Huang, Massey University, New Zealand
Luigi Villani, University di Napoli Federico II, Italy
Mahasweta Sarkar, San Diego State University, USA
Maki K. Habib, Saga University, Japan
Manuel Ortigueira Faculdade de Cinciase, Tecnologia da Universidade Nova de Lisboa, Portugal
Marek Zaremba, UQO, Canada
Maria Gini, University of Minnesota, USA
Mario Ricardo Arbulu Saavedra, University Carlos III of Madrid, Spain
Masao Ikeda, Osaka University, Japan
Matthias Harders, Computer Vision Laboratory ETH Zurich, Switzerland
Mehmet Sahinkaya, University of Bath, UK
Michael Jenkin, York University, Canada
Mitsuji Sampei, Tokyo Institute of Technology, Japan
Nitin Afzulpurkar, Asian Institute of Technology, Thailand
Olaf Stursberg, Technische Universitaet Muenchen, Germany
Panagiotis Petratos, California State University, Stanislaus, USA
Pani Chakrapani, University of Redlands, USA

Paul Oh, Drexel University, USA
Peng-Yeng, National Chi Nan University, Taiwan
Peter Xu, Massey University, New Zealand
Pieter Mosterman, The Mathworks, Inc.
Plamen Angelov, Lancaster University, UK
Prabhat K. Mahanti, University of New Brunswick, Canada
Qinggong Meng, Research School of Informatics, UK
Qurban A Memon, United Arab Emirates University, UAE
Radu Bogdan Rusu, Technical University of Munich, Germany
Ragne Emardson, SP Technical Research Institute of Sweden
Ren C. Luo, National Taiwan University, Taiwan
Rezia Molfino, Università degli Studi di Genova, Italy
Riad I. Hammoud DynaVox and Mayer-Johnson, Innovation Group, USA
Richard J. Duro, Universidade da Coruña, Spain
Robert Babuska, Delft University of Technology, Netherlands
Rolf Johansson, Lund University, Sweden
Romeo Ortega, LSS Supelec, France
Ruediger Dillmann, University of Karlsruhe, Germany
Ryszard Tadeusiewicz, AGH University of Science and Technology, Poland
Saeid Nahavandi, Alfred Deakin Professor; Director, CISR, New Zealand
Sarath Kodagoda, University of Technology, Sydney, Australia
Sean McLoone, National University of Ireland (NUI) Maynooth, Ireland
Selahattin Ozcelik, University-Kingsville, USA
Sergej Fatikow, University of Oldenburg, Germany
Seth Hutchinson, University of Illinois, USA
Shu-Ching Chen, Florida International University, USA
Shugen Ma, Ritsumeikan University, Japan
Shuro Nakajima, Chiba Institute of Technology, Japan
Shuzhi Sam Ge, National University of Singapore, Singapore
Simon G. Fabri, University of Malta, Malta
Stjepan Bogdan, University of Zagreb, Faculty of EE&C, Croatia
Tariq Shehab, California State University, Long Beach, USA
Taskin Padir, Worcester Polytechnic Institute, USA
Thira Jearsiripongkul, Thammasat University, Thailand
Thomas C. Henderson, University of Utah, USA
Tomonari Furukawa, Virginia Polytechnic Institute and State University, USA
Tong Heng Lee, NUS, Singapore
Tongwen Chen, University of Alberta, Canada
Tsai-Yen Li, National Chengchi University, Taiwan
Uwe R. Zimmer, The Australian National University, Australia
Venketesh N Dubey, Bournemouth University, UK
Ventzeslav (Venny) Valev, Bulgarian Academy of Sciences, Bulgaria
Wail Gueaieb, University of Ottawa, Canada
Wang Qing-Guo, National University of Singapore, Singapore
Waree Kongprawechnon, Thammasat University, Thailand

Weihua Sheng, Oklahoma State University, USA
Weizhong Dai, Louisiana Tech University, USA
Wen-Hua Chen, Loughborough University, UK
Wladyslaw Homenda, Warsaw University of Technology, Poland
Wolfgang Halang, Fernuniversitaet, Germany
Won-jong Kim, Texas A&M University, USA
Xun W Xu, University of Auckland, New Zealand
Yang Dai, University of Illinois at Chicago, USA
Yangmin Li, University of Macau, Macao
Yantao Shen, University of Nevada, USA
Yugeng Xi, Shanghai Jiaotong University, China
Yun Wang, University of California, Irvine, USA
Zhijun Yang, University of Edinburgh, UK
Zidong Wang, Brunel University, UK
Zongli Lin, University of Virginia, USA
Vasily Sachnev, The Catholic University of Korea, Korea
Elena Tsomko, Namseoul University, Korea
Jin Young Kim, Kwangwoon University, Korea
Ki-Hyung Kim, Ajou University, Korea
Namme Moon, Hoseo University, Korea
Taesam Kang, Konkuk University, Korea
Hwa-Jong Kim, Kangwon National University, Korea
Yeonseok Lee, Kunsan National University, Korea
Cheong Ghil Kim, Namseoul University, Korea
Sanghyun Joo, ETRI, Korea
Wei Wei, Xi'an Jiaotong University, China

ITMUE 2011 Welcome Message from Workshop Organizers ITCS 2011

It is our pleasure to welcome you to The FTRA International Workshop on Information Technology and Multimedia for Ubiquitous Environments (ITMUE 2011), held in Gwangju, Korea, October 20–22.

The ITMUE 2011 provides a forum for academic and industry professionals to present novel ideas on ITMUE. We expect that the ITMUE technologies have become state-of-the-art research topics and are expected to play an important role in human life in the future. ITMUE 2011 aims to advance ubiquitous multimedia techniques and systems research, development, and design competence, and to enhance international communication and collaboration. The workshop covers traditional core areas of information technology and multimedia for ubiquitous and Intelligent Recommendation and Personalization.

We would like to send our sincere appreciation to all participating members who contributed directly to ITMUE 2011. We would like to thank all Program Committee members for their excellent job in reviewing the submissions. We also want to thank the members of the organizing committee, all the authors and participants for their contributions to make ITMUE 2011 a grand success.

Yanming Shen, Ali Asghar Nazari Shirehjini, Jung-Sik Cho
ITMUE 2011 Chairs

Workshop Chairs

Yanming Shen, Dalian University of Technology, China

Ali Asghar Nazari Shirehjini, University of Ottawa, Canada

Jung-Sik Cho, Chuang-Ang University, Korea

Program Committee

Tobias Bürger, Capgemini SD&M, Germany

Yiwei Cao, RWTH Aachen, Germany

Minoru Uehara, Toyo University, Japan

Fatos Xhafa, Polytechnic University of Catalonia, Spain

Muhammad Younas, Oxford Brookes University, UK

- Hahn Le**, University of Cape Town, South Africa
Qun Jin, Waseda University, Japan
Hung-Min Sun, National Tsing Hua University, Taiwan
Li-Ping Tung, Academia Sinica, Taiwan
Eric Na, LG Electronics, Korea
S. Raviraja, University of Malaya, Malaysia
Matthias Rauterberg, Eindhoven University of Technology, Netherlands
Claudio Biancalana, Roma Tre University, Roma
Ernesto William De Luca, TU Berlin, Germany
Hao Wang, Nokia Research Center, China
Hyuk Cho, Sam Houston State University, USA
Jingyu Sun, Taiyuan University of Technology, China
Marius Silaghi, Florida Institute of Technology, USA
Nurmamat Helil, Xinjiang University, China
Okkyung Choi, Sejong University, Korea
Se Joon Park, SK C&C, Korea
Seunghwan Kim, Korea Atomic Energy Research Institute, Korea
Sten Govaerts, Katholieke Universiteit Leuven, Belgium
Yangjin Seo, SECUI, Korea

PCT 2011 Welcome Message from Workshop Organizers ITCS 2011

On behalf of the 2011 International Workshop on Personal Computing Technologies (PCT 2011), we are pleased to welcome you to Gwangju, Korea.

The workshop will foster state-of-the-art research in the area of personal computing technologies. The PCT 2011 will also provide an opportunity for academic and industry professionals to discuss the latest issues and progress in the area of personal computing technologies.

Due to many high quality paper submissions and the lack of space in proceedings, the review process was very tough and we had no choice but to reject several good papers. Finally, we would like to sincerely express gratitude to all the people who have contributed directly or indirectly to make PCT 2011 a grand success. We would like to express our appreciation to all TPC members for the valuable time and their professional supports to this workshop. Particularly, we would like to thank ITCS 2011 General Chairs (Prof. Hang-Bae Chang and Prof. Hamid R. Arabnia) who allow us to hold this workshop in conjunction with ITCS 2011.

Thank you

Jeunwoo Lee, Electronics and Telecommunications Research Institute, Korea
Changseok Bae, Electronics and Telecommunications Research Institute, Korea
Chanik Park, Pohang University of Science and Technology, Korea
PCT 2011 Chairs

General Chair

Jeunwoo Lee, Electronics and Telecommunications Research Institute, Korea

Workshop Chairs

Changseok Bae, Electronics and Telecommunications Research Institute, Korea
Chanik Park, POSTECH, Korea

Program Committee

Dong-oh Kang, Electronics and Telecommunications Research Institute, Korea

Yuk Ying Chung, University of Sydney, Australia

Xiang Jian He, University of Technology Sydney, Australia

Wei-Chang Yeh, National TsingHua University, Taiwan

Jinho Yoo, Baekseok University, Korea

Mohd Afizi Mohd Shukran, National Defense University of Malaysia, Malaysia

Noorhaniza Wahid, University Tun Hussein Onn Malaysia (UTHM), Malaysia

SAE 2011 Welcome Message from Workshop Organizers ITCS 2011

On behalf of the FTRA International Workshop on Security and Application for Embedded systems (SAE 2011), we are pleased to welcome you to Gwangju, Korea.

The SAE 2011 will be the most comprehensive workshop focused on the various aspects of Security and Application for Embedded smart systems (SAE 2011). The SAE 2011 provides a forum for academic and industry professionals to present novel ideas on SAE. We expect that the workshop and its publications will be a trigger for further related research and technology improvements in this important subject.

We would like to thank many people who have generously made contributions for this workshop. First of all, we thank the Program Committee members for their excellent job in reviewing the submissions and thus guaranteeing the quality of the workshop under a very tight schedule. We also want to thank the members of the organizing committee, all the authors and participants for their contributions to make SAE 2011 a grand success.

Jongsung Kim, Sang Oh Park, Jung-Sik Cho
SAE 2011 General and Program Chairs

General Chair

Jongsung Kim, Kyungnam University, Korea

Program Chairs

Sang Oh Park, Chuang-Ang University, Korea

Jung-Sik Cho, Chuang-Ang University, Korea

Program Committee

Axel Poschmann, Nanyang Technological University, Singapore

Emmanuelle Anceaume, IRISA, France

Frederik Armknecht, Institute for Computer Science at the University of Mannheim, Germany

- Guy Gogniat**, Universite de Bretagne Sud, France
Houcine Hassan, Polytechnic University of Valencia, Espana
Kris Gaj, George Mason University, USA
Kurt Rothermel, University of Stuttgart, Germany
Leandro Buss Becker, Federal University of Santa Catarina (UFSC), Brazil
Meng-Yen Hsieh, Providence University, Taiwan
Pinit Kumhom, King Mongkut's University of Technology Thonburi, Thailand
Raimund Kirner, Vienna University of Technology, Austria
Shangping Ren, Illinois Institute of Technology, USA
Shlomi Dolev, Ben Gurion University, Israel
Srinivasa Vemuru, Ohio Northern University, USA
Thumrongrat Amornraksa, King Mongkut's University of Technology Thonburi, Thailand
Tilman Wolf, University of Massachusetts Amherst, USA
Zebo Peng, Linkoping University, Sweden
Zhijie Jerry Shi, University of Connecticut, USA

Smartphone 2011 Welcome Message from Workshop Organizers ITCS 2011

Welcome to the International Workshop on Smartphone Applications and Services (Smartphone 2011), held in Gwangju, Korea, during October 20–22, 2011. Smartphone 2011 follows on the success of the Smartphone 2010 in Gwangju, Korea, held December 9–11, 2010.

First, we are very grateful to the 3rd FTRA International Conference on Information Technology Convergence and Services (ITCS 2011) organizing committee, which is sponsored by the National IT industry Promotion Agency (NIPA) and the Gwangju Convention & Visitors Bureau, for their support of the Smartphone 2011. It's our great pleasure to include the papers of Smartphone 2011 in the ITCS 2011 proceedings.

Smartphone 2011 is the second-year event of the Smartphone conference series and it has attracted a small number of submissions. Nevertheless, all submitted papers have undergone blind reviews by at least three reviewers from the technical program committee, which consists of leading researchers from around the globe. Without their hard work, achieving such high-quality proceedings would not have been possible. We take this opportunity to thank them for their great support and cooperation.

We hope the Smartphone 2011 will be the most comprehensive workshop focused on advances in Smartphone applications and services. This year's Smartphone event is very small, but we are sure that the conference will provide an opportunity for academic and industry professionals to discuss the latest issues and progress in the areas of mobile technologies that includes highly capable handheld device or cell-phone with advanced features such as iPhone OS, Android, Linux Mobile, Windows Mobile/Phone operation system, access to the Internet, and other computer-like processing capabilities similar to personal computer.

We would like to thank many people who have generously made contributions for this workshop. First of all, we thank the Program Committee members for their excellent job in reviewing the submissions and thus guaranteeing the quality of the workshop under a very tight schedule. We also want to thank the members of the organizing committee.

Finally, we would like to thank all of the authors and participants for their contributions to make the Smartphone 2011 a grand success.

James J. (Jong Hyuk) Park , Sang Oh Park, Fernando Ferri
Smartphone 2011 General and Program Chairs

Steering Chair

James J. (Jong Hyuk) Park, Seoul National University of Science and Technology,
Korea

Program Chairs

Sang Oh Park, Chuang-Ang University, Korea

Fernando Ferri, IRPPS-CNR, Rome, Italy

Publicity Chairs

Jung-Sik Cho, Chuang-Ang University, Korea

Taeshik Shon, Ajou University, Korea

Nitendra Rajput, IBM Research, India

Program Committee

Alexander De Luca, Ludwig-Maximilians-Universitat, Germany

Ana Belen Lago, University of Deusto, Spain

Chan Yeun Yeob, Khalifa University of Science Technology and Research, UAE

Deborah Dahl, Conversational Technologies, USA

Deok Gyu Lee, ETRI, Korea

Edward Hua, QED Systems, USA

Florian Michahelles, ETH Zurich, Switzerland

Jeong Heon Kim, Chung-Ang University, Korea

Jeong Hyun Yi, Soongsil University, Korea

Jonathan M. McCune, Carnegie Mellon University, USA

Jongsub Moon, Korea University, Korea

Jose A. Onieva, University of Malaga, Spain

Kyusuk Han, KAIST, Korea

Mark Billingham, University of Canterbury, New Zealand

Mark Shaneck, Liberty University, USA

Michael Rohs, Ludwig Maximilian University of Munich, Germany

Mucheol Kim, Chung-Ang University, Korea

Oliver Amft, Eindhoven University of Technology, Netherlands

Rene Mayrhofer, University of Vienna, Austria

Ruben Rios del Pozo, University of Malaga, Spain

Soo Cheol Kim, Chung-Ang University, Korea

Thomas Strang, German Aerospace Center (DLR), Germany

Thomas Wook Choi, Hankuk University of Foreign Studies, Korea

Vishal Kher, VMware, USA

Yong Lee, ChungJu University, Korea

Contents

Part I IT Convergence and Services

Analysis of Security Vulnerability and Authentication Mechanism in Cooperative Wireless Networks	3
Ki Hong Kim	
Spam Host Detection Using Ant Colony Optimization.	13
Arnon Rungsawang, Apichat Taweessiriwate and Bundit Manaskasemsak	
Location Estimation of Satellite Radio Interferer Using Cross Ambiguity Function Map for Protection of Satellite Resources	23
Chul-Gyu Kang, Chul-Sun Park and Chang-Heon Oh	
Korean Voice Recognition System Development	31
Soon Suck Jang	
Availability Management in Data Grid	43
Bakhta Meroufel and Ghalem Belalem	
Mobi4D: Mobile Value-Adding Service Delivery Platform	55
Ishmael Makitla and Thomas Fogwill	
The Security Management Model for Small Organization in Intelligence All-Things Environment	69
Hangbae Chang, Jonggu Kang and Youngsub Na	
Simulation Modeling of TSK Fuzzy Systems for Model Continuity . . .	77
Hae Young Lee, Jin Myoung Kim, Ingeol Chun, Won-Tae Kim and Seung-Min Park	

A New Method of Clustering Search Results Using Frequent Itemsets with Graph Structures 87
I-Fang Su, Yu-Chi Chung, Chiang Lee and Xuanyou Lin

A Data Gathering Scheme Using Mobile Sink Dynamic Tree in Wireless Sensor Networks 99
Kilhung Lee

An Enhanced Resource Control Scheme for Adaptive QoS over Wireless Networks for Mobile Multimedia Services 109
Moonsik Kang and Kilhung Lee

An Analysis of Critical Success Factor of IT based Business Collaboration Network Implementation 119
Hangbae Chang, Hyukjun Kwon and Jaehwan Lim

Study of Generating Animated Character Using the Face Pattern Recognition. 127
Seongsoo Cho, Bhanu Shrestha, Bonghwa Hong and Hwa-Young Jeong

Enhancing Performance of Mobile Node Authentication with Practical Security 135
Kyunuk Han and Taeshik Shon

A Study on Turbo Coded OFDM System with SLM for PAPR Reduction 141
Mashhur Sattorov, Sang-Soo Yeo and Heau-Jo Kang

A Context Information Management System for Context-Aware Services in Smart Home Environments 149
Jong Hyuk Park

Enhanced Security Scheme for Preventing Smart Phone Lost Through Remote Control. 157
Jae Yong Lee, Ki Jung Yi, Ji Soo Park and Jong Hyuk Park

SSP-MCloud: A Study on Security Service Protocol for Smartphone Centric Mobile Cloud Computing 165
Ji Soo Park, Ki Jung Yi and Jong Hyuk Park

Self-Adaptive Strategy for Zero-Sum Game 173
Keonsoo Lee, Seungmin Rho and Minkoo Kim

Effect of Light Therapy of Blue LEDs Irradiation on Sprague Dawley Rat 181
 Taegon Kim, Yongpil Park, Yangsun Lee and Minwoo Cheon

Fast Cancer Classification Based on Mass Spectrometry Analysis in Robust Stationary Wavelet Domain 189
 Phuong Pham, Li Yu, Minh Nguyen and Nha Nguyen

Part II Future Security Technologies

An Improved User Authentication Scheme for Wireless Communications 203
 Woongryul Jeon, Jeeyeon Kim, Junghyun Nam, Youngsook Lee and Dongho Won

An Improved Protection Profile for Multifunction Peripherals in Consideration of Network Separation. 211
 Changbin Lee, Kwangwoo Lee, Namje Park and Dongho Won

Security Improvement to an Authentication Scheme for Session Initiation Protocol 221
 Youngsook Lee, Jeeyeon Kim, Junghyun Nam and Dongho Won

A Study on the Development of Security Evaluation Methodology for Wireless Equipment. 231
 Namje Park, Changwhan Lee, Kwangwoo Lee and Dongho Won

Computer Application in Elementary Education Bases on Fractal Geometry Theory Using LOGO Programming 241
 Jaeho An and Namje Park

Construction of a Privacy Preserving Mobile Social Networking Service 251
 Jaewook Jung, Hakhyun Kim, Jaesung You, Changbin Lee, Seungjoo Kim and Dongho Won

Part III IT–Agriculture Convergence

Standardization Trend of Agriculture-IT Convergence Technology in Korea 265
 Se-Han Kim, Chang Sun Shin, Cheol Sig Pho, Byung-Chul Kim and Jae-Yong Lee

Design and Implementation of Greenhouse Control System Based IEEE802.15.4e and 6LoWPAN 275
 Se-Han Kim, Kyo-Hoon Son, Byung-Chul Kim and Jae-Yong Lee

Accuracy Estimation of Hybrid Mode Localization Method Based on RSSI of Zigbee 285
 HoSeong Cho, ChulYoung Park, DaeHeon Park and JangWoo Park

A Study on the Failure-Diagnostic Context-Awareness Middleware for Wireless Sensor Networks 295
 In-Gon Park and Chang-Sun Shin

Livestock Searching System on Mobile Devices Using 2D-Barcode 305
 ChulYoung Park, HoSeong Cho, DaeHeon Park, ChangSun Shin, Yong Yun Cho and JangWoo Park

Towards a Context Modeling for a Greenhouse Based on USN 315
 Daeheon Park, Kyoungyong Cho, Jangwoo Park and Yongyun Cho

Ad-Hoc Localization Method Using Ranging and Bearing 321
 Jang-Woo Park and Dae-Heon Park

Part IV Intelligent Robotics, Automations, Telecommunication Facilities, and Applications

An Improved Localization Algorithm Based on DV-Hop for Wireless Sensor Network 333
 Long Chen, Saeyoung Ahn and Sunshin An

A Design of Intelligent Smart Controller for Object Audio-based User’s Active Control Service 343
 Jong-Jin Jung and Seok-Pil Lee

The Method of Main Vocal Melody Extraction Based on Harmonic Structure Analysis from Popular Song 351
 Chai-Jong Song, Seok-Pil Lee, Kyung-Hack Seo and Hochong Park

The Fusion Matching Method for Polyphonic Music Feature Database 359
 Chai-Jong Song, Seok-Pil Lee, Kyung-Hack Seo and Kang Ryoung Park

Towards an Autonomous Indoor Vehicle: Utilizing a Vision-Based Approach to Navigation in an Indoor Environment 367
Edward Mattison and Kanad Ghose

Artificial Pheromone Potential Field Built by Interacting Between Mobile Agents and RFID Tags 377
Piljae Kim and Daisuke Kurabayashi

Proposed Network Coding for Wireless Multimedia Sensor Network (WMSN) 387
A. A. Shahidan, N. Fisal, Nor-Syahidatul N. Ismail and Farizah Yunus

Alternative Concept for Geometry Factor of Frequency Reuse in 3GPP LTE Networks. 397
Modar Safir Shbat and Vyacheslav Tuzlukov

Cognitive Radio Simplex Link Management for Dynamic Spectrum Access Using GNU Radio 407
M. Adib Sarijari, Rozeha A. Rashid, N. Fisal, A. C. C. Lo, S. K. S. Yusof, N. Hija Mahalin, K. M. Khairul Rashid and Arief Marwanto

Do Children See Robots Differently? A Study Comparing Eye-Movements of Adults vs. Children When Looking at Robotic Faces 421
Eunil Park, Ki Joon Kim and Angel P. del Pobil

Relative Self-Localization Estimation for Indoor Mobile Robot. 429
Xing Xiong and Byung-Jae Choi

Q(λ) Based Vector Direction for Path Planning Problem of Autonomous Mobile Robots. 433
Hyun Ju Hwang, Hoang Huu Viet and TaeChoong Chung

Registered Object Trajectory Generation for Following by a Mobile Robot 443
Md Hasanuzzaman and Tetsunari Inamura

An Improved Algorithm for Constrained Multirobot Task Allocation in Cooperative Robot Tasks. 455
Thareswari Nagarajan and Asokan Thondiyath

Simulation-Based Evaluations of Reinforcement Learning Algorithms for Autonomous Mobile Robot Path Planning. 467
 Hoang Huu Viet, Phyo Htet Kyaw and TaeChoong Chung

Control Mechanism for Low Power Embedded TLB 477
 Jung-hoon Lee

Part V IT Multimedia for Ubiquitous Environments

A Noise Reduction Method for Range Images Using Local Gaussian Observation Model Constrained to Unit Tangent Vector Equality. 485
 Jeong Heon Kim and Kwang Nam Choi

Group-Aware Social Trust Management for a Movie Recommender System 495
 Mucheol Kim, Young-Sik Jeong, Jong Hyuk Park and Sang Oh Park

Collaborative Filtering Recommender System Based on Social Network 503
 Soo-Cheol Kim, Jung-Wan Ko, Jung-Sik cho and Sung Kwon Kim

Considerations on the Security and Efficiency of RFID Systems 511
 Jung-Sik Cho, Soo-Cheol Kim, Sang-Soo Yeo and SungKwon Kim

A Development Framework Toward Reconfigurable Run-time Monitors 519
 Chan-Gun Lee and Ki-Seong Lee

Part VI Personal Computing Technologies

Web Based Application Program Management Framework in Multi-Device Environments for Personal Cloud Computing 529
 Hyewon Song, Eunjeong Choi, Chang Seok Bae and Jeun Woo Lee

Hands Free Gadget for Location Service 537
 Jinho Yoo, Changseok Bae and Jeunwoo Lee

Biologically Inspired Computational Models of Visual Attention for Personalized Autonomous Agents: A Survey 547
 Jin-Young Moon, Hyung-Gik Lee and Chang-Seok Bae

Mobile Health Screening Form Based on Personal Lifelogs and Health Records 557
 Kyuchang Kang, Seonguk Heo, Changseok Bae and Dongwon Han

Remote Presentation for M Screen Service in Virtualization System 565
 Joonyoung Jung and Daeyoung Kim

Lifelog Collection Using a Smartphone for Medical History Form 575
 Seonguk Heo, Kyuchang Kang and Changseok Bae

Simplified Swarm Optimization for Life Log Data Mining 583
 Changseok Bae, Wei-Chang Yeh and Yuk Ying Chung

The Design and Implementation of Web Application Management on Personal Device 591
 Eunjeong Choi, Hyewon Song, Changseok Bae and Jeunwoo Lee

Ad Hoc Synchronization Among Devices for Sharing Contents 597
 Eunjeong Choi, Changseok Bae and Jeunwoo Lee

A Framework for Personalization of Computing Environment Among System on-Demand (SoD) Zones 603
 Dong-oh Kang, Hyungjik Lee and Jeunwoo Lee

Part VII Security and Application for Embedded Smart Systems

Facsimile Authentication Based on MAC 613
 Chavinee Chaisri, Narong Mettripun and Thumrongrat Amornraksa

Dynamic Grooming with Capacity aware Routing and Wavelength Assignment for WDM based Wireless Mesh Networks 621
 Neeraj Kumar, Naveen Chilamkurti and Jongsung Kim

Weakness in a User Identification Scheme with Key Distribution Preserving User Anonymity 631
 Taek-Youn Youn and Jongsung Kim

A Compact S-Box Design for SMS4 Block Cipher 641
 Imran Abbasi and Mehreen Afzal

Part VIII Smartphone Applications and Services

iTextMM: Intelligent Text Input System for Myanmar Language on Android Smartphone 661
 Nandar Pwint Oo and Ni Lar Thein

A Novel Technique for Composing Device Drivers for Sensors on Smart Devices 671
 Deok hwan Gim, Seng hun Min and Chan gun Lee

Various Artistic Effect Generation From Reference Image 679
 Hochang Lee, Sang-Hyun Seo, Seung-Taek Ryoo and Kyung-Hyun Yoon

A Photomosaic Image Generation on Smartphone 687
 Dongwann Kang, Sang-Hyun Seo, Seung-Taek Ryoo and Kyung-Hyun Yoon

Erratum to: IT Convergence and Services E1
 James J. Park, Hamid Arabnia, Hang-Bae Chang and Taeshik Shon

Author Index 695

Part I
IT Convergence and Services

Analysis of Security Vulnerability and Authentication Mechanism in Cooperative Wireless Networks

Ki Hong Kim

Abstract In this paper, we study the security vulnerabilities a CoopMAC faces and authentication mechanisms suitable for cooperative networks to be achieved. We identify various security attacks against control packets of CoopMAC and security vulnerabilities caused by these attacks, and discuss channel-based non-cryptographic mechanisms for user authentication in CoopMAC using physical layer characteristics.

Keywords CoopMAC · Cooperative communication · Security vulnerability · Physical layer security · Channel-assisted authentication

1 Introduction

Cooperative communication is indispensable for making ubiquitous communication connectivity a reality. Cooperative network is an innovative communication networks that takes advantages of the open broadcast nature of the wireless channel and the spatial diversity to improve channel capacity, robustness, reliability, and coverage. In the cooperative network, when the source node transmits data packet to the destination node, some nodes that are close to source node and destination node can serve as relay nodes by forwarding replicas of the source's data packet. The destination node receives multiple data packet from the source node and the relay nodes and then combines them to improve the communication quality [1, 2].

K. H. Kim (✉)

The Attached Institute of ETRI Yuseong, P. O. Box 1Daejeon 306-600,
The Republic of Korea
e-mail: hong0612@ensec.re.kr

A MAC protocol called CoopMAC is designed to improve the performance of the IEEE 802.11 MAC protocol [3] with minimal modification. It is able to increase the transmission throughput and reduce the average data delay. It also utilizes the multiple transmission rate capability of IEEE 802.11b, 1–11 Mbps, and allows the source node far away from the access point (AP) to transmit at a higher data rate by using a relay node [4, 5].

Although cooperative communication has recently gained momentum in the research community, there has been a great deal of concern about cooperative communication mechanism and its security issues. There have been several previous related works regarding communication techniques and security issues for cooperative network. The work in [1, 2] described wireless cooperative communication and presented several signaling schemes for cooperative communication. In [4, 5], a new MAC protocol for the 802.11 wireless local area network (WLAN), namely CoopMAC, was proposed and its performance was also analyzed. The potential security issues that may arise in a CoopMAC were studied in [6], and various security issues introduced by cooperating in Synergy MAC were also addressed in [7]. The [8] suggested cross-layer malicious relay tracing method to detect signal garbling and to counter attack of signal garbling by compromised relay nodes, while the [9] presented the distributed trust-assisted cooperative transmission mechanism handle relays' misbehavior as well as channel estimation errors. Also, a performance of cooperative communication in the presence of a semi-malicious relay which does not adhere to strategies of cooperation at all time was analyzed in [10], and a statistical detection scheme to mitigate malicious relay behavior in decode-and-forward (DF) cooperative environment was developed [11]. The examination of the physical consequences of a malicious user which exhibits cooperative behavior in a stochastic process was discussed in [12]. The [13] described a security framework for leveraging the security in cognitive radio cooperative networks. However, most of the works on cooperative communication is focused on efficient and reliable transmission schemes using the relay and identification of general security issues caused by the malicious relay node. No work has been done on the analysis of denial of service (DoS) vulnerability caused by an attacker node in cooperative networks.

In this paper, a case study of DoS attack in CoopMAC is presented for the first time. Security vulnerabilities at each protocol stage while attacking a cooperative communication is analyzed and compared. The authentication approaches, conventional mechanism using cryptographic algorithm and emerging mechanism using physical layer characteristics, are also discussed to verify entities in cooperative networks. This study differs from previous works in that it concentrates on one significant aspect of security vulnerability in the CoopMAC, namely DoS vulnerability of CoopMAC caused by the Dos attack of attacker node. This is believed to be the first comprehensive analysis and comparison of the security vulnerability from possible DoS attack and its authentication mechanisms in CoopMAC.

The remainder of this paper is organized as follows. In Sect. 2, we identify some possible security attacks against CoopMAC and then analyze the security

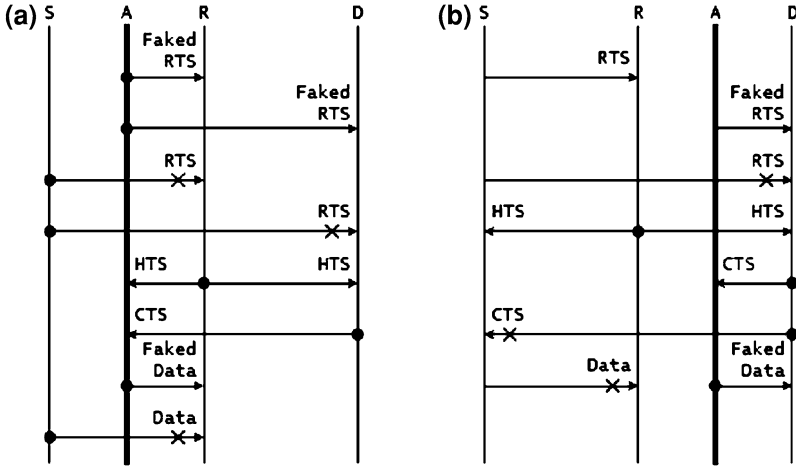


Fig. 1 Security vulnerability by RTS packet attack. a Faked RTS to R and D. b Faked RTS to D

vulnerabilities. Next, we discuss that it is possible to achieve a channel-based non-cryptographic authentication mechanism that uses physical layer properties to provide authentication service. Finally, in Sect. 4, we review our conclusion and detail plan for future work.

2 Security Vulnerability in CoopMAC

Due to broadcast nature of the wireless channel and cooperative nature, cooperative communication suffers from various attacks. The goal of the attacker node is to obstruct the communication between source and destination. These attackers would exploit the weakness in cooperation procedures, especially in the control packet exchange, and disguise themselves as legitimate relays. We will analyze and compare some cases of attacks according to the control packet of CoopMAC next.

2.1 Attacks on RTS Control Packet

In the CoopMAC as shown in Fig. 1a, attacker A sends the faked RTS to relay R and destination D, and then waits for the HTS from relay R as well as CTS from destination D. After the attacker A receives the HTS and the CTS, it sends a faked data to the relay R. Consequently, this attack results in a transmission disturbance in the RTS and the data packet from source S.

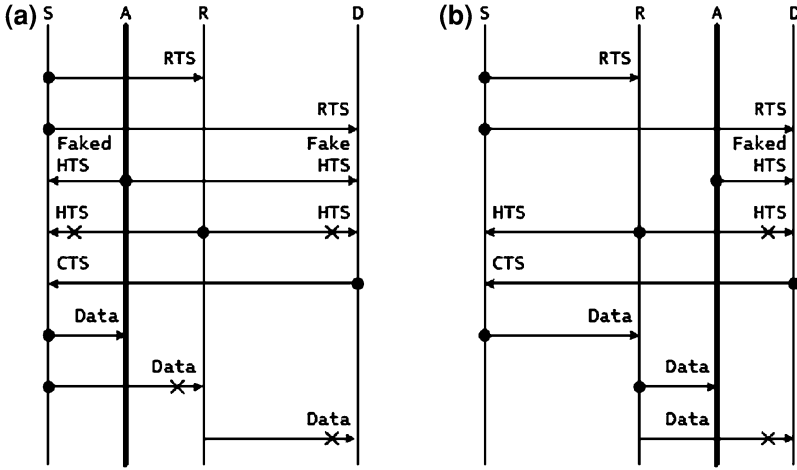


Fig. 2 Security vulnerability by HTS packet attack. **a** Faked HTS to S and D. **b** Faked HTS to D

On the other hand, as shown in Fig. 1b, attacker A intentionally sends the faked RTS to only destination D. The legal RTS from source S can be rejected by destination D due to an illegal previous RTS received from attacker A. Hence, CTS is sent from the destination D to attacker A, which causes source S to continuously wait for the CTS from destination D. As a result, normal cooperative communication between source S and destination D cannot be guaranteed.

2.2 Attacks on HTS Control Packet

As shown in Fig. 2a, the faked HTS is sent from attacker A to source S and destination D. Accordingly, the legal HTS from relay R is denied by source A and destination D. Then, destination D sends CTS to source A. After receiving the faked HTS and CTS, source S starts data transmission to attacker A, but relay R. Due to this false transmission to the attacker A, cooperative communication between source S and destination D via relay R is not established.

The potential vulnerability from faked HTS is also shown in Fig. 2b. In the case of sending faked HTS to only destination D, since the destination D is typically not come to know of this, although the legal HTS is sent from the relay R to destination D, it is denied by destination D. Then, the destination D sends a CTS to source S in order to notify that it successfully receives the control packet. This also means that attacker A is an intended legitimate relay forwarding data packet. Therefore, if relay R receives the data packet from source S, it does not forward data packet to the destination D, but forwards it the attacker A. Finally, the attacker A denies cooperative communication to the source S by simply dropping the data packet it receives. It also spoofs an ACK, causing the source S to wrongly conclude a successful transmission.

The potential vulnerability from faked HTS is also shown in Fig. 2b. In the case of sending faked HTS to only destination D , since the destination D is typically not come to know of this, although the legal HTS is sent from the relay R to destination D , it is denied by destination D . Then, the destination D sends a CTS to source S in order to notify that it successfully receives the control packet. This also means that attacker A is an intended legitimate relay forwarding data packet. Therefore, if relay R receives the data packet from source S , it does not forward data packet to the destination D , but forwards it the attacker A . Finally, the attacker A denies cooperative communication to the source S by simply dropping the data packet it receives. It also spoofs an ACK, causing the source S to wrongly conclude a successful transmission.

2.3 Attacks on CTS Control Packet

Figure 3 shows a security vulnerability which caused by the faked CTS from attacker A . The attacker A sends a faked CTS to the source S , informing the source S that it is an intended recipient of future data packet. And, since the authentication is not applied to CTS packet, the legal CTS from destination D can be rejected by source S due to previous illegal CTS from attacker A . After receiving the CTS from attacker A , source S transmits data packet to relay R . Subsequently, the relay R receives the data packet and then forwards received data packet to attacker A . The attacker A may try to deny communication service to the source S by deliberately not forwarding data packet received from the relay R . Consequently, cooperative communication is not established.

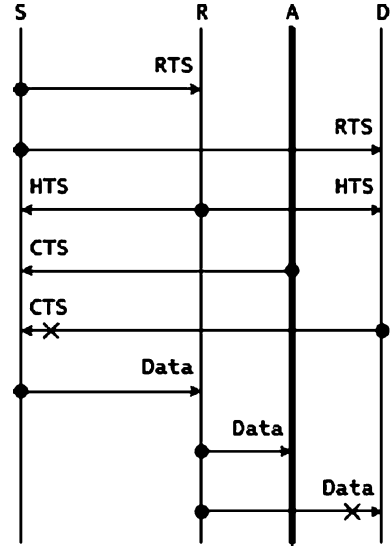
3 Cryptographic & Non-cryptographic Authentication

In order to prevent the security attacks inherent in cooperative networks and to verify communication entities more efficiently, we discuss authentication approaches and the practical implementation issues. Such authentication approach can be achieved by one of two approaches: (1) conventional approach using cryptographic algorithm, or (2) channel-assisted approach using physical layer properties of wireless channel [14–16].

3.1 Conventional Cryptographic Authentication

Authentication provides the assurance that users are who they claim to be or that data come from where they claim to originate. Most conventional cryptographic mechanisms of authentication are accomplished at a higher layer, namely above

Fig. 3 Security vulnerability by CTS packet attack



the physical layer. Although these conventional mechanisms can potentially provide authenticity in static networks, they are inefficient in dynamic networks including cooperative wireless networks.

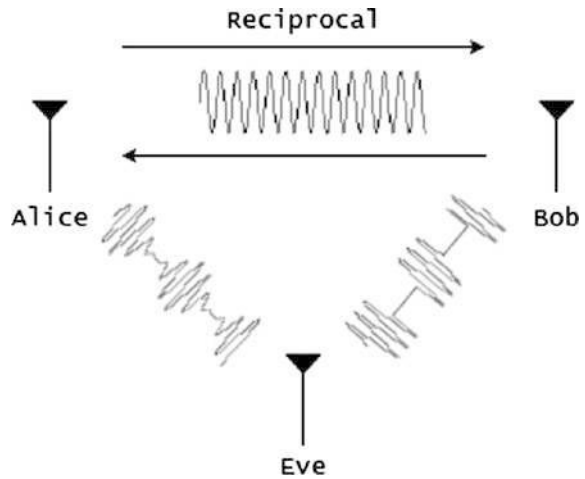
A few demerits can be identified as follows. First, most conventional cryptographic mechanisms are not suited for less equipped wireless networks due to large computational complexity. Second, the conventional cryptographic techniques need key management system which generates, distributes, and refreshes the keys. However, it is difficult in dynamic wireless networks where entities frequently join and leave the network. Third, wireless communication devices are subjects to physical compromises in adversarial communication environment. Therefore, these constraints of dynamic networks can cause the conventional cryptographic authentication not to work well in cooperative networks [17].

3.2 Channel-Assisted Non-cryptographic Authentication

Due to the main characteristics in cooperative networks or in its communication systems, namely dynamic network topology, variable channel capacity, limited bandwidth, limited processing capacity, and limited power, the authentication mechanism in cooperative networks should be lightweight and scalable.

In light of these constraints, there is increasing concern in enhancing or complementing conventional cryptographic authentication techniques in wireless networks using physical layer authentication mechanisms. The physical layer authentication mechanism is the channel-assisted non-cryptographic authentication scheme using the inherent and unique properties of wireless channel. The following

Fig. 4 Typical scenario of security community with Alice (legitimate), Bob (legitimate), and Eve (illegitimate)



four main characteristics of wireless channels can allow the wireless channel to be used as a means to authenticate the legitimate entity [14–17].

- The time-variant wireless channel impulse response $h(t, \tau)$ decorrelates quite rapidly in space. It implies that if the one of the entities changes its location in space by the order of a wavelength or more, the resulting channel response will be uncorrelated with the previous one.
- Wireless channel also changes in time. It results in a natural refresh for a channel-assisted security mechanism.
- The wireless channel is reciprocal in space, which means that the channel between two transceivers has the same frequency response in either communication direction at the same time instant.
- The time variation is slow enough so that the channel response can be accurately estimated within the channel coherence time. The channel state is considered to be stable, predictable, or highly correlated during the coherence time of the channel.

As depicted in Fig. 4, three different entities, Alice, Bob, and Eve, are potentially located in spatially separated positions. Alice and Bob are the two legitimate entities, and Eve is the illegitimate entity. Alice is the transmitter that initiates communication and sends data packet, while Bob is the intended receiver. Eve is an adversary that injects false signals into the channel in the hope of spoofing Alice. In this typical communication environment, our major security goal is to provide authentication service between Alice and Bob. The legitimate receiver Bob should have to distinguish between legitimate signals from transmitter Alice and illegitimate signals from illegitimate Eve because Eve locates within range of Alice and Bob so that it is capable of injecting undesirable signals into the wireless channel to impersonate Alice.

In the environment as shown in Fig. 4, suppose that Alice transmits data packet to Bob at a sufficient rate to ensure temporal coherence between successive data

packets and that Bob estimate the Alice-Bob channel prior to Eve's arrival. In addition, while trying to impersonate Alice, Eve wishes to convince Bob that she is Alice. To provide authentication between Alice and Bob, Bob first uses the received signal from Alice to estimate the channel response. He then compares this signal with a previous signal version of the Alice-Bob channel. If the two channel responses are close to each other, Bob conclude that the source of the data packet is the same as the source of the previously transmitted data. Otherwise, Bob concludes that the transmitter is not Alice [14–17]. Using this uniqueness of the Alice-Bob wireless channel, it is possible to distinguish between a legitimate transmitter and illegitimate one. It is caused by the fact that the wireless channel decorrelates in space, so the Alice-Bob channel is totally uncorrelated with the Alice-Eve and Bob-Eve channels if Eve is more than an order of a wavelength away from Alice and Bob.

4 Conclusion and Future Work

This paper presented the first comprehensive case study of DoS attack in the CoopMAC, which analyzed security vulnerabilities at each protocol stage while attacking a control packet exchanged among nodes. It also discussed that a channel-assisted authentication approach is applicable to enhance and supplement conventional cryptographic authentication mechanisms for cooperative networks. These channel-assisted non-cryptographic mechanisms exploit physical layer information of wireless media, such as the rapid spatial, spectral, and temporal decorrelation properties of the radio channel. In this way, legitimate entities can be reliably authenticated and illegitimate entities can be reliably detected. Our analytical results can be applied not only to cooperative network security, but also wireless sensor network (WSN) security design in general.

In the future, the authors will attempt to design and implement physical layer authentication mechanism suitable for cooperative networks. The plan is then to examine the effect that the proposed authentication mechanism has on the performance and efficiency of the cooperative transmission.

References

1. Nosratinia A, Hunter TE, Hedayat A (2004) Cooperative communication in wireless networks. *IEEE Commun Mag* 42(10):74–80
2. Laneman JN, Tse DNC, Wornell GW (2004) Cooperative diversity in wireless networks: efficient protocols and outage behavior. *IEEE Trans Inform Theory* 50(12):3062–3080
3. Part 11: (2003) Wireless LAN medium access control (MAC) and Physical layer (PHY) specifications, ANSI/IEEE Std 802.11, 1999 Edition (R2003)
4. Liu P, Tao Z, Panwar S (2005) A cooperative MAC protocol for wireless local area networks. *Proceedings of the 2005 IEEE ICC*, pp 2962–2968

5. Korakis T, Tao Z, Makda S, Gitelman B (2007) It is better to give than to receive— Implications of cooperation in a real environment. Springer LNCS 4479, pp 427–438
6. Makda S, Choudhary A, Raman N, Korakis T, Tao Z, Panwar S (2008) Security implications of cooperative communications in wireless networks. In: Proceedings of the 2008 IEEE sarnoff symposium, pp 1–6
7. Kulkarni S, Agrawal P (2010) Safeguarding cooperation in synergy MAC. In: Proceedings of the 2010 SSST, pp 156–160
8. Mao Y, Wu M (2007) Tracing malicious relays in cooperative wireless communications. IEEE Trans Inform Forensics Secur 2(2):198–207
9. Han Z, Sun YL (2007) Securing cooperative transmission in wireless communications. In: Proceedings of the 2007 IEEE MobiQuitous, pp 1–6
10. Dehnie S, Sencar HT, Memon N (2007) Cooperative diversity in the presence of a misbehaving relay: performance analysis. In: Proceedings of the IEEE Sarnoff Symposium, pp 1–7
11. Dehnie S, Sencar HT, Memon N (2007) Detecting malicious behavior in cooperative diversity. In: Proceedings of the 2007 IEEE CISS, pp 895–899
12. Dehnie S, Memon N (2008) A stochastic model for misbehaving relays in cooperative diversity. In: Proceedings of the 2008 IEEE WCNS, pp 482–487
13. Marques H, Ribeiro J, Marques P, Zuquete A, Rodriguez J (2009) A security framework for cognitive radio IP based cooperative protocols. In: Proceedings of the 2009 IEEE PIMRC, pp 2838–2842
14. Zeng K, Govindan K, Mohapatra P (2010) Non-cryptographic authentication and identification in wireless networks. IEEE Wirel Commun 17(5):56–62
15. Xiao L, Greenstein L, Mandayam N, Trappe W (2008) Using the physical layer for wireless authentication in time-variant channels. IEEE Wirel Commun 7(7):2571–2579
16. Yu PL, Baras JS, Sadler BM (2008) Physical-layer authentication. IEEE Trans Inform Forensics Secur 3(1):38–50
17. Mathur S (2010) Exploiting the physical layer for enhanced security. IEEE Wirel Commun 17(5):63–70

Spam Host Detection Using Ant Colony Optimization

Arnon Rungsawang, Apichat Taweewirawate
and Bundit Manaskasemsak

Abstract Inappropriate effort of web manipulation or spamming in order to boost up a web page into the first rank of a search result is an important problem, and affects the efficiency of a search engine. This article presents a spam host detection approach. We exploit both content and link features extracting from hosts to train a learning model based on ant colony optimization algorithm. Experiments on the WEBSpAM-UK2006 dataset show that the proposed method provides higher precision in detecting spam than the baseline C.45 and SVM.

Keywords Spam host detection · Ant colony optimization algorithm · Content and link features · Search engine

1 Introduction

Search Engine has been developed and used as a tool to locate web information and resources. For a given query, the ranking result on the first page of a famous search engine is highly valuable to commercial web sites. Current competitive business then gives birth to aggressive attempts from web engineers to boost the

A. Rungsawang (✉) · A. Taweewirawate · B. Manaskasemsak
Massive Information and Knowledge Engineering Department of Computer Engineering,
Kasetsart University, Bangkok 10900, Thailand
e-mail: arnon@mikelab.net

A. Taweewirawate
e-mail: ball@mikelab.net

B. Manaskasemsak
e-mail: un@mikelab.net

ranking of web pages in search results to increase the return of investment (ROI). Manipulating search engine ranking methods to obtain a higher than deserved rank of a web page is called search engine (or web) spam [1]. Besides degrading the quality of search results, the large number of pages explicitly created for spamming also increases the cost of crawling, and inflates both index and storage with many useless pages.

As described by Gyöngyi and Garcia-Molina in [1], there are many varieties of spamming techniques. Often, most of them exploit the weakness of the search engine's ranking algorithm, such as inserting a large number of words that are unrelated to the main content of the page (i.e., content spam), or creating a link farm to spoil the link-based ranking results (i.e., link spam). Many researchers have concentrated on combating spam. For example, Gyöngyi et al. [2] propose an idea to propagate trust from good sites to demote spam, while Wu and Davison [3] expand from a seed set of spam pages to the neighbors to find more suspicious pages in the web graph. Dai et al. [4] exploit the historical content information of web pages to improve spam classification, while Chung et al. [5] propose to use time series to study the link farm evolution. Martinez-Romo and Araujo [6] apply a language model approach to improve web spam identification.

In this paper, we propose to apply the ant colony optimization algorithm [7, 8] in detecting spam host problem. Both content and link based features extracted from normal and spam hosts have been used to train the classification model in order to discover a list of classification rules. From the experiments with the WEBSHAM-UK2006 [9], the results show that rules generated from ant colony optimization learning model can classify spam hosts more precise than the baseline decision tree (C4.5 algorithm) and support vector machine (SVM) models, that have been explored by many researchers [10–12].

2 Related Work and Basic Concept

2.1 *Web Spam Detection Using Machine Learning Techniques*

Web spam detection became a known topic to academic discourse since the Davison's paper on using machine learning techniques to identify link spam [13], and was further reasserted by Henzinger et al. [14] as one of the most challenges to commercial search engines. Web spam detection can be seen as a binary classification problem; a page or host will be predicted as spam or not spam.

Fetterly et al. [15] observe the distribution of statistical properties of web pages and found that they can be used to identify spam. In addition to content properties of the web pages or hosts, link data is also very helpful. Becchetti et al. [10] exploit the link features, e.g., the number of in- and out-degree, PageRank [16], and TrustRank [2], to build a spam classifier. Following the work in [10, 12], Castillo et al. [11]

extract link features from the web graph and host graph, and content features from individual pages, and use the simple decision tree C4.5 to build the classifier. Recently, Dai et al. [4] extract temporal features from the Internet Archive's Wayback Machine [17] and use them to train a cascade classifier built from several SVM^{light} and a logistic regression implemented in WEKA [18].

2.2 Basic Concept of Ant Colony Optimization

Naturally, distinct kind of creatures behaves differently in their everyday life. In a colony of social ants, each ant usually has its own duty and performs its own tasks independently from other members of the colony. However, tasks done by different ants are usually related to each other in such a way that the colony, as a whole, is capable of solving complex problems through cooperation [8, 19]. For example, for survival-related problems such as selecting the shortest walking path, finding and storing food, which require sophisticated planning, are solved by ant colony without any kind of supervisor. The extensive study from ethologists reveals that ants communicate with one another by means of pheromone trails to exchange information about which path should be followed. As ants move, a certain amount of pheromone is dropped to make the path with the trail of this substance. Ants tend to converge to the shortest trail (or path), since they can make more trips, and hence deliver more food to their colony. The more ants follow a given trail, the more attractive this trail becomes to be followed by other ants. This process can be described as a positive feedback loop, in which the probability that an ant chooses a path is proportional to the number of ants that has already passed through that path [7, 8].

Researchers try to simulate the natural behavior of ants, including mechanisms of cooperation, and devise ant colony optimization (ACO) algorithms based on such an idea to solve the real world complex problems, such as the travelling salesman problem [20], data mining [19]. ACO algorithms solve a problem based on the following concept:

- Each path followed by an ant is associated with a candidate solution for a given problem.
- When an ant follows a path, it drops varying amount of pheromone on that path in proportion with the quality of the corresponding candidate solution for the target problem.
- Path with a larger amount of pheromone will have a greater probability to be chosen to follow by other ants.

In solving an optimization problem with ACO, we have to choose three following functions appropriately to help the algorithm to get faster and better solution. The first one is a problem-dependent heuristic function (η) which measures the quality of items (i.e., attribute-value pairs) that can be added to the current partial solution (i.e., rule). The second one is a rule for pheromone updating (τ) which specifies how to modify the pheromone trail. The last one is a

probabilistic transition rule (P) based on the value of the heuristic function and on the contents of the pheromone trail that is used to iteratively construct the solution.

3 Spam Detection Based on Ant Colony Optimization Algorithm

3.1 Graph Representation

In a learning process based on the ACO algorithm, problems are often modeled as a graph. Thus, we let $\{A_1, A_2, \dots, A_m\}$ represent a set of m features, i.e., both content and link features, extracted from hosts. If we denote $\{a_{i1}, a_{i2}, \dots, a_{ini}\}$ to a set of n_i possible values belonged to a feature A_i . Therefore, we can construct a graph $G = (V, E)$ including a set of nodes $V = \{A_1, A_2, \dots, A_m\} \cup \{S\}$ and a set of edges $E = V^2$, where S is a virtual node set to a starting point. This graph can be illustrated in Fig. 1.

3.2 Methodology

Consider the graph in Fig. 1, when we assign artificial ants to start walking from node S , behavior of those ants will decide to choose a path to walk in each step, from one node to others, using some probabilistic transition function calculated based on the value of a heuristic function and pheromone information value. The following probabilistic transition P_{ij} is denoted a probability value for an ant to walk from any current node to node a_{ij} :

$$P_{ij} = \frac{\eta_{ij}\tau_{ij}(t)}{\sum_{i=1}^m x_i \cdot \sum_{j=1}^{n_i} (\eta_{ij}\tau_{ij}(t))}, \quad (1)$$

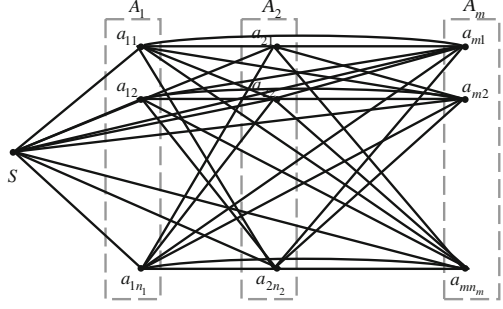
where η_{ij} denotes a heuristic function, $\tau_{ij}(t)$ denotes a pheromone information value obtained at iteration time t , and

$$x_i = \begin{cases} 1 & \text{If the node } a_{i*} \text{ has never been passed by that ant,} \\ 0 & \text{Otherwise.} \end{cases} \quad (2)$$

In this paper, we use an open source software package called GUIAnt-Miner [21] which provides an implementation of the ACO algorithm [19] used for classification problems in data mining. The default heuristic function with the value of disorder (i.e., the entropy function) between nodes is defined by:

$$\eta_{ij} = \frac{\log_2 k - H(W|A_i = a_{ij})}{\sum_{i=1}^m x_i \cdot \sum_{j=1}^{n_i} (\log_2 k - H(W|A_i = a_{ij}))}, \quad (3)$$

Fig. 1 The problem represented as a graph



where

$$H(W|A_i = a_{ij}) = - \sum_{w=1}^k (P(w|A_i = a_{ij}) \cdot \log_2 P(w|A_i = a_{ij})). \quad (4)$$

Here, we define W as a set of target classes and k as the number of classes (i.e., $|W|$), so that $W = \{spam, normal\}$ and $k = 2$ in this case. $P(w|A_i = a_{ij})$ is the probability of class w given $A_i = a_{ij}$. Consequently, the range of a value obtained from Eq. 4 is $(0, \log_2 k)$.

Since the ACO algorithm iteratively finds the optimal solution, the pheromone in Eq. 1 which controls the movement of ants will be changed for each run. For GUIAnt-Miner, the pheromone information function has been defined as:

$$\tau_{ij}(t+1) = \frac{\tau_{ij}(t) + \tau_{ij}(t)Q}{\sum_{i=1}^m \sum_{j=1}^{n_i} \tau_{ij}(t)}, \quad (5)$$

where Q measures the quality of prediction rules over the training data set. This measure is defined as the product of the sensitivity and specificity:

$$Q = \frac{TP'}{TP' + FN'} \cdot \frac{TN'}{FP' + TN'}. \quad (6)$$

Note that TP' is the number of hosts covered by rule that has the class predicted by that rule, FP' is the number of hosts covered by rule that has a class different from the class predicted by that rule, FN' is the number of hosts that is not covered by rule but has the class predicted by that rule, and TN' is the number of hosts that is not covered by rule and that does not have the class predicted by that rule.

For the first iteration, the initial pheromone value is normally set to:

$$\tau_{ij}(t=0) = \frac{1}{\sum_{i=1}^m n_i} \quad (7)$$

After each iteration run, a result of the model can be expressed by a path of ant walking from a value of a feature through one of the other feature. However, this result does not specify to any target class yet and then cannot be utilized. We therefore check all hosts covered by the result from the training data set again to obtain a target class by majority vote, and subsequently create a rule as follows.

IF ($A_i = a_{ix}$ AND $j = a_{jy}$ AND \dots) THEN ($W = w_z$)

Consequently, the iterative computation in Eq. 1 will terminate if it produces the set of rules covering all hosts in the training data set.

4 Experimental Results

4.1 Data Set Preparation

We use the WEBSpAM-UK2006 [9] containing hosts within .uk domain. From these, there are 1,803 hosts labeled as spam and 4,409 hosts labeled as normal. The data set contains several features including both content- and link-based features, as well as a spamicity value of each host. We further process this data set as follows (see Fig. 2):

- For the 1,803 spam hosts, we first sort them by ascending order of the spamicity values. Each host will be assigned with an identification number beginning from 0. We then decompose spam hosts into to 3 buckets by considering the remainder from dividing its identification number with 3. Eventually, we will have “bucket1”, “bucket2”, and “bucket3”, in which each contains equally 601 spam hosts.
- Similarly, for the 4,409 normal hosts, we sort them by descending order of the spamicity values. We equally divide them into 10 portions, and assign an identification number beginning from 0 to each host in each portion separately. For each portion, each identification number is again modulo by 7. The normal hosts whose remainder is 0, 2 and 5, will then be assigned into “bucket1”, “bucket2”, and “bucket3”, respectively. Note that the host with less identification number will be first assigned. To avoid data imbalance of normal and spam hosts in training set, we will stop the assigning process if each bucket contains 601 normal hosts. For all remaining hosts, we will put them into a new “bucket4”.

4.2 Host’s Feature Selection

We use the information gain as a criterion to select the host’s features. Figure 3 shows the 10 highest information gain features used to train the machine learning models. Of these, the first nine features are the link-based features; but only the last one is the content-based feature. Since all these features have continuous-range values, which cannot directly exploit in the GUIAnt-Miner program; we therefore discretize those values into 10 equal ranges.

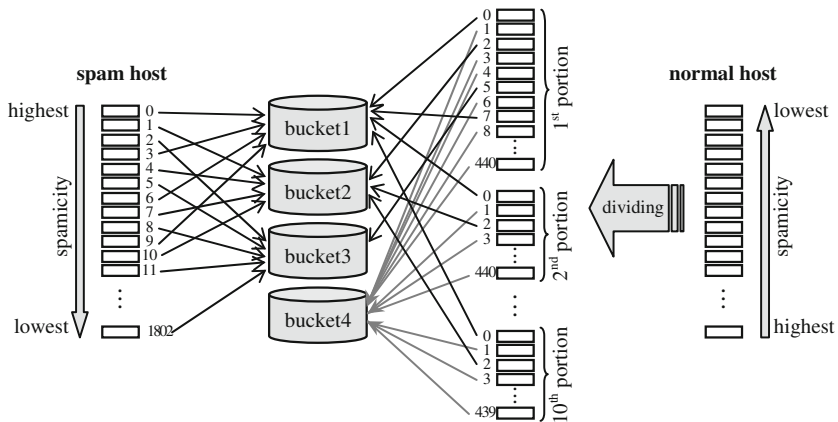


Fig. 2 Data preparation

-
1. Logarithm value of TrustRank/PageRank of homepage
 2. Logarithm value of TrustRank/in-degree of homepage
 3. Logarithm value of TrustRank/PageRank of max PageRank
 4. Logarithm value of TrustRank of homepage
 5. Logarithm value of TrustRank/in-degree of max PageRank
 6. Logarithm value of TrustRank of max PageRank
 7. Logarithm value of number of different supporters (sites) at distance 4 from homepage
 8. Logarithm value of number of different supporters (sites) at distance 4 from max PageRank
 9. Logarithm value of number of different supporters (sites) at distance 3 from homepage
 10. Top 200 corpus recall (standard deviation for a ll pages in the host)
-

Fig. 3 Features used to train the machine learning models

4.3 Results

From the set of data described in Sect. 4.1, we design 3 set of experiments according to the following scenarios:

- Scenario 1: we use bucket1 for training, while use bucket2, bucket3, and bucket4 for testing.
- Scenario 2: we use bucket2 for training, while use bucket1, bucket3, and bucket4 for testing.
- Scenario 3: we use bucket3 for training, while use bucket1, bucket2, and bucket4 for testing.

We compare performance of the ACO model with two other baselines, i.e., the decision tree (C4.5) and the support vector machine (SVM), using two standard measures: the positive predictive value (i.e., precision) and false positive rate (i.e., fall-out). To train the ACO model, we use 5 artificial ants. The terminating condition is either uncovered hosts by the rules are less than 10, or the number of

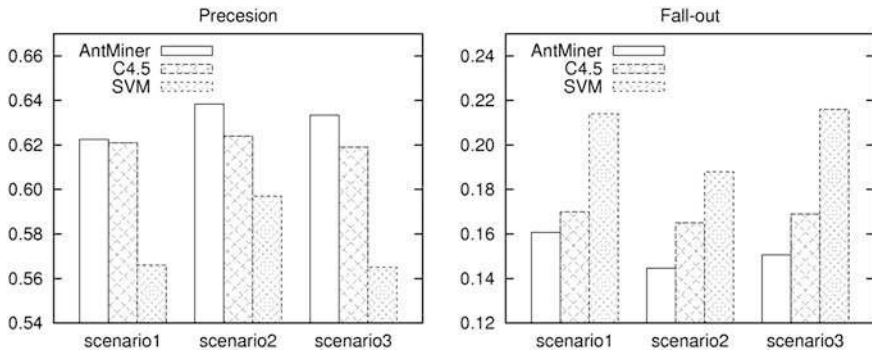


Fig. 4 Spam host detection performance

iterations reaches 100. Generated rules that can cover at least 5 hosts will be kept as a candidate set of usable rules. These rules will finally be checked with the training data set again to obtain a target class by majority vote. For the C4.5 model, the rule pruning is disabled. For all other remaining parameters of C4.5 and SVM, the default setting in WEKA software [18] has been assigned.

The precision results in Fig. 4 show that the ACO learning model has the ability to detect spam hosts more accurate than C4.5 and SVM in all experiments. This is consistent with the fall-out results that the ACO learning model yields the least error prediction.

5 Conclusions

In this article, we propose to apply the ant colony optimization based algorithm to build a set of classification rules for spam host detection. Both content and link features extracted from normal and spam hosts have been exploited. From the experiments with the WEBSpAM-UK2006 dataset, the proposed method provides higher precision in detecting spam than the basic decision tree C4.5 and SVM models. However, we currently just run our experiments using the default heuristic and basic pheromone updating function setting in the GUIAnt-Miner. In future work, we are looking forward to doing further experiments using other types of heuristic and pheromone updating functions, and hope to obtain higher quality set of classification rules.

References

1. Gyöngyi Z, Garcia-Molina H (2005) Web spam taxonomy. In: Proceedings of the 1st international workshop on adversarial information retrieval on the web
2. Gyöngyi Z, Garcia-Molina H, Pedersen J (2004) Combating web spam with TrustRank. In: Proceedings of the 30th international conference on very large data bases

3. Wu B, Davison BD (2005) Identifying link farm spam pages. In: Proceedings of the 14th international world wide web conference
4. Dai N, Davison BD, Qi X (2009) Looking into the past to better classify web spam. In: Proceedings of the 5th international workshop on adversarial information retrieval on the web
5. Chung Y, Toyoda M, Kitsuregawa M (2009) A study of link farm distribution and evolution using a time series of web snapshots. In: Proceedings of the 5th international workshop on adversarial information retrieval on the web
6. Martínez-Romo J, Araujo L (2009) Web spam identification through language model analysis. In: Proceedings of the 5th international workshop on adversarial information retrieval on the web
7. Dorigo M, Di Caro G, Gambardella LM (1999) Ant algorithms for discrete optimization. *Artif Life* 5(2):137–172
8. Dorigo M, Maniezzo V, Coloni A (1996) Ant system: optimization by a colony of cooperating agents. *IEEE Trans Syst Man Cybern* 26(1):29–41
9. Castillo C, Donato D, Becchetti L, Boldi P, Leonardi S, Santini M, Vigna S (2006) A reference collection for web spam. *ACM SIGIR Forum* 40(2):11–24
10. Becchetti L, Castillo C, Donato D, Leonardi S, Baeza-Yates R (2006) Link-based characterization and detection of web spam. In: Proceedings of the 2nd international workshop on adversarial information retrieval on the web
11. Castillo C, Donato D, Gionis A, Murdock V, Silvestri F (2007) Know your neighbors: web spam detection using the web topology. In: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval
12. Ntoulas A, Najork M, Manasse M, Fetterly D (2006) Detecting spam web pages through content analysis. In: Proceedings of the 15th international world wide web conference
13. Davison BD (2000) Recognizing nepotistic links on the web. In: Proceedings of AAAI workshop on artificial intelligence for web search
14. Henzinger MR, Motwani R, Silverstein C (2002) Challenges in web search engines. *ACM SIGIR Forum* 36(2):11–22
15. Fetterly D, Manasse M, Najork M (2004) Spam, dam spam, and statistics: using statistical analysis to locate spam web pages. In: Proceedings of the 7th international workshop on the web and databases
16. Page L, Brin S, Motwani R, Winograd T (1999) The PageRank citation ranking: bringing order to the web. Technical report, Stanford InfoLab
17. Internet archive. The wayback machine. <http://www.archive.org/>
18. Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques with Java implementations, 2nd edn. Morgan Kaufmann, San Francisco
19. Parpinelli RS, Lopes HS, Freitas AA (2002) Data mining with an ant colony optimization algorithm. *IEEE Trans Evol Comput* 6(4):321–332
20. Dorigo M, Gambardella LM (1997) Ant colony system: a cooperative learning approach to the traveling salesman problem. *IEEE Trans Evol Comput* 1(1):53–66
21. Dorigo, M (2004) Ant colony optimization public software. <http://iridia.ulb.ac.be/~mdorigo/ACO/aco-code/public-software.html/>

Location Estimation of Satellite Radio Interferer Using Cross Ambiguity Function Map for Protection of Satellite Resources

Chul-Gyu Kang, Chul-Sun Park and Chang-Heon Oh

Abstract In this paper, a scheme using Cross Ambiguity Function (CAF) map is proposed to estimate the location of an unknown interferer which emits harmful radio signal in the satellite communication network. In conventional CAF based TDOA, FDOA location, TDOA and FDOA are determined by location the peak in the CAF plane and then the peak's information is fed into a least squares like location tool to determine the emitter's location. However, this proposed scheme omits the step in which the location is determined with the post processed CAF peak information and instead maps the CAF surface directly to the earth surface. In simulation results, the distance error of about 800 m is occurred at $E_b/N_0 = 4-10$ dB and the distance error of about 1.3 km is occurred at -20 dB of E_b/N_0 .

Keywords Cross ambiguity function · TDOA · FDOA · Location estimation · Satellite interference

C.-G. Kang (✉) · C.-H. Oh
School of Electrical, Electronics and Communication Engineering,
Korea University of Technology and Education, Cheon-An, Korea
e-mail: swing98@kut.ac.kr

C.-H. Oh
e-mail: choh@kut.ac.kr

C.-S. Park
Network Planning and Protection Division,
Korea Communications Commission, Seoul, Korea
e-mail: poempark@kcc.go.kr

1 Introduction

In all around world, USA('77), Germany('80), Japan('98), and China are operating the satellite radio monitoring system to protect their satellite resources from other country's satellites and paper satellites. In case of our country, to secure satellite resources and protect the right at the state level, the satellite radio monitoring center was constructed in 2002 and it has been contributed to the policy establishment of the satellite and the development of the satellite industry as eliminating harmful interferer, supplying all sort of measurement data, and monitoring the domestic and foreign satellite what they fulfill international telecommunication union (ITU) international regulation. The accurate location of the illegal interferer has to be estimated first to perform those roles.

There are time difference of arrival (TDOA), frequency difference of arrival (FDOA) and cross ambiguity function (CAF) schemes, which is using TDOA and FDOA both, to estimate the interferer location in satellite radio interferer searching system so far. However, the estimation scheme using TDOA is affected by the shape of receiver position for the estimation performance, and FDOA can not estimate when there is no movement of receivers or interferer as the frequency offset is not happened [1, 2]. The cross ambiguity function is somewhat free from these problems having TDOA and FDOA scheme as it is using TDOA and FDOA both to estimate the interferer location. In CAF, TDOA and FDOA are determined by the peak location in the CAF plane and then the peak's information is fed into a least squares like location tool to determine the emitter's location. Therefore, the computational complexity becomes a problem [3].

To solve these kinds of the problems and get the high performance in the interferer location estimation, the scheme using CAF map is proposed. The proposed scheme in this paper omits the step in which the location is determined with the post processed CAF peak information and instead maps the CAF surface directly to the earth surface.

2 Searching Technique of Satellite Radio Interferer

2.1 Interference Scenario

As shown in Fig. 1, the signal transmitted at the same interferer notated as interference earth station is received at the contiguous two satellite SAT1 and SAT2.

We assume that $s_1(t)$ is a received signal with high signal to noise ratio when an arbitrary signal source $s(t)$ is transmitted to the earth station of satellite SAT1. At the same time, $s_2(t)$ is also transmitted from the side lobe of the interference earth station and it is received at the satellite SAT2 with low signal to noise ratio. We assume this signal is $s_2(t)$. Even though signal $s_1(t)$ and $s_2(t)$ are transmitted at

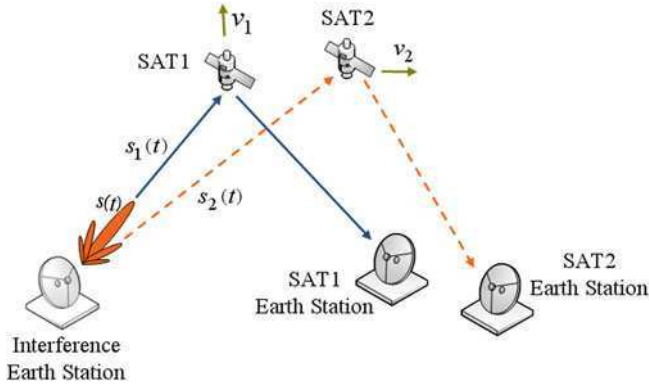


Fig. 1 The location estimation scenario of the interferer using CAF

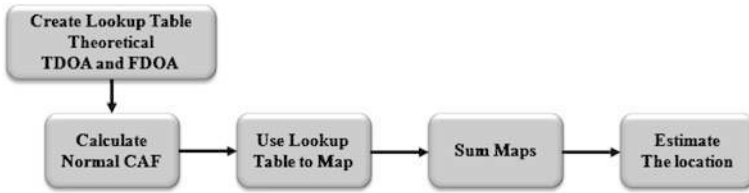


Fig. 2 The algorithm of the cross ambiguity function map

the same source $s(t)$, they have time differences of arrival caused by difference transmission paths each other. Furthermore, they have difference doppler frequencies of arrival because SAT1 and SAT2 move along their own orbit and own speed.

2.2 Cross Ambiguity Function Map

Figure 2 shows the steps of the cross ambiguity function map algorithm. The algorithm used in this approach follows:

1. Calculate the theoretical TDOA and FDOA value for points on the X, Y coordinates for the current geographic area to create a lookup table of FDOA and TDOA.
2. Calculate the normal cross ambiguity function.
3. Use the lookup table in step 1 to map the amplitude of the CAF in step 2 to a new X, Y coordinates.
4. Repeat 1–3 and sum maps

2.3 Cross Ambiguity Function

In Fig. 1, two transmitted signal $s_1(t)$ and $s_2(t)$ can be written like Eqs. (1) and (2).

$$s_1(t) = s(t) + n_1(t) \quad (1)$$

$$s_2(t) = s(t - f)e^{jf(t-\tau)} + n_2(t) \quad (2)$$

$n_1(t)$ and $n_2(t)$ are additive white gaussian noise (AWGN). $s_1(t)$ is received signal with added AWGN at $s(t)$ and $s_2(t)$ is received signal that is added time delay, τ , frequency difference, f , and AWGN to $s(t)$. The time delay and frequency difference are only decided according to the location of the signal source transmitting the signal. If TDOA and FDOA value are calculated comparing the received two signal $s_1(t)$ and $s_2(t)$, the location of the signal source is decided automatically. As previously stated, $s_2(t)$ includes the TDOA value and the FDOA value to $s_1(t)$ which are able to decide the location of the signal source. As long as TDOA value and FDOA value are compensated to $s_2(t)$, it could be the same signal as $s_1(t)$ except for AWGN. But the real TDOA and FDOA value are not calculated as making a simple comparison between the received two signals so cross ambiguity function are used.

$$CAF(\tau, f) = \int_0^T s_1(t)s_2^*(t + \tau)e^{-j2\pi ft} dt \quad (3)$$

In Eq. (3), T is signal time period and $*$ is conjugation. To modify continuous time signal like Eq. (3) to discrete time signal, time $t = nT_s$ and $f = kf_s/N$, where T_s is the sample period, $f_s = 1/T_s$ is the sampling frequency, n represents the individual sample numbers, and N is the total number of samples. Once these are inserted back into Eq. (3), we get Eq. (4):

$$CAF(\tau, k) = \sum_{n=0}^{N-1} [s_1(n)s_2^*(n - \tau)]e^{-j2\pi \frac{kn}{N}} \quad (4)$$

where s_1 and s_2 are the sampled signal in analytic format, τ is the time delay in samples, and k/N is the frequency difference in digital frequency, or fraction of the sample frequency. Note the similarity with the discrete fourier transform (DFT) in Eq.(5).

$$X(k) = \sum_{n=0}^{N-1} [x(n)]e^{-j2\pi \frac{kn}{N}} \quad (5)$$

Now replace $x(n)$ with $s_1(n)s_2^*(n - \tau)$ and we get the discrete form of the CAF Eq. (4). Cross ambiguity function, $|CAF(\tau, k)|$, has the peak value like showing Fig. 3 when the TDOA and FDOA value of s_1 and s_2 are the same. In Fig. 3, X axis, Y axis, and Z axis are TDOA value, FDOA value, and CAF value.

Fig. 3 The estimated TDOA and FDOA with cross ambiguity function

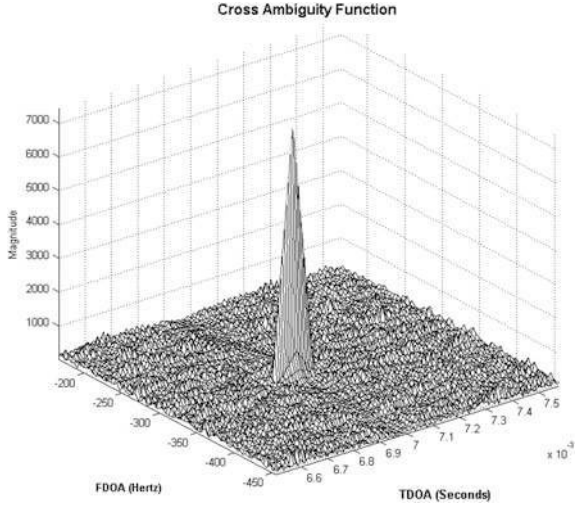


Table 1 The simulation parameters for the interferer searching system

Parameters	Values
Carrier and sampling frequency	11.8 GHz, 160 MHz
Symbol rate	13,333 ksymbol/sec
Signal to noise ratio	$s_1 = 10$ dB, $s_2 = -30$ -10 dB
Satellite 1 & 2 geodetic coordinates	Koreasat3(E:116°, N: 0°), Koreasat5(E:113°, N: 0°)
Satellite 1 & 2 velocity(m/s)	$x = 150, y = 0, z = 0$
Interferer geodetic coordinates	E:139°39' 16", N: 35° 12' 12"
Interferer velocity (m/s)	$x = 0, y = 0, z = 0$

3 Simulation and Results

For the computer simulation, we assume the parameters as Table 1. We assume that the interferer transmits signal $s_1(t)$ to Koreasat-3, signal $s_2(t)$ is transmitted from a side lobe of the same antenna to Koreasat-5 at the same time.

Figure 4 shows the TDOA and FDOA error according to E_b/N_0 changing -20 to 10 dB. In these results, the time error and frequency error is dramatically increased from 4 dB of E_b/N_0 because the correlation value between two signals transmitted from the interferer is decreased.

The left side of Fig. 5 shows the estimated interferer location using CAF map at 10 dB of E_b/N_0 . In theoretical calculation, the TDOA value and FDOA value are 6.9722 ms and -311.872 Hz in Table 1 parameters but the estimated values using CAF map are 6.9699 ms and -312.5305 Hz. The right side of Fig. 5 shows the estimated distance error of the interferer according to changing E_b/N_0 .

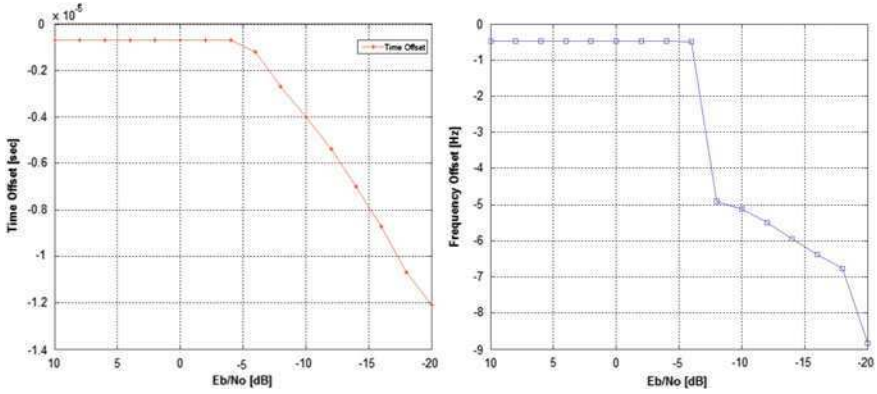


Fig. 4 The time error (left) and frequency error (right) according to E_b/N_0

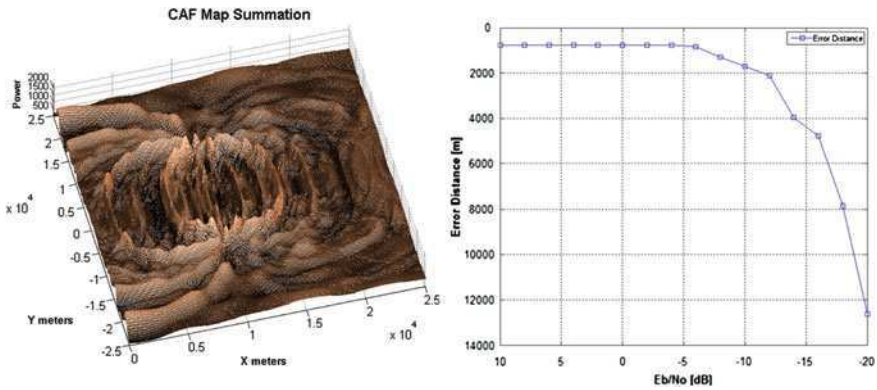


Fig. 5 The error distance (right) according to E_b/N_0 and the estimated interferer location with CAF Map (left) at 10 dB of E_b/N_0

The distance error is about 800 m at 10 -4 dB of E_b/N_0 and about 13 km of distance error is occurred at $E_b/N_0 = -20$ dB. It is also caused by the correlation value of two received signals.

4 Conclusion

In this paper, a scheme using CAF map is proposed to estimate the location of an unknown interferer which emits harmful radio signal in the satellite communication network. In this proposed scheme omits the step in which the location is

determined with the post processed CAF peak information and instead maps the CAF surface directly to the earth surface.

In simulation results, the distance error of about 800 m is occurred at 4–10 dB of E_b/N_0 and the distance error of about 1.3 km is occurred at $E_b/N_0 = -20$ dB. The reason showing the large distance error at low E_b/N_0 is that the correlation value between received signals is low. From these results, we confirm that this proposed scheme is very useful for the satellite radio monitoring system with the low computational complexity and high accuracy.

“This paper was partially supported by the education and research promotion program of KUT”

References

1. Dulman S, Havinga P, Baggio A, Langendoen K (2008) Revisiting the cramer-rao bound for localization algorithms. In: 4th IEEE/ACM DCOSS Work-in-progress paper, June
2. Vesely J (2010) Differential doppler target position fix computing methods. In: IEEE proceedings of the international conference on circuits, systems, signals, pp 284–287, Dec 2010
3. Stein S (2003) Algorithms for ambiguity function processing. *IEEE Trans Acoust Speech Signal Process* 29(3):588–599 Jan 2003
4. Wax M (1982) The joint estimation of differential delay, doppler, and phase. *IEEE Trans Inf Theory* IT-28:817–820 Sept 1982
5. Friedlander B (1984) On the Cramer-Rao bound for time delay and doppler estimation. *IEEE Trans Inf Theory* IT-30:575–580 May 1984

Korean Voice Recognition System Development

Soon Suck Jarng

Abstract In this paper, the voice recognition algorithm based on Hidden Markov Modeling (HMM) is analyzed in detail. The HMM voice recognition algorithm is explained and the importance of voice information DB is revealed for better improvement of voice recognition rate. An algorithm designed to extract syllable parts from continuous voice signal is introduced. This paper shows the relationship between recognition rates and number of applying syllables and number of groups for applying syllables.

Keywords Hidden Markov Model · Voice Recognition Algorithm

1 Introduction

Voice recognition was attempted in the 1960s based on Motor theory presented by Liberman and others [1]. The theory was as simple as that a voice was generated through the trachea and the speech was decoded in the brain. Even the voice spectrogram was not considered. In the 1970s Cole and Scott presented a progressive Multiple-Cue model where they suggested that a voice might be classified as an independent or dependent cue from a sentence, and that a phonemic shift would happen in the sentence [2]. Fletcher, who was studying about human auditory sensation on telephone speech, found that the non-sensation rate of a certain frequency band was the same as the non-sensation rate of a narrow band

S. S. Jarng (✉)

Department of Control and Instrumentation, Robotics Engineering,

Chosun University, 375 Seoseok-Dong, Dong-Ku, Gwang-Ju, South Korea

e-mail: ssjarng@chosun.ac.kr

multiplied by the number of non-sensible narrow bands [3]. Allen suggested a Fletcher–Allen algorithm in which voice recognition should be independently done in the frequency domain. However even though voice recognition is done partly in frequency domain, a still unknown brain-like functioning algorithm should be discovered to explain how the voice is divided into syllables and phonemes for recognition. Since there are too many unknown facts about how the brain recognizes the voice through different paths and processes, it may be still better to approach the problem by probabilistic algorithm than analytic algorithm. For this reason, two different voice recognition algorithms have been studied while the common feature in both these algorithms is to extract the feature parameters of the speech signal. The Neural Network (NN) recognition algorithm first generates a large-sized coefficient matrix through training of characteristic feature parameters representing syllables or words, then calculates an output index by directly applying the feature parameters of an unknown new syllable or word to the huge coefficient matrix [4, 5]. Recognition using a neural network speech recognition method with a large coefficient matrix for the whole learning process is time-consuming. If you add a new speech signal to the recognition algorithm, the entire process should be repeated from the beginning which is time consuming [6]. In the second method, Hidden Markov Model (HMM) recognition algorithm, for every new input voice signal, voice feature parameters are generated which are used in the learning process to create a new HMM model. So with each new HMM model created for every word, during the testing phase, all these models are compared with the test word to find out the matching voice sample [7, 8].

The disadvantage that a HMM model has is, that for every new voice that is added to the model, a new individual HMM model needs to be created, and each model should be compared with all the existing HMM models to get a match, slowing down the recognition process speed. But HMM method is fast in initial training, and when a new voice information is added into the HMM database, only the new voice is used in the training process to create a new HMM model [6]. Compared to the neural network algorithm, for a large number of speech samples, the HMM algorithm provides a higher speech recognition rate.

In both these recognition algorithms, in order to increase the recognition rate, unique feature parameters (Feature) of the signal should be extracted. Even similar words with the same meaning spoken by the same user at different time intervals differ in sound intensity, pitch, and timbre. The voice waveform varies in vocalizing speed, personnel style, and is masked by environmental noise. Speech itself strongly depends on the language. Rate of speech and ambient noise, are considered to be the biggest factors that reduce recognition rate and ways to overcome these challenges are presented [9]. In this paper, a detailed explanation of the HMM speech recognition algorithm along with the challenges to improve speech recognition rate are explained.

2 Voice Feature Parameter Extraction

First, an important issue in speech recognition algorithm is to identify the unique features of speech and to extract the quantitative parameters, which are interwoven in the selection of parameters. Until now, Mel-Frequency Cepstral Coefficients (MFCC) parameters are the most commonly used, but extraction of new parameters of the speech signal can dramatically improve voice recognition. MFCC calculated from a given speech signal to know the hourly cepstrum is usually expressed as the coefficient matrix. MFCC feature parameters are extracted from the voice signal and the procedure used for parameter extraction is described below. During MFCC calculation, the sampling frequency is set to 16,000 Hz for each 10 ms frame. Mel frequency bands were split into 24 bands.

1. Time interval Voice signals (Frame) are multiplied by a Hamming window, after which a FFT power spectra is obtained by conversion.
2. Apply the triangular window of (1) to the Power spectrum and convert it to Mel frequency units.

$$w(n) = \frac{2}{N} \left(\frac{N}{2} - \left| n - \frac{N-1}{2} \right| \right). \tag{1}$$

3. Take the logarithm of the power you have in the Mel frequency units.
4. Take the discrete cosine transform (DCT) of the Mel log spectral power.

3 HMM Recognition Algorithm Overview

For better understanding of HMM a 6×10 MFCC matrix is assumed, where 6 is the number of MFCC coefficients and 10 is the corresponding number of time coefficients.

$$W1P1 = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 2 & 1 & 0 & 1 & 4 & 1 & 2 & 2 & 2 & 1 \\ 2 & 1 & 5 & 1 & 1 & 1 & 2 & 2 & 2 & 1 \\ 2 & 1 & 3 & 1 & 1 & 1 & 2 & 2 & 2 & 1 \\ 2 & 1 & 3 & 1 & 1 & 1 & 2 & 2 & 0 & 1 \\ 2 & 1 & 2 & 1 & 1 & 0 & 2 & 2 & 2 & 1 \end{bmatrix} \tag{2a}$$

$$W1P2 = \begin{bmatrix} 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 2 & 1 & 5 & 1 & 1 & 1 & 2 & 2 & 2 & 1 \\ 2 & 1 & 4 & 1 & 1 & 1 & 2 & 2 & 2 & 1 \\ 2 & 1 & 0 & 1 & 1 & 1 & 2 & 2 & 2 & 1 \\ 2 & 1 & 3 & 1 & 1 & 1 & 0 & 2 & 2 & 1 \\ 2 & 1 & 3 & 1 & 1 & 1 & 2 & 2 & 2 & 1 \end{bmatrix} \tag{2b}$$

$$W2P1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 2 & 2 & 4 & 2 & 2 & 3 & 1 & 1 & 2 & 1 \\ 2 & 5 & 2 & 2 & 2 & 1 & 0 & 1 & 2 & 1 \\ 2 & 2 & 1 & 2 & 2 & 1 & 1 & 1 & 2 & 1 \\ 2 & 2 & 1 & 2 & 2 & 4 & 1 & 1 & 2 & 1 \\ 2 & 2 & 1 & 2 & 2 & 7 & 1 & 1 & 2 & 1 \end{bmatrix} \quad (2c)$$

$$W2P2 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \\ 2 & 2 & 7 & 2 & 2 & 1 & 1 & 1 & 2 & 1 \\ 2 & 6 & 2 & 3 & 2 & 4 & 1 & 1 & 2 & 1 \\ 2 & 2 & 0 & 2 & 2 & 1 & 1 & 1 & 2 & 1 \\ 2 & 2 & 1 & 2 & 2 & 1 & 5 & 1 & 2 & 1 \\ 2 & 2 & 1 & 3 & 2 & 1 & 1 & 1 & 2 & 1 \end{bmatrix} \quad (2d)$$

The four MFCC coefficient matrices shown above are assumed to be two syllables or words (W1, W2) spoken by two (P1, P2) different people. For speaker independent voice recognition, [W1P1] and [W1P2] can be concatenated into [W1], that is, [W1] = [W1P1 W1P2]. Likewise, W2 is denoted by, W2 = [W2P1 W2P2]. The more people we gather speech samples from, W1 can be extended to be, W1 = [W1P1 W1P2 W1P3... W1PN]. This will produce better results due to better convergence. W1 and W2 will each be 6×20 matrices.

As shown in Fig. 1, W1 and W2 MFCC coefficient matrices are transformed into several states and transients, then are modified into sequential probabilistic models. We call the total time, T and the discrete-time is set to $t = \{1, 2, 3, \dots, T\}$. N number of states are denoted as $q = \{q_1, q_2, q_3, \dots, q_N\}$ and M number of events are denoted as $o = \{o_1, o_2, o_3, \dots, o_M\}$. Figure 1 shows two different states (State) and the resulting state transition probability of four a_{ij} is shown. HMM models have an initial steady state, the probability of which is the initial probability, and each probability of transition from one state to another is called transition probability. In addition, the probability of observing a state refers to the probability of another event.

$$\text{Initial probability } \pi_j = P[q_1 = j] \quad 1 \leq j \leq N. \quad (3a)$$

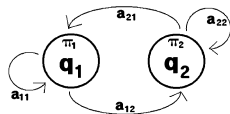
$$\text{Transition probability } a_{ij} = P[q_t = j \mid q_{t-1} = i] \quad 1 \leq i, \quad j \leq N. \quad (3b)$$

$$\text{Observation probability } b_j(k) = b_j(o_t) = P[o_t = e_k \mid q_t = j] \quad 1 \leq j \leq N, \quad 1 \leq k \leq M. \quad (3c)$$

As shown for convenience, A represents a matrix, a_{ij} and B represents a set $b_j(o_t)$, and π represents π_j . Thus the HMM model is denoted by $\lambda = (A, B, \pi)$. If $o = \{o_1, o_2, o_3, \dots, o_T\}$ and $q = \{q_1, q_2, q_3, \dots, q_T\}$, the simultaneous probability of both states and observations happening together is

$$P[o, q \mid \lambda] = \pi_{q_0} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) a_{q_2 q_3} \cdots a_{q_{t-1} q_t} b_{q_t}(o_T).$$

Fig. 1 HMM transforms W1 and W2 MFCC coefficients matrix into some states and state transitions



And the sequential probability of continuous observations $o = \{o_1, o_2, o_3, \dots, o_T\}$ for the model is

$$\begin{aligned}
 P[o|\lambda] &= \sum_{all\ q} P[o|q, \lambda]P[q|\lambda] \\
 &= \sum_{q_1, q_2, q_3, \dots, q_T} \pi_{q_1} b_{q_1}(o_1) a_{q_1 q_2} b_{q_2}(o_2) a_{q_2 q_3} \dots a_{q_{T-1} q_T} b_{q_T}(o_T). \quad (4)
 \end{aligned}$$

The above state representations are defined as vector quantization (VQ), so as to statistically quantize the two-dimensional MFCC coefficients. Each word/syllable goes through the HMM routine individually to produce three variables that help in differentiating between different words/syllables (Learning process, $\lambda_i = (A_i, B_i, \pi_i)$). During the testing phase, the test word is compared with the different HMM models that were calculated during the training phase to find the match that produces the Maximum Log Likelihood.

After the learning process, during the testing phase, each individual word is passed through the HMM (Recognition process, $\lambda_j = (A_j, B_j, \pi_j)$) and the word with the highest similarity (Maximum log likelihood value) against all the words tested is the likely match. For every single word or every syllable (W1 or W2), the same HMM technique is applied separately.

4 HMM Algorithm’s Programming

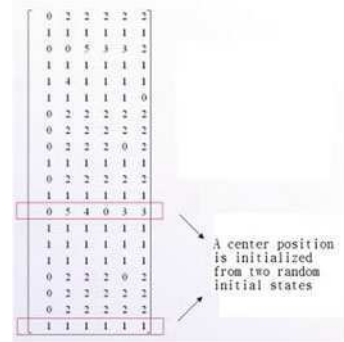
Let us reconsider W1. Both W1 and W2 of the 6×20 matrix are transposed. Initial values of a and π are taken at random, and using W1 we calculate a more accurate a and π . To do this, we start by selecting a random center point μ (Center Points) and σ (Covariance) and follow the set of procedures described below based on W1

1. From W1 by VQ, new μ (Center Points) yields σ^2 (Covariance)

$$\mu_i = \frac{\sum_{i=1}^N x_i}{N}. \quad (5a)$$

$$\sigma_{ii}^2 = \frac{\sum_{i=1}^N (x_i - \mu_i)^2}{N - 1} = \frac{\sum_{i=2}^N x_i^2 - \left(\left(\sum_{i=2}^N x_i \right)^2 / N \right)}{N - 1}. \quad (5b)$$

Fig. 2 If W1 of the 6×20 matrix is transposed, a 20×6 matrix of is formed as MFCC observed data



Where x_i is the input data set, an element of W1 and N is the total number of set elements. The initial values of HMM technique are used to calculate the exact μ and this accelerates the rate of convergence. For example, the centers in Fig. 2 are calculated from the W1

The two initial random center points are given by,

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 5 & 4 & 0 & 3 & 3 \end{bmatrix}. \quad (6)$$

$$\sqrt{\sum_{f=0}^F (x(f) - x_t(f))^2}. \quad (7)$$

The next set of center points of the state which are calculated by the K-means technique is more accurate. The next set of the center points are

$$\begin{bmatrix} 0.5263 & 1.5263 & 1.6316 & 1.5263 & 1.3158 & 1.4211 \\ 0 & 5.0000 & 4.0000 & 0 & 3.0000 & 3.0000 \end{bmatrix}. \quad (8)$$

2. $\pi = \pi_j$ and $A = a_{ij}$ are taken to be random initial matrices.
3. Using the calculated μ , ρ values, we calculate the observation probability, b

$$b = -0.5 * [D * \log(2\pi) + \log(|\sigma|) + d]. \quad (9)$$

Where d is the Squared Euclidean Distance between W1 and μ , D is the length of the spectrum of MFCC, 13, $\log(2\pi) \cong 1.8379$ and $|\sigma|$ is a matrix Determinant.

4. Improved values of π , a, b are obtained from α , β , γ calculations.

When calculating $P[o | \lambda]$, using Eq. 4 to reduce the amount of computation of the $2T \times NT$, we introduce a inductive operation. The variable $\alpha_i(i)$ indicates the probability of observations $o = \{o_1, o_2, o_3, \dots, o_t\}$, and being in state i and time t.

$$\alpha_t(i) = P[o, q_t = i | \lambda]. \quad (10a)$$

$$\alpha_1(i) = \pi_i b_i(o_1) \quad 1 \leq i \leq N. \quad (10b)$$

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^N \alpha_t(i) a_{ij} \right] b_j(o_{t+1}) \quad (10c)$$

$$1 \leq t \leq T-1, 1 \leq j \leq N.$$

$$P[O|\lambda] = \sum_{i=1}^N \alpha_T(i) = \alpha. \quad (10d)$$

$\log(\alpha)$ is defined as the Log Likelihood.

$\beta_t(i)$, the observed probability is obtained as a result of the forward-backward algorithm. $\beta_t(i)$ is defined as the probability of observations, $o = \{o_T, o_{T-1}, o_{T-2}, \dots, o_{t+1}\}$, given that we are in state i at time t

$$\beta_t(i) = P[o | q_t = i, \lambda]. \quad (11a)$$

$$\beta_T(i) = 1 \quad 1 \leq i \leq N. \quad (11b)$$

$$\beta_t(i) = \sum_{j=1}^N \beta_{t+1}(j) a_{ij} b_j(o_{t+1}) \quad (11c)$$

$$t = T-1, T-2, \dots, 1, \quad 1 \leq i \leq N.$$

$$\beta_o(\cdot) = \sum_{i=1}^N \beta_1(j) \pi_j b_j(o_1) = \sum_{i=1}^N \alpha_T(i) = P[o|\lambda]. \quad (11d)$$

The observed data, $\beta_o(\cdot)$ is defined as a total observation probability, that is, o (sequential MFCC data) which occurs sequentially.

And $\gamma_t(i)$ is the probability of being in a state i at time t given an observation sequence, $o = \{o_1, o_2, o_3, \dots, o_T\}$ and a HMM model state.

$$\begin{aligned} \gamma_t(i) &= P[q_t = i | o, \lambda] = P[o, q_t = i | \lambda] / P[o | \lambda] \\ &= P[o, q_t = i | \lambda] / \sum_{j=1}^N (P[o, q_t = i | \lambda]) = \alpha_t(i) \beta_t(i). \end{aligned} \quad (12a)$$

$$\gamma_t(i) = \frac{\alpha_t(i) \beta_t(i)}{\sum_{j=1}^N \alpha_t(j) \beta_t(j)}. \quad (12b)$$

5. a, and b are calculated from $\Sigma \chi$

$\chi_t(i, j)$ is the probability of being in state i at time t , and in state j at time $t+1$ given the observations, $o = \{o_1, o_2, o_3, \dots, o_T\}$ and the HMM model state.

$$\chi_t(i, j) = P[q_t = i, q_{t+1} = j | o, \lambda] \quad 1 \leq t \leq T - 1. \quad (13a)$$

$$= P[q_t = i, q_{t+1} = j, o | \lambda] / P[o | \lambda]. \quad (13b)$$

$$= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{P[o | \lambda]}. \quad (13c)$$

$$= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{k=1}^N \sum_{p=1}^N \alpha_t(k) a_{kp} b_p(o_{t+1}) \beta_{t+1}(p)}. \quad (13d)$$

$$\gamma_t(i) = \sum_{j=1}^N \chi_t(i, j). \quad (14)$$

$\Sigma \chi = \Sigma \chi + \text{norm}(a^*(\alpha(:,t)^*(\beta^*b)'))$ and is calculated as.

6. Log Likelihood (=log($\Sigma \alpha$)) calculations

$$\alpha = P[o | \lambda] = \sum_{i=1}^N \alpha_T(i). \quad (15)$$

7. Improved values of a' and π' are obtained from the following calculations

$$a' = \text{norm}\left(\sum \chi\right) \therefore a'_{ij} = \frac{\sum_{t=1}^{\gamma-1} \chi_t(i, j)}{\sum_{t=1}^{\gamma-1} \gamma_t(i)}. \quad (16a)$$

$$\pi' = \text{norm}(\gamma) \therefore \pi'_j = \gamma_1(j). \quad (16b)$$

8. Improved values of μ' and ρ' are obtained by the following calculations

$$\text{op} = \text{op} + \text{wobs} * W1'. \quad (17a)$$

$$m = m + \Sigma \text{wobs}. \quad (17b)$$

$$\text{op} = \text{op} + \text{wobs} * W1'. \quad (17c)$$

$$\mu' = m / \sum \gamma \therefore \mu'_j = \frac{\sum_{t=1}^T \gamma_t(j) o_t}{\sum_{t=1}^T \gamma_t(j)}. \quad (17d)$$

$$\sigma' = \text{op}' / \sum \gamma - \left(\mu' \times (\mu')^T\right) \therefore \quad (17e)$$

$$\sigma'_j = \frac{\sum_{t=1}^T \gamma_t(j) (o_t - u'_j) (o_t - u'_j)^T}{\sum_{t=1}^T \gamma_t(j)}$$

To improve μ' , σ' , π' and a' , repeat steps three to six and this will provide a maximum likelihood estimate (Maximum Log Likelihood). The above process is also performed for W2. At the end of the training process for each trained word, we have improved set of variables μ' , σ' , π' and a' which are stored.

Now, the test procedure is similar to the training process for newly learned words. For example, consider W'1, which undergoes the process of recognition for the HMM. It undergoes the learning process described above, but steps 1 and 2 are omitted.

Using the variables learned from the previous stage μ' , σ' , π' and a' and following the steps three to six provided in the HMM learning process, Log likelihood ($=\log(\Sigma \alpha)$) is calculated. Since we used two words W1 and W2, in HMM training, the following variables μ' , σ' , π' and a' were calculated for each word. As a result two Log likelihood ($=\log(\Sigma \alpha)$) values were obtained. The value with largest maximum likelihood (Maximum Log likelihood) indicates the recognized word.

5 Apply Theory and Analysis

In order to apply the HMM theory to Korean syllables, I selected the most frequently used 72 Korean syllables, and then they were listed based on the highest frequency of use.

“I HA E GA RA EUL EUI GUI NA NI NEUN RO YEO A LI REUL GI GO SEO GAE
 DEUL JAR SA DA WA NAE EU KI EUN SI KWA DO NEO GEOK HEU DEO HAN
 MYEO HO MAL DAE JOO KKE SEU REU WOO RAM IL MO GEO BO IN SUNG
 SOO DEO JEO YO YEOT IT JE DEUN O SIN NIM SO HAM MOO EL REO SE WE”
 (in English)

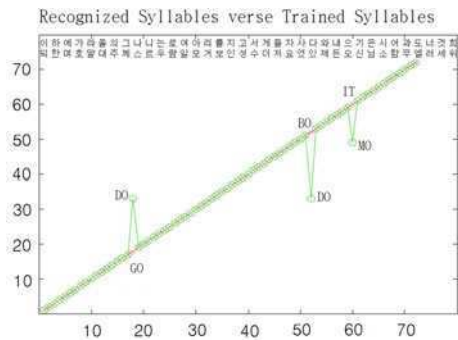
For verifying the speech recognition rate, I recorded the voice of four adult males with normal hearing, then detected an envelope curve of voice signal waveform as you see in Fig. 1a. Then the signals over a certain amplitude were extracted and applied to each corresponding syllable. The following Table 1 lists the recognition rate by the HMM algorithm, after the syllables went through a learning phase, and eventually were tested. I tried different number of Mel frequency indices like 13 or 24 and compared many cases, but for the sake of brevity I have written only the most important result in this paper.

As the number of word samples W1 ($=\{W1P1\ W1P2\ W1P3\ \dots\ W1PN\}$) increase, we see that the recognition rate increases too. This shows that the more data we use in training the higher recognition rate we can achieve. When the Mel frequency index increases, from 13 to 24, the recognition rate increases, but the computation time increases too. Figure 3 shows the recognition rate of 72 syllables tested against the same 72 syllables that were trained in HMM. Table 1 and Fig. 3 show the result of testing syllables with a Mel frequency index of 24. There are three errors out of a total of 72 syllables. Three syllables were recognized incorrectly. “DO” was recognized as “GO” and “BO” was recognized

Table 1 Voice Recognition Rate Result by HMM Algorithm

The number of syllables	The number of melfrequency index	Number of syllable utterance	Recognition rate (%)
32	13	8	56
32	13	16	78
32	13	24	94
72	13	24	94
72	24	24	96

Fig. 3 72 tested syllables against 72 HMM trained syllables



as “DO,” and “IT” was recognized as “MO”. To decrease the speech syllable recognition error rate, increasing the count from 24 to 40 would be reported as the most direct way to reduce the error rate [11]. we can see that it is better to introduce a parameter with more stable accuracy by increasing the number of syllable, when the similarity level of voice signals for the same syllable has large deviation.

6 Conclusion

In this paper, speech recognition HMM technique was applied for the Korean language. For speech recognition, first we should develop a speech recognition engine software program, and then record a person’s or several people’s voice. Subsequently we should divide the speech into sub-units (syllables), and then the syllables should go through the learning process. Increasing the number of samples of the same syllable during learning tends to increase the recognition rate. At the end of the learning process, the re-recorded syllables go through the speech recognition process. This paper describes the core engine of the HMM method, and simple syllables were used for the recognition process. In order to achieve a high recognition rate for different syllables, significant quantitative information of syllables is required. In this paper MFCC parameters were used. MFCC with a Mel frequency index of 24 provides a higher recognition rate (96%/72 syllables).

Speaker dependent recognition requires only a mel frequency index of 14 during training in comparison to the 24 required for speaker independent recognition training. And as the number of training syllables are increased, more significant characteristic features of voice samples need to be developed.

Acknowledgments This study was supported by a grant of the Overseas Buyer Request Related Technology R&D Project for Small & Medium Business Administration, Republic of Korea (Project Number SJ112664).

References

1. Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. *Psychol Rev* 74(6):431–461
2. Cole RA, Scott B (1974) Toward a theory of speech perception. *Psychol Rev* 81:348–374
3. Allen JB (1994) How do humans process and recognize speech? *Proc IEEE* 4:567–577
4. Han HY, Kim JS, Huh KI (1999) A study on speech recognition using recurrent neural networks. *J Acoust Soc Korea* 18(3):62–67
5. Jarnng SS (2009) Application view of voice recognition programming for hearing aids. *Conf J Acoust Soc Korea* 28(2s):76–79
6. Jarnng SS (2010) Speech recognition algorithm understanding about HMM. *Conf J Acoust Soc Korea* 29(1):260–261
7. Rabiner LR, Juang BH (1986) An introduction to Hidden Markov Models. *IEEE ASSP Magazine*, January 1986
8. Ku MW, Eun JK, Lee HS (1991) A comparative study of speaker adaptation methods for HMM-based speech recognition. *J Acoust Soc Korea* 10(3):37–43
9. Ahn TO (2008) HMM-based speech recognition using DMS model and fuzzy concept. *J Korea Ind Acad Technol Soc* 9(4):964–969
10. Jung MH (2010) QoLT Technology Development Projects Candidate Proposal Subplan. Korea Evaluation Institute of Industrial Technology
11. Hosom JP (2009) Speech Recognition with Hidden Markov Models. Oregon Health & Science University <http://www.cslu.ogi.edu/people/hosom/cs552/>

Availability Management in Data Grid

Bakhta Meroufel and Ghalem Belalem

Abstract The data grids are highly distributed environments where nodes are geographically distributed across the globe and shared data are generally very large. The use of replication techniques ensure better availability and easy access to data handled in the grids. In this article, we propose a dynamic replication strategy based on availability and popularity, this replication takes into account failures in the system. The minimum degree of replication is specified by a certain probability of availability and the maximum degree is controlled by the popularity of the data, we introduced also the concept of dynamic primary replica that is used to ensure availability without increasing recovery time. We show in this article that the proposed strategy improves the availability of data according to its popularity and at the same time it improves system performance.

Keywords Data grid · Hierarchical topology · Replication · Availability · Popularity

1 Introduction

Availability is a very important parameter for evaluating a system. Several studies in the literature suggest techniques to ensure the availability, an improvement of 1% of availability are important, corresponding to about 3.5 additional days of

B. Meroufel · G. Belalem (✉)
Department of Computer Science, Faculty of Sciences,
University of Oran (Es Sénia), Oran, Algeria
e-mail: ghalem1dz@gmail.com

B. Meroufel
e-mail: bakhtasba@gmail.com

uptime per year [1]. Replication is a technique used to guarantee the availability of data in the system. In this work we focalize on the availability of data. Static replication sets from the beginning the number of replicas in the system which makes the availability fixe, there are systems that use this strategy as: CFS [2], Glacier [3], GFS [4], IrisStore [5] and MOAT [6]. Unfortunately this type of replication does not take into account the popularity of data in the system. But the dynamic replication [7, 8] is an effective strategy that adapts to large scale systems. It can be used for several reasons such as: reducing response time, improved communication costs, preservation of bandwidth, assurance of data availability and fault tolerance. Although this type of replication ensures good availability for the requested data, in large environments that replication can not avoid some problems such as:

Loss of data due to the unpopularity of these: if there is a change in the popularity of a data (data that is not popular at the time t become popular at the moment t') it may be unavailable (for deletion) or it will be very rare, which increases the response time, increases the server load and degrades availability. There are systems where the availability is defined as the availability of the least available object in the whole system such that the system FARSITE [9] and in the case of a data loss, availability becomes 0%. To resolve this problem there is researches that propose the idea of the primary copy. Each data has one or more replicas that can not be deleted whatever the number of access on this data, in this way, the system guarantees the existence of this data in the system.

Loss of data may be due to the failure of nodes: the failure of a node that contains rare data may cause their unavailability. To resolve this problem, some studies propose that in case of failure, the node replicates all the data it stores, which increases the recovery time. Other works propose to replicate only the primary copies. But if the data is already popular in the system, it will be replicated several times at different nodes; in this case it is inefficient to replicate the primary replica elsewhere in case of failure.

So the primary copies are a good strategy to solve the problems of dynamic replication but they may increase the recovery time in the system and they do not ensure availability. To solve this problem and ensure the availability of data regardless of the popularity of the data, we proposed an approach that combines between replication based on availability offered in the work [1, 10] and replication based on the popularity of the work given in [7, 8] with some improvements. In this approach the minimum number of replicas for each data is the number of replicas that meets the availability desired by the administrator, but the number of replicas can increase depending on the number of petitions requesting this information (popularity). We have also introduced the idea of dynamic primary copy to minimize recovery time in case of failure. At the end we consider the choice of threshold level to control the degree of replication. Our replication approach is articulated on a semi-centralized hierarchical topology.

The remainder of this article is organized as follows: [Sect. 2](#) presents the used topology. [Section 3](#) defines our service of dynamic replication. [Section 4](#) will be reserved for the experimental part; we show in this section that the results are

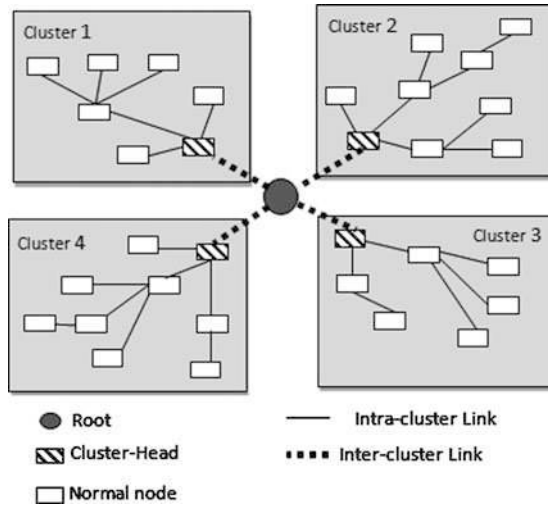


Fig. 1 The topology of work

encouraging from a performance standpoint. We conclude this paper with a summary and some future work.

2 Description of the Used Topology

In our work, we used a hierarchical topology (see Fig. 1). The choice of such a topology is motivated by: minimizing the time of reception of each message and minimizing the number of messages exchanged through the tree architecture. Several global systems use this type of topology (Internet and DIET) [1, 11].

3 Replication Manager

Our proposal for replication manager exploited the powers of dynamic replication and dynamic primary copy taking into account the availability of data and its popularity as well. In our approach, the administrator requires the minimum level of availability for each data. But this availability may increase depending on the popularity of the data. The replication Manager consists of two collaborative sub services: the first subs service is “dynamic replication”, it creates replicas depending on the availability and popularity. Each node in the system can use this service. The second sub service is “availability monitoring”, it minimizing the number of copies without degrading availability. Only the Cluster-Head can use this service to monitor the availability in the cluster.

3.1 Dynamic Replication

The definition of availability is the measure of the frequency or duration in which a service/system is available to the user. To calculate the availability of data we assume the following case: Each node (the component that contains the replica) has a certain probability of stability. The availability of a replica is the stability of the node where it is stored. So if p is the probability of availability of data noted M in a node and if α is the number of replicas of data M then the availability $Avail(M)$ can be calculated as follows:

$$Avail(M) = 1 - (1 - p)^\alpha \quad (1)$$

From this formula we can calculate the number of replicas α needed to have some probability of availability. At first, the administrator requires a certain probability of availability. The desired availability for each data is specified by some parameters such as: access history in prior periods and the importance of the data. The replicas that assure the availability are primary copies. As soon as the number of replicas is known (using the formula 1), the replication manager with its centralized management at the Cluster-Head starts creating primary replicas. We associate with each replica a boolean variable D .

- $D = \text{False}$: indicates that the replica is primary and is created by cluster-head to meet availability. The node is not allowed to delete this replica even if not requested. Nodes that have this type of replicas are the most stable in the system. In case of failure, the node that contains this data will replicate it among the best responsible.

The best responsible are the nodes that have the smallest degree of responsibility and a good stability, there will always be the destination of the primary replica created by the CH or replicated by another node in case of failure. The degree of responsibility is the sum of the sizes of primary replicas in the node (in Bytes).

But what is missing in this strategy is that the data have not the same popularity. The popularity of the data M is calculated by the following formula:

$$pop(M) = \frac{\text{Number of requests demanding } M}{\text{Number of all the requests}}. \quad (2)$$

So it is not to assure the same degree of availability for all data. For this reason we add another type of replication based on popularity. This replication is a non-centralized replication (unlike the case of replication based on availability) because it is triggered by the local replication manager of each node. Each node has a history table that stores the number of accesses to its data. If the number of total access exceeds a certain threshold, the node replicates the data in the best client. The best client is the node that has the greatest number of access on a given data [17]. The replica created in this case is a non-primary copy that has $D = \text{True}$.

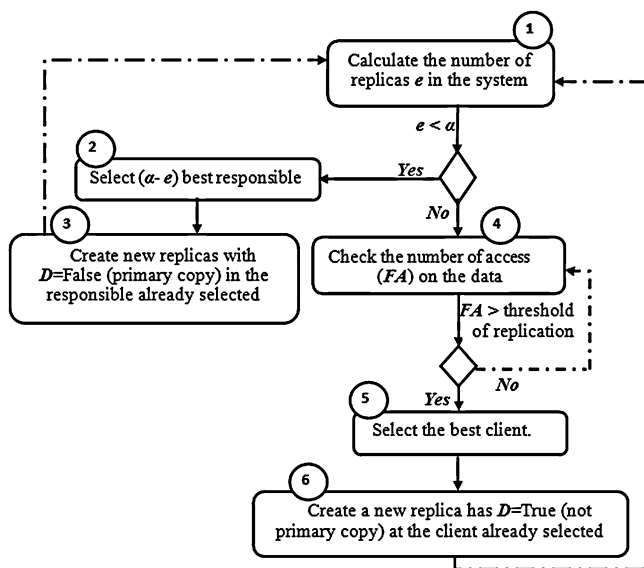


Fig. 2 Steps of dynamic replication sub service in our approach

- $D = \text{True}$: indicates that the replica is not primary and it is created by a node to meet the demands on the data. The node that contains the replica can remove it to store more popular data in local disk. In case of failure, it is not necessary to replicate this data elsewhere.

The diagram of the Fig. 2 shows the steps of replication in our approach. The first step is to compare the number of replicas that really exists in the system (e) with the number of replicas that meets the desired availability (α).

- If $e < \alpha$: then the system performs steps 2 and 3 to satisfy the availability.
- If $e \geq \alpha$: then the system executes the steps 4–6 to satisfy popularity.

The broken lines in the diagram indicate that the system must wait some time before executing the next stage (between 6 and 1 for example).

A best responsible can store the new replica if it has a sufficient memory size, if it has not; it removes the non-primary data starting with the lower frequency data access. In the case of best client, that client can remove only the non-primary data that have access rate less the rate of access of the data they want to store.

In this strategy, the number of primary copies is static which increases the degree of responsibility of the nodes and also the recovery time after each failure. We also note that in case of failure it is unnecessary to replicate elsewhere a primary copy of a popular data because it already exceeds availability desired. To resolve this problem, we added a service known as: monitoring availability.

3.2 Availability Monitoring

The objectives of this sub service consist of: ensures that the number of replicas α that satisfies the desired availability is always respected, whatever the popularity of the data taking into account failures in the system and minimize system recovery time. The fact that the monitoring service availability is localized within the cluster, it can be triggered by each CH. This CH uses three types of messages:

- Request message of replication: is sent to node to replicate the data.
- Message “Fixed” is sent to nodes that have the reply with $D = \text{True}$ (non-primary copy) to transform to $D = \text{False}$ (primary copy).
- Message “Relax” is sent to nodes that have the reply with $D = \text{False}$ (primary copy) to be transformed into $D = \text{True}$.

We call $e(M)$ is the number of replicas of the data M which exists in a cluster. Each period, the CH checks whether:

- The first case: If $e(M) < \alpha(M)$, the CH sends a fixe message to all replica sets that exist in the cluster. And sends requests to the best responsible for storing $(\alpha(M) - e(M))$ replicas that ensure the availability desired.
- The second case: If $e(M) = \alpha(M)$, the CH sends a fixe message to the replicas that exist in the cluster.
- The third case: If $e(M) > \alpha(M)$, the CH sends a message to relax $\delta(M)$ of the replicas, this number is calculated by formula (3)

$$\delta(M) = \text{Min}(\beta(M), r(M) - 2(\alpha(M) - \beta(M))) \quad (3)$$

where $\beta(M)$ is the number of primary copies that exist in the system. The process of monitoring availability is summarized in the following diagram (See Fig. 3).

4 Experimental Results

To validate the proposed approach we used the simulator FTsim (Fault Tolerance simulator) developed in Java [12]. In this section, we simulated different scenarios to study our approach and the impact of various parameters on system performance. We also Compare this approach (DR + Av: Dynamic Replication + Availability) with the classical dynamic replication approach (CDR) proposed in [7].

4.1 Response Time

The number of requests affects the response time of the system (see Fig. 4). We note that for both approaches (ours and CDR) the response time decreases if the number of requests increases. We also note that the response time of our approach

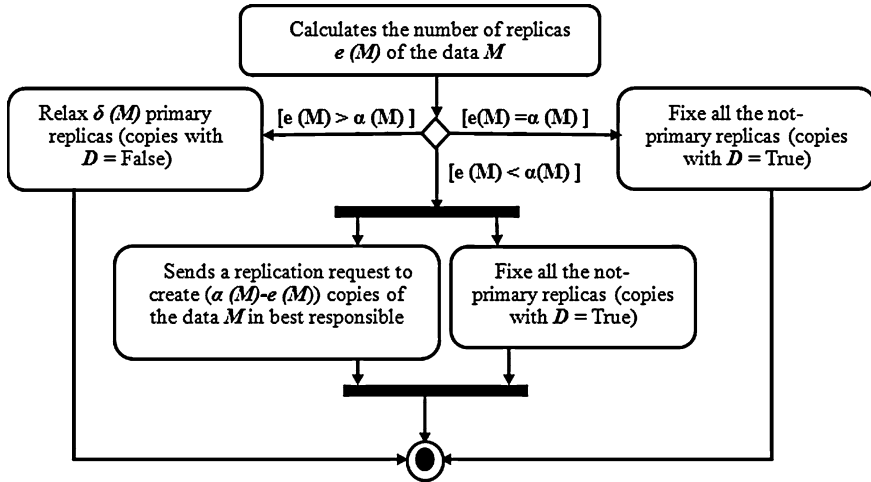
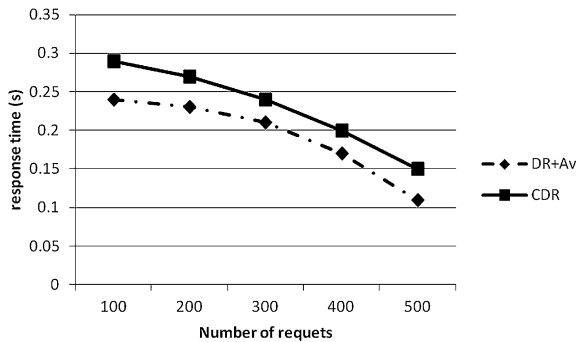


Fig. 3 Steps of availability monitoring sub-service

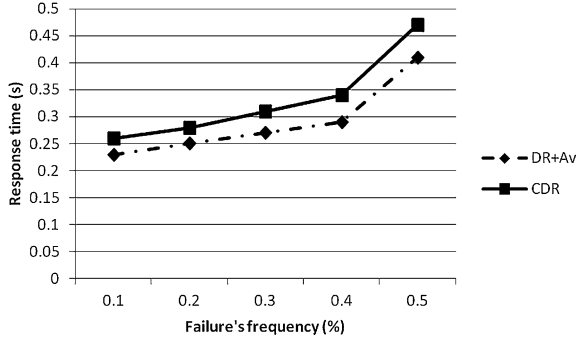
Fig. 4 Impact of number of requests on response time



is better than the approach of CDR guarantees the existence of the data in the system if it is popular otherwise this data will be lost or at least rare (because of the unpopularity). In the event of a change in the frequency access on a given data (if a data unpopular who becomes popular), the access on nodes will be overload before begin to create new replicas in the system, which increases the response time. Our approach (DR + Av) guarantees the existence of a data regardless of its popularity, which minimizes the response time.

In second series of experiments, we study the impact of the frequency of failures (ratio between the number of failed nodes and the number of all nodes in the system) on the response time, and the results are shown in Fig. 5. The number of failures increases the response time because there are lost replicas, but our approach gives good results compared to traditional replication. In CDR: unlike a rare data, the

Fig. 5 Impact of Failures' frequency on response time



failure of a node that contains a popular data shall not cause the loss of this data but it minimizes its availability. In our approach (DR + Av), whatever the number of failures or unpopularity of the data, the system always assures the desired availability which improves response time. The strategy of best responsible guarantees a good distribution of primary copies. The average gain in our approach is 12%.

4.2 SFMR (System File Missing Rate)

The second metric in our simulations is *SFMR*. *SFMR* is the ratio between the number of unavailable data and the number of data requested by queries we call it also (unavailability of requests). This parameter is proposed in work [13] where the authors proved that the minimization of this parameter indicates a good availability for the data system. According to [13] *SFMR* is calculated by the following function:

$$SFMR = \frac{\sum_{i=1}^n \sum_{j=1}^m (1 - P_j)}{\sum_{i=1}^n m_i} \quad (4)$$

where n is the total number of jobs, each job request to access m data. P_j is the availability of the data. In our case the job is a request and each request requires access to a single given time ($m = 1$). We studied the impact of the number of requests on the unavailability of requests. The results in Fig. 6 show that the SFMR decrease if the number of request increase in both strategy of replication because if the number of requests augment, the data concerned will be replicated and its availability will increase. We remark also that our dynamic replication (DR + Av) minimize the SFMR parameter, especially in the case where the frequency of access to the requested data is changed. The gain is estimated by 14%.

The number of failures also infects the unavailability of requests; this result is confirmed by the simulation of the second scenario (see Fig. 7). Queries laced in a

Fig. 6 Impact of number of requests on SFMR

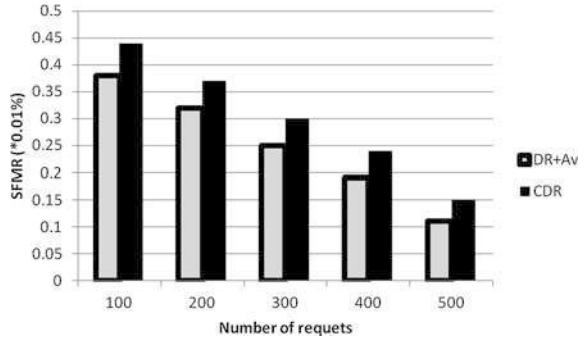
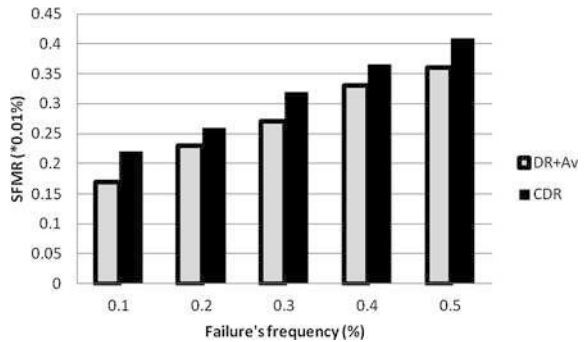


Fig. 7 Impact of failure's frequency on SFMR



system that uses CDR are less satisfied compared to queries made in the system that adopts our approach that ensures availability of data.

4.3 Availability and Recovery Time

In these experimentations, system availability is the average availability of all data. We measured the availability of our approach and that of the CDR. The results illustrated in Fig. 8 shows that the average availability of data in a system that uses our approach is that better then the availability assured by the approach CDR. Our approach (DR + Av) provides at least the desired availability, but it increases the availability by the popularity of the data. In case of CDR, data are available only if it is requested otherwise it may be lost because of failures or unpopularity.

The recovery time is the time required to replicate the primary data elsewhere (best responsible) in case of failure. In this experiment, we studied the recovery time in two cases: a replication manager that uses the monitoring of availability

Fig. 8 Impact of failure's frequency on availability

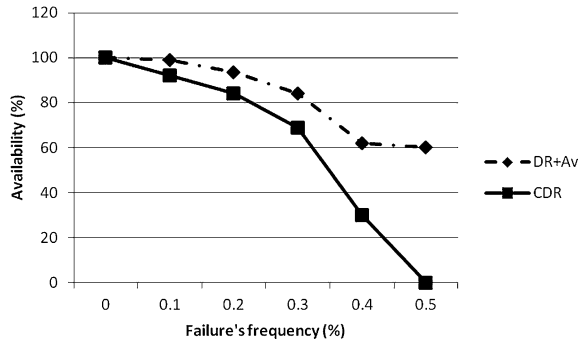
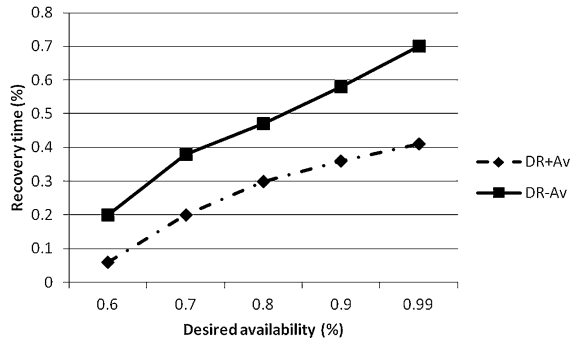


Fig. 9 Impact of desired availability on recovery time



(DR + Av) and another manager who does not use the monitoring of availability (DR - Av). In both cases with a replication manager monitoring availability and replication without the availability monitoring, increasing the desired availability increases the recovery time because the number of primary copies is also increasing and therefore increases the time needed to replicate these copies at best responsible (see Fig. 9). Despite the replication with availability monitoring has minimized recovery time by 42% compared with replication alone, because in the DR + Av cluster-head begins to relax the primary replicas of popular data.

5 Conclusion

In this work we presented our replication manager that uses the dynamic replication service that increases the availability of the data according to its popularity and the availability monitoring service that assure the desired availability without

increasing the response time. The experiments we did prove that our replication manager is better than a classical dynamic replication manager in terms of response time and availability. In future work, we propose to extend the approach proposed by a multi-agent decision-making within each Cluster-Head for the removal of replicas, by taking into account the replicas consistency, and the medium term, we propose to integrate our solution in the Globus middleware.

References

1. Lei M, Vrbsky S (2008) An on-line replication strategy to increase availability in data grids. *Future Gener Comput Syst* 24(2):85–98
2. Dabek F, Kaashoek MF, Karger D, Morris R, Stoica I (2001) Wide-area cooperative storage with CFS. In: *Proceedings of the 18th ACM symposium on operating systems principles*, Banff, Canada, Oct 2001, pp 202–215
3. Haeberlen A, Mislove A, Druschel P (2005) Glacier: Highly durable, decentralized storage despite massive correlated failures. In: *Proceedings of the Second USENIX symposium on networked systems design and implementation*, Boston, May 2005, pp 143–158
4. Ghemawat S, Gobioff H, Leung ST (2003) The google file system. In: *Proceedings of the 19th ACM symposium on operating systems principles*, Bolton Landing, NY, Oct 2003, pp 29–43
5. Nath S, Yu H, Gibbons PB, Seshan S (2006) Subtleties in tolerating correlated failures in wide-area storage systems. In: *Proceedings of the third USENIX symposium on networked systems design and implementation*, San Jose, CA, May 2006, pp 225–238
6. Yu H, Gibbons PB, Nath S (2006) Availability of multi-object operations. In: *Proceedings of the third USENIX symposium on networked systems design and implementation*, San Jose, CA, May 2006, pp 211–224
7. Min Park S, Kim J-H, Ko Y-B, Yoon W-S (2003) Dynamic data grid replication strategy based on internet hierarchy. *Second international workshop on grid and cooperative computing (GCC'2003)* Shanghai, China, Dec
8. Madi KM, Hassan S (2008) Dynamic replication algorithm in data grid: survey. In: *International conference on network applications, protocols and services 2008 (NetApps2008)*, ISBN 978-983-2078-33-3, on 21–22 Nov 2008
9. Douceur JR, Wattenhofer RP (2001) Competitive hill-climbing strategies for replica placement in a distributed file system. In: *Proceedings of the 15th international symposium on distributed computing*, Lisboa, Portugal, Oct 2001, pp 48–62
10. Huu T, Segarra M-T, Gilliot J-M (2008) Un système adaptatif de placement de données. In: *CFSE'6*, Fribourg, Switzerland, 11–13 Feb 2008
11. Lamhamedi H, Szymansky B, Shentu Z, Deelman E. (2002) Data replication strategies in grid environments. In: *Proceedings of the 5th international conference on algorithms and architectures for parallel processing (ICA3PP'02)* IEEE CS Press, Los Alamitos
12. Meroufel B (2011) Fault tolerance in data grid. These Master. University of Oran, Alegria, March
13. Lei M, Vrbsky S (2006) A data replication strategy to increase availability in data Grids. In: *Grid computing and applications*, Las Vegas, NV, pp. 221–227
14. Foster I (2002) The grid: a new infrastructure for 21st century science. *Phys Today* 55(2): 42–47

Mobi4D: Mobile Value-Adding Service Delivery Platform

Ishmael Makitla and Thomas Fogwill

Abstract Mobi4D is a generic mobile services delivery platform that simplifies development of mobile value-added services by offering reusable communication and shared resource components as part of an extendible IP-centric service delivery framework. As a communication service delivery platform, it is based on the JAIN SLEE specification which was developed by the Java Community Process under the JAIN Initiative. The JAIN SLEE architecture provides an abstraction between end user services and the underlying telecommunication networks and their protocols, thus simplifying the development of converged Information Technology and telecommunication applications. This paper gives a technical overview of the Mobi4D platform that is being developed within the Next Generation ICT and Mobile Architectures and Systems research group of the CSIR Meraka Institute. It also highlights the opportunities that such a platform presents to the developing world, particularly in light of the rapid penetration of mobile phones and related technologies in these regions.

Keywords Mobi4D · JSLEE JAIN SLEE · Service delivery platform · Converged communications applications

The Mobi4D Project is undertaken by the Meraka Institute of the South African Council for Scientific and Industrial Research (CSIR) and sponsored by the Department of Science and Technology of the Republic of South Africa.

I. Makitla (✉) · T. Fogwill
Council for Scientific and Industrial Research (CSIR),
Meraka Institute, Pretoria, South Africa
e-mail: imakitla@csir.co.za

T. Fogwill
e-mail: tfogwill@csir.co.za

1 Introduction

The wide-spread adoption and usage of mobile phones in developing regions makes mobile communication and computing a viable platform for development. The mobile phone offers great opportunities for developmental impact, potentially allowing ordinary phone users to use the cellular phone as a crucial ICT tool for empowerment and development in various sectors such as education, health, government and business. The Mobi4D platform leverages this potential, by enabling non-telecommunications developers to easily and rapidly build mobile services. It offers them a robust framework, together with a library of re-usable and integrated components, which together abstract the technical details of the underlying telecommunications and mobile technologies. This alleviates the need for developers to possess in-depth technical knowledge of mobile technologies, allowing them to rather focus on the business and interaction logic of their applications and services.

Mobi4D is a communication service delivery platform based on the Java API for Integrated Networks Service Logic Execution Environment's (JAIN SLEE, or JSLEE) specification developed through the Java Community Process (JCP). Mobi4D is based on the open source Mobicents platform. Mobicents is, as of the writing of this paper, the only open source and certified JSLEE implementation [1]. A technical overview of the platform and its technologies are discussed in the remainder of this paper.

2 Technology Description

2.1 The JSLEE Over View

Java Community Process (JCP) and the Java Specification Participation Agreement (JSPA) carry out the development of Java APIs for Integrated networks (JAIN) [2]. The objectives of the JAIN initiative are to define application programming interfaces APIs for application development, as well as a set of lower-level APIs for signalling protocols such as Session Initiation Protocol (SIP) and Signaling System #7 (SS-7). The JAIN program had to ensure that the following requirements were met:

- Service portability, to allow services to run on any JAIN-compliant environment.
- Network independence, to provide APIs that abstract the complexity of the underlying network infrastructure from the service logic.
- Open development, to provide Java industry standards to transform telecommunication systems into more open environments.

The development of the JAIN API specification led to the creation of a specification and architecture for an environment for execution of service logic,

known as the JAIN Service Logic Execution Environment (JSLEE). The JAIN API specifications covered the following two aspects with regard to the service logic execution environment:

- A specification for container interfaces which specifies APIs for service execution environment (SLEE) that can support low-latency, high throughput and other stringent requirements of the communications domain.
- A specification for service development APIs for distributed communication applications.

The challenge of ensuring service interoperability required a standardised execution environment that could host services from different vendors, or developed using different technologies, provided they comply with the SLEE standard. SLEE standardisation thus ensures and promotes service portability and interoperability.

The Devoteam white paper [3, p. 21] lists the following as key features that a SLEE should have in order to support interoperability:

- Portability of services over different SLEE vendors that support the standard, through standardised APIs, objects and methods.
- Operating Systems (OS), hardware, platform, and network architecture independence.
- A common framework providing the generic services of a SLEE (timers, statistics, fault tolerance, etc.).
- A modular architecture, allowing interoperability with legacy, state-of-the-art, and next generation service networks.

The JSLEE which is specified through the JCP meets these requirements [3]. JSLEE provides “tools” for building a service execution framework [3]. According to the JSLEE Specification 1.1 document [2], JSLEE brings service portability, convergence and secure network access to telephony and data networks.

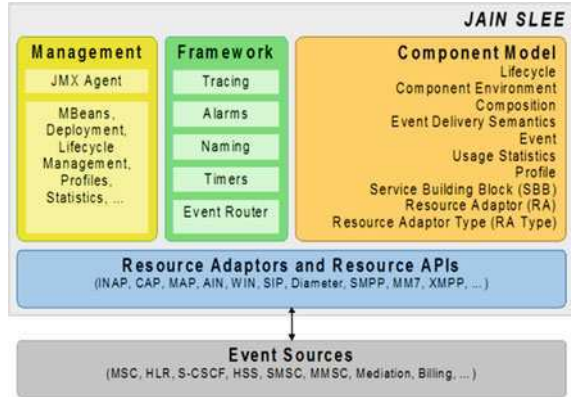
The JSLEE Specification 1.1 document [2, p. 23] identifies some of the goals of the JSLEE architecture as follows:

- Defining the standard component architecture for building distributed, object-oriented communication applications using the Java programming language.
- Allowing the development of these distributed communication applications by combining different components from different vendors, developed using different tools.
- Adopting the “Write Once, Run Anywhere” philosophy of Java to support portability of service components.
- Defining interfaces that enable communication applications from multiple vendors to interoperate.

The JSLEE specification describes a number of key elements of the architecture, including the following:

- Resource adaptors (RAs): resources are technologies and systems outside the SLEE, that the SLEE interacts with. Examples include networks, protocol

Fig. 1 JAIN SLEE architecture [4]



stacks, directories and databases. RAs are software components that adapt and translate the interfaces and requirements of these resource into interfaces and requirements understood by JSLEE.

- **Service Building blocks (SBBs):** the JSLEE is a component architecture. Whereas RAs are components that encapsulate access to and control of external resources, SBBs are atomic service components, and contain the actual service and application logic. SBBs are intended to be atomic, self-contained, reusable and portable across compliant platforms. They represent the smallest self-contained units of service functionality (i.e. components). SBBs are combined, composed and orchestrated to form larger services that are consumed by users/subscribers.
- **Events:** JSLEE has an event-oriented component model, and uses an asynchronous invocation model based on events. SBBs send and receive asynchronous messages as events, which are queued, prioritised and managed by the SLEE on an internal event bus. The SLEE offers sophisticated event distribution and management mechanisms, and uses type and an event subscription model to map events onto the appropriate processing components (SBBs). SLEE events are typically fine-grained, and of high frequency.

Figure 1 depicts the JSLEE architecture and indicates how the JSLEE architecture addresses the requirements of the SLEE, as well as the network abstraction concerns.

2.2 *Mobi4D and the JSLEE*

Mobi4D is a services delivery platform that simplifies development of value-adding mobile services by offering reusable communication, service and shared resource components as part of an extendible IP-centric service framework. It is based on JSLEE and is based on Mobicents, which is an open source, certified implementation of the JSLEE 1.1 specification. The Mobi4D platform and architecture are described in the next section.

3 The Mobi4d Platform

Mobi4D is a communication service delivery platform based on the JSLEE architecture. The rationale behind the development of Mobi4D was to realise the advantage of growing penetration of the mobile phone as primary ICT devices in Africa, by empowering developers to create converged, IP-centric applications and services and to easily deliver those services to mobile devices. Mobi4D strives to lower the barrier to entry for non-telecommunications developers by shielding them from the lower-level technical details of the telecommunications protocols, allowing them to focus on the logic of their service.

Building on the conceptual platform depicted in Fig. 2, Mobi4D was envisaged to be network protocol agnostic. This implies that a request coming into the platform could come from any network, using any protocol. The Resource Adaptor (RA) layer adapts this external protocol into a format understandable internally by the platform and by service components. The resource in this case could be a protocol stack that represents the network from which a request came; it could also be an interface into external application servers through an API. Figure 2 depicts a simplified view of the Mobi4D internal architecture.

Discussions of the major components of the platform are given in subsequent sub-sections.

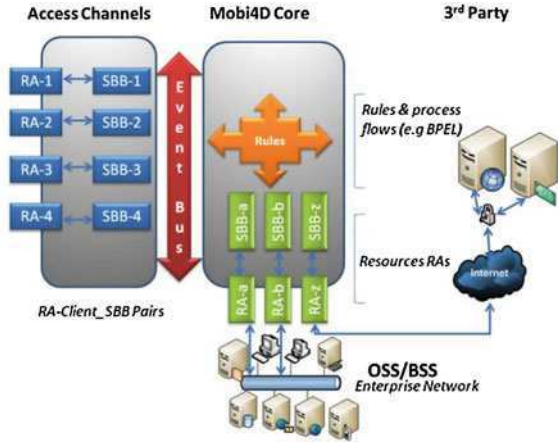
3.1 Platform Access Channels

The first layer (Access channels) connects Mobi4D to underlying access networks. It comprises a set Resource Adaptors (RAs) that serve as network protocol abstraction mechanism and adapts Mobi4D to different communication protocols supported by the underlying access networks. These RAs are technology-specific protocol stack implementations and will be called protocol-RAs throughout this document. Each protocol-RA is paired with a corresponding service building block (SBB). These SBBs (SBB-1 to SBB-4 in the Fig. 4) are called protocol-SBBs and serve as an additional abstraction layer. They act as clients to the protocol-RAs, process events that are passed to and from the protocol_RAs, and translate these events into requests that are understood by the non-protocol SBBs within the platform. The protocol-SBBs have some knowledge of the protocols supported by the protocol-RAs to which they are attached, and define Resource Adaptor bindings in their descriptors, which represent the logical links between the SBBs and the RAs.

3.2 Event Bus

The Event Bus (EB) is placed between the Access Channel layer and the Mobi4D Core in the conceptual architecture. The EB is where all the events defined and

Fig. 2 Simplified Mobi4D architecture



supported by the JSLEE container, RAs and SBBs are fired, fetched, distributed and managed. JSLEE follows the Subscribe-Notify event model, therefore the SBBs that are interested in receiving certain events must specify these events explicitly and when these events occur, the SLEE event router will deliver them to the interested SBBs.

3.3 *Mobi4D Core*

The Mobi4D core is the architectural layer within which service and internal application logic is executed. It contains a class of SBBs that are completely independent of the underlying RAs. These SBBs contain the necessary business logic to provide services regardless of the RAs from which the request came, and regardless of the channel on which the response should be sent. In fact, these SBBs only communicate with protocol-RAs via the protocol-SBBs, not directly. An example of such an SBB is a lookup service that receives a lookup key and returns a corresponding value from some dictionary. The lookup SBB is solely responsible for its own service logic (find and returning the correct value for the requested key), and carries no knowledge of the communication channels or networks on which the request was received. It is this design principle that makes it possible for Mobi4D to provide network and protocol agnostic delivery of services.

The core controller and processing engine that orchestrates, controls and coordinates the flow of execution of services hosted within the Mobi4D core using a rules engine. The processing logic is specified as a set of rules, together with a set of “commands”. The rules are interpreted by the rules engine, which determines the appropriate flow of execution, and is responsible for invoking the correct “command”. For each “command”, there is an SBB implementing the Command Pattern, that is responsible for processing that “command”—it is to these SBBs that

the rules engine delegates control. The use of the rules engine allows for great flexibility, as the flow of execution for services can be changed through the rules editor, without having to modify or redeploy any SBB code. The rules engine is currently implemented using Business Logic Integration Platform called Drools [5].

The Mobi4D core SBBs are either service endpoints themselves, or they forward requests (as events) through to the third party, external services in the enterprise Information Technology (IT) domain through the resource layer, which is described in the next section.

3.4 Mobi4D Resources Domain

The final Mobi4D architectural layer is the resource layer. It defines pairs of RAs (resource-RAs) and SBBs (resource-SBBs), similar to the protocol-RA and protocol-SBB pairs. In this case, the underlying resource-RAs are not communication protocol abstractions, but rather APIs for accessing resources within the IT enterprise, and external parties. Essentially, these resource-RA resource-SBB pairs allow Mobi4D to access internal and external systems, information sources and services. This is typically achieved through the use of technologies such as Service Oriented Architectures (SOA), Web Services and standard protocols such as Simple Object Access Protocol (SOAP). Examples include: web news feeds over Hypertext Transfer Protocol (HTTP), directory services over Lightweight Directory Access Protocol (LDAP), and Operations Support Systems/Business Support Systems (OSS/BSS) like accounting and Customer Relationships Management (CRM) systems.

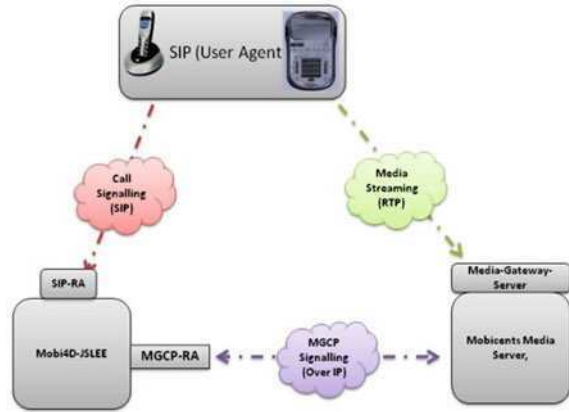
SBBs in core layer access external resource via the resource-SBBs, which in turn delegate to the resource-RAs. The function of the resource-SBBs and resource-RAs is to channel the requests to the appropriate service providers (third parties or internal IT systems). By doing this, the core SBBs are completely independent of the underlying networks via which the resources are accessed (whether HTTP, SMPP, XMPP, SIP, etc.) and it is this design principle that further contributes towards making Mobi4D truly network and protocol agnostic.

4 Mobi4d: State of the Project

4.1 Current Capabilities

The first phase of the platform development was aimed at providing a sufficient proof of concept by developing Resource Adaptors (RAs) for popular mobile services such as Short Message Service (SMS), Unstructured Supplementary Service Data (USSD), and Extensible Messaging and Presence Protocol (XMPP)

Fig. 3 Access-agnostic weather service implementation example



used in Instant Messaging (IM). Currently the SMS, USSD and IM RAs and their respective protocol-SBBs are fully functional, and a Simple Short Message Interface (SSMI) RA has been developed which connects these components to a mobile network aggregator. This aggregator acts as a gateway for sending and receiving SMS and USSD using the proprietary SSMI protocol.

For the IM/chat service access, libpurple and XMPP protocol-RAs have been developed. Along with the protocol-SBBs, these protocol-RAs enable an instance of service to connect to multiple IM service providers using multiple accounts for each of these IM services. This means that an end-user is able to “chat” with the platform through different IM accounts by adding the relevant service address as one of his/her contacts.

The Voice over Internet Protocol (VoIP) capability has been added as one of the platform’s access mechanisms. In the current design, Mobicents Media Server and Mobicents JSLEE [6] are used as media gateway and call control agent, respectively. VoIP functionality is made possible using a Session Initialization Protocol (SIP) protocol-RA, a Media Gateway Control Protocol (MGCP) protocol-RA, and a single protocol-SBB that handles both the SIP messages and simultaneously acts as call control agent using the MGCP to interact with the media server. In addition, a resource-RA and resource-SBB have been developed to access an external speech synthesis server to provide text-to-speech (TTS) services. The TTS server provides speech synthesis for a number of indigenous South African languages. The VoIP and TTS capabilities enable the development of Interactive Voice Response (IVR) applications on top of Mobi4D. The basic architecture of the Mobi4D voice capabilities is depicted in Fig. 3.

A Keyword service was developed to provide an easy-to-configure keyword lookup service; the lookup service was designed to allow its owner to define how keyword request responses are to be rendered back to the end user. The Keyword SBB provides text-based user-system interaction as well as lookup and delegation services to other SBBs within the platform.

In addition, the Authentication, Authorization and Accounting module (AAA) access control module has also been developed. The current module uses open LDAP, an open source directory service implementing LDAP, for user white- or blacklisting and call-blocking. An LDAP resource-RA was developed with an API client to enable the platform to communicate with the open LDAP directory server to perform directory lookups. This implementation uses the group concept of LDAP to define service access groups and adds users to these groups to grant or block access to services. The Diameter Credit Control Application [1] is used to provide accounting and charging for resource usage as part of the AAA functionality of Mobi4D.

4.2 Opportunities for Developmental Initiatives

This section seeks to advocate the case for Mobi4D in supporting developmental projects.

Extending the reach—The World Bank’s report on ICT4D [3] discusses the developmental impact of ICTs and highlights the need to expand the reach and increasing impact of ICT4D initiatives in developing countries.

Mobi4D provides a technology-agnostic service delivery platform, allowing users to access content from IM clients, SMS, USSD or HTTP, whichever technology the user’s device is able to support. This is particularly helpful for making information accessible to end-users in environments where higher-end phones are not pervasive, and where only the so-called bottom-of-the-pyramid mobile capabilities (SMS, USSD, Voice) can be assumed to exist. To cater for textually-illiterate portion of the user population, the use of voice to access information offers promise, particularly when paired with local-language TTS services. Mobi4D’s IVR service is used to achieve this end.

Capitalizing on available technological capabilities—it is important to determine which technologies the resource-constrained user communities already possess, and to offer services supported by the capabilities of those technologies [7]. Although full device agnostic service delivery may pose many technical challenges, particularly relating to the type, format and size of media (text, audio, video), network abstraction is a key milestone towards this goal. The Mobi4D platform achieves network and protocol abstraction through the Resource Adaptor framework of the JSLEE architecture.

4.3 Mobi4D Value Proposition

JAIN SLEE is known for its steep learning curve [8]. Mobi4D addresses this by enabling a non-telecommunications expert developer to define new service

functionalities by defining set of rule and processes using rules engine in the Mobi4D core (see Fig. 2).

Furthermore, one of the known limitations of the JAIN SLEE architecture is the tight coupling between the protocol-to-java object mapping and the SBB, which limits the portability of the SBBs [8]. Mobi4D addresses this by defining the protocol-RA-SBB pairs that handle protocol-specific signalling (at the RA level) and the protocol-specific Java event objects (at the RA-bound SBB). These protocol-RA-SBB pairs transform protocol-specific events into access-agnostic events understood by the SBBs within the Mobi4D core. The SBBs within the Mobi4D core are independent of both the protocols and service functionalities and are highly reusable and portable across services and even JAIN SLEE containers.

4.4 Mobi4D Demonstrators

One of the demonstrator services hosted on Mobi4D platform is a “Weather service”, developed as a means to demonstrate the network agnostics of the platform; it makes use of the Keyword service to delegate control to the Weather SBB, which performs an external lookup via the resource layer to a website containing weather information. The weather information is repackaged and returned to the user. Due to the network agnostics of Mobi4D, it is possible to send a keyword “Weather” from an IM application, via SMS, USSD, and even through a web browser (using HTTP), to access the same weather information, via the same Weather SBB. It is also possible to access the weather service through a SIP-based Interactive Voice Response (IVR) system which uses Text-To-Speech technology to render audio version of the retrieved weather information, this audio functionality is particularly useful to cater for the textual illiterate users in developing regions such as Africa’s rural areas. [Section 4.5](#) presents the implementation example.

Another demonstrator service involves the use of SMS and USSD at a local academic conference hosted at the South African Council for Scientific and Industrial Research (CSIR). This allowed the conference delegates to access the conference programme and to comment on speaker presentations using SMS and USSD, in real-time. Questions posed were viewed by the session chair, who would then read them out to the speaker to address the audience.

4.5 Mobi4D Implementation Example: Weather-Service

This section presents the implementation example of Mobi4D and discusses how a weather service, which is traditionally accessible only through the provider’s website, has been made accessible through USSD, SMS, IM and voice using a SIP-based IVR.

Scenario:

Context: a farming rural community with basic communications infrastructure. Individual residents have personal computing and communications devices of various technological capabilities; some have powerful smart phones, others have very basic SMS-Call-Only mobile phones, while yet others have computers with Internet access.

Potential access-technologies supported by the collective technological capabilities are:

- SMS
- USSD
- IM (e.g. Mxit)
- Voice-Calls
- Web (HTTP)

Purpose: the community would like to get on-time, up-to-date weather information in order to properly plan their farming activities.

Typical challenge: the weather service is only available online from website for free. However phones that do not have Internet browsing capabilities cannot access this service. How might the same weather service be made accessible through all other access-technologies supported by other personal computing and communication devices within the community?

Solution approach: deliver the weather service through an access-agnostic service delivery platform.

Platform access channels/protocol-RA-SBB pairs:

- SSMI-RA and SSMI SBB (handle both USSD and SMS)
- SIP-RA and SBB handle SIP voice calls
- HTTP-Servlet-RA and SBB handle incoming web-based requests (HTTP)
- MxitGateway-RA and SBB handles incoming Mxit chats (Instant Messaging)

Access-agnostic Mobi4D core:

- *Weather-Service-SBB:* receives weather requests from any underlying access network represented by the protocol-RA-SBB pairs. Once it retrieves the weather information from the service provider, it sends the response (as a JSLEE Event) back to the requesting protocol-RA-SBB pairs.
- *Text-to-Speech Service SBB:* this helper service receives requests from within the delivery platform to convert from text to speech (audio). These requests can originate from the Protocol-RA-SBB pairs or from other access-agnostic SBBs.

Mobi4D resource domain:

- HTTP-client RA and SBB sends access-agnostic SBBs' HTTP requests (GET or POST) to remote Web-based services hosted at the service provider's domain.

The schematic representation is provided in Fig. 4:

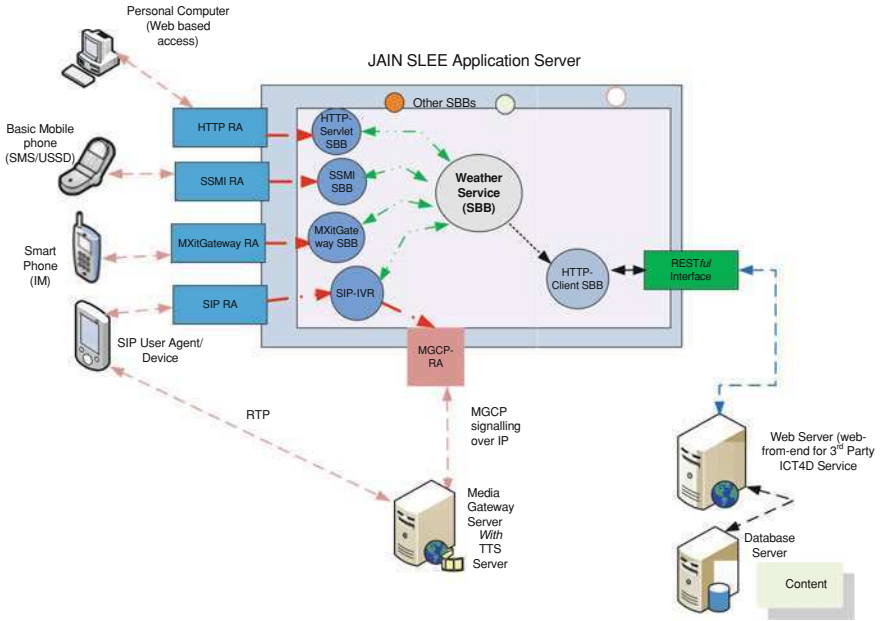


Fig. 4 Basic Mobi4D IVR architecture

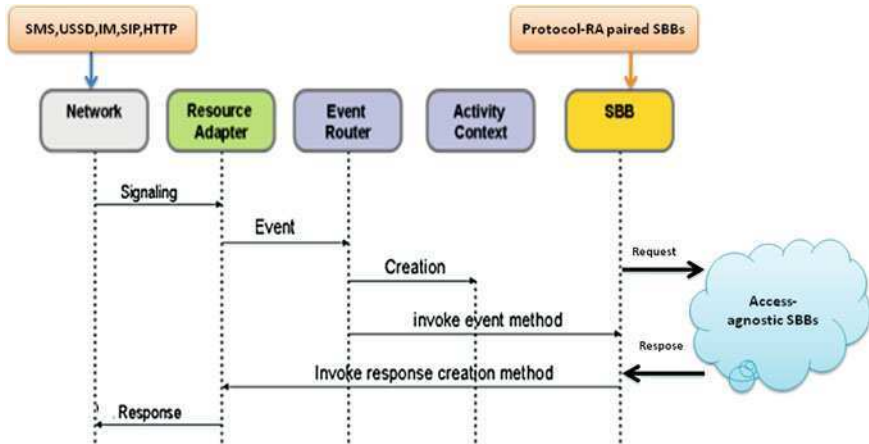


Fig. 5 Event flow and processing within Mobi4D

The flow of events and processing that happens within the platform is presented in Fig. 5

When a user send a request either through SMS, USSD, Instant Messenger (MXit), VoIP voice call (SIP) or Web browser client (HTTP), an access-technology specific signalling begins, the protocol-RA for the specific access network

receives the request, it generates a JAIN SLEE typed event and hands it over to the event router, the event router follows the event delivery semantics to invoke event processing logic on the SBBs (these SBBs are bound to the protocol-RAs as part of a protocol-RA-SBB pair). The event processing logic invoked on these SBBs involves creating access-technology agnostic request events and sending these requests to access-agnostic SBBs. The access-agnostic SBBs handle the request and send back responses to the requesting SBBs which in turn invoke response creation methods on the RA to which it is bound (this is achieved using a Custom RASbb Interface Java Interface class as defined by the JAIN SLEE specification).

4.6 Future Work

Development is ongoing on the Mobi4D platform. The short term roadmap includes plans to add the following capabilities:

- Location-based services, requiring the development of a location-lookup-resource-RA to access secure external services that provide cell-based, estimated geo-location for mobile phone users, as well as base service components for processing and extracting geo-spatial information.
- Multimedia message services (MMS), requiring an MMS-protocol-RA to send and receive multipart messages containing multimedia.
- Full integration into the MXit mobile communications and IM system.
- Automated Speech Recognition (ASR), Language Detection and Speaker Verification, and other human language technology capabilities.

In addition, a number of demonstrators and applications may be developed, including:

- A demonstrator application showing the potential of Mobi4D in MHealth.
- A demonstrator application showing the potential of Mobi4D for enhanced access to educational resources.

A converged demonstrator for a call centre application, combining all the capabilities of the platform.

5 Conclusion

Mobi4d presents numerous opportunities for developing mobile services or for adding mobility to existing services. As this paper describes, socio-economic developmental initiatives can also benefit from the use of Mobi4D as a value-adding mobile services delivery platform. The growing penetration rate of mobile phones in developing countries strengthen the case for adopting platforms such as Mobi4D, as it lowers the technological barriers to serving that rapidly growing

user-base, and allows service developers to focus on developing good services, without worrying about the complexity of the underlying protocols. Through technologies such as text messaging (SMS) and USSD, which are supported even by the most basic mobile phones, access to information can be significantly improved for disadvantaged individuals, particularly in developing and under-resourced areas. To enable such improved information access, value added services such as those demonstrated by the Mobi4D capabilities are key.

Acknowledgments The members of the Mobi4D project are duly acknowledged, including the software developers, researchers and project management team, all of whom play a major role in the success of the platform.

References

1. Deruelle J (2008) JSLEE and SIP-servlets interoperability with mobicents communication platform. In: Second international conference on next generation mobile applications, services, and technologies, IEEE NGMAST
2. Java Community Process. JSR 22: JAIN service logic execution environment API specification [Online]. Available from. <http://jcp.org/en/jsr/detail?id=22>. Accessed 18 Jan 2011
3. Service Delivery Platforms: the key to service convergence [Online]. Available from. http://www.devoteam.fr/images/File/Livres_Blancs/ServiceDeliveryPlateforms.pdf. Accessed 17 Feb 2011
4. Open Cloud. 2008. Scope of the JAIN SLEE specification.[Online]. Available from. <https://developer.opencloud.com/devportal/display/RD2v0/1.4.2+Scope+of+the+JAIN+SLEE+Specification>. Accessed 18 Jan 2011
5. World Bank (2009) Information and communications for development: extending reach and increasing impact. World Bank, Washington
6. Jboss, Mobicents communications platform, See <http://www.mobicents.org>
7. Heeks R (2008) ICT4D 2.0: the next step in applying ICT for international development. *Computer* 41(6):26–33
8. Maretzke M (2008). Java telecommunication application server technology comparison
9. RFC 4006—diameter credit-control application [Online]. Available from. <http://tools.ietf.org/html/rfc4006>. Accessed 17 Feb 2011
10. JBOSS DROOLS, “The business logic integration platform” [Online]. Available from: <http://www.jboss.org/drools>. Accessed 18 Jan 2011
11. Open Cloud (2007) A SLEE for all seasons: a discussion on JAINSLEE as an execution environment for new revenue generating services across current and future networks. Open Cloud Limited

The Security Management Model for Small Organization in Intelligence All-Things Environment

Hangbae Chang, Jonggu Kang and Youngsub Na

Abstract Since organizations have recognized needs for industrial technique leakage prevention, they tend to construct information security system causing huge consumption of budget, yet many of them are not affordable to organize information security team to operate integrated information security management system with consistent investment and maintenance. It is fact that there only occur instant introductions of certain system. In this study, we designed information security management system for organizations' industrial technology leakage prevention which is differentiated from those of large enterprises based on current status of small and medium-sized organizations' industrial technology leakage. Specifically we analyzed current status and vulnerability of organizations' industrial technique leakage and we designed industrial technique leakage prevention management system for organizations. Then we applied Delphi method to validate appropriateness of study result. We strongly believe that organizations may estimate an appropriate level of investment on information security and develop countermeasures for control by utilizing this study result.

Keywords Information security · Information security management system for small organization · Vulnerability of information security

H. Chang (✉) · J. Kang · Y. Na
Department of Business Administration, Daejin University,
1007 Hogukro, Pocheon-Si, Gyeonggi-Do, Korea
e-mail: hbchang@daejin.ac.kr

J. Kang
e-mail: jgkang@daejin.ac.kr

Y. Na
e-mail: nangsub@daejin.ac.kr

1 Introduction

It seems that ICT paradigms are standing on the brink of a new internet era, 'Internet of Things', which will radically evolve to the world of Intelligent All-Things. The concept of the Intelligent All-Things will entail the connection of real world myriad things and intelligent devices to all kinds of networks. Building the new world, Intelligent All-Things, however, will pose important challenges. Concerns over privacy on the Intelligent All Things Environment are newly emerged and widespread.

As technique-based Small and Medium Business (SMBs) that are usually venture businesses tend to retain world class techniques, the number of industrial technique leakage and the amount of damage of those incidents for SMBs are increasing rapidly than large enterprises. This tendency attributes to increasing interest of Korean and overseas competitors in high technologies that SMBs possess. These damages caused by industrial technique leakage delay the development pace of SMBs in the knowledge-information-based society where a level of retaining technology directly influences enterprises' competitiveness. It also deteriorates the competitiveness of SMBs.

Preliminary studies regarding this tendency generally possess limitations as below. Firstly, technology-based approach was centered and there have existed a lack of study on managerial and environmental factors regarding information security. Secondly, existing studies concerning information security are just introducing research methodology and deal with a necessity of implementing information security. Only some of recent studies tried to investigate information security management system and level evaluation. Thirdly, due to the stagnation of preliminary researches that are basic level as explained previously, there emerges a lack of research for characteristics of SMBs' information security. Different characteristics of information security should be perceived and different counter-measures are needed for SMBs in comparison with large enterprises, due to SMBs' a limitation of resources and workforce for SMBs in comparison with large enterprises which have large scale of fund and have abundant workforce to utilize.

In this study, we expect to provide an adequate level of investment on information security and control tool for SMBs to progress information security by designing information security management system for SMBs industry technology leakage prevention, based on investigation of current status of SMBs' industrial technology leakage.

2 Characteristics of Small Organization's Industrial Leakage

2.1 *Patterns of Industrial Technology Leakage*

The patterns of recent illegal industrial technology leakage which happen frequently are divided into four types. The first one is the industrial technology leakage caused by labor mobility. Some try to attract competitors' employees with

offer of high annual salary and incentives or bribe Korean engineers who are on the overseas business trip for product demonstration or else purpose. Installing a regional branch to headhunt competitors' core workforce could be another way. The second type of information leakage is a transfer of empirical knowledge concerning components and equipment. This case may take place when partners export core component or equipment that were developed with associate development or when partner collects technical information regarding equipment factory and finished products. The third type of industrial technology leakage may arise by technology transaction. When the overseas firm where technology is transferred grants other company a technology without previous warning or in case of contract conclusion of technology license with third-country firm, the third type may take place. Last type of industrial technology leakage is an industrial spy. Some foreign employees hired by Korean enterprises as researcher, technical counselor, and etc. could be directed to thief information concerning industrial technology by foreign governments or companies. This sort of information leakage may be classified as fourth type of information leakage.

2.2 The Current Status of SMBs' Industrial Technology Leakage

SMBs possess higher risk of possibility of industrial technology leakage because SMBs feature that they have a relative importance of core technology compared with large enterprises. According to (SMBA, Small and Middle Business Administration)'s data, the core industrial technologies that SMBs retain have been categorized as manufacturing technology, knowhow, research and development technology/results, industrial property rights, and sale methods. Amongst those core technologies, research and development technology/results account for the biggest part. The methods of industrial technology leakage are duplication/theft and headhunting core workforce. This is a stereo type of leakage that retiree outflow a related confidential information and provide it to headhunting firm. Other types of leakage channel can be identified as e-mail transmission, the person concerned, tour or observation, cooperative research, and joint business. However SMBs' countermeasures for those risk factors are insufficient except for confidentiality agreement and access control.

3 Study on Information Security Management System

3.1 Preliminary Study on Information Security Management System

Information Security Management System (ISMS) is a certain process and activity to actualize 3 factors (confidentiality, integrity, availability) and it is a systematic management system which is including human resources, process, and information

system to protect firms' sensible information safely [1]. To achieve those objectives, ISMS systematically establishes procedure and process to raise a level of reliability and safety of organizations' asset and put in writing them to maintain consistent management and operation.

'BS7799' has been developed to provide universal reference materials as security standard, consisting of document form to managers who realize information security of organization and are responsible for maintenance of the system under the title of 'Information Security Management Working Standard' by UK's domestic major enterprises and UK Department of Commerce. 'BS7799' provides the unitary reference to identify a necessary and appropriate control for the circumstance that firms are facing with and it has been designed to help not only SMBs but also large enterprises apply it to the extensive areas [2]. The design purpose of 'BS7799' is to cultivate reliability amongst business organizations by providing common information security management document. But 'BS7799' is the authentication regarding management system which is regardless of authentication of information security product/system. Furthermore there exists difficulty to apply itself to SMBs because it just supplies rigid level of standard which lacks flexibility and adaptability under the current information security environment.

Korea Internet and Security Agency has developed information security management system which enables comprehensive application to various environments, based on managerial method in perspectives of organization or environment rather than terms of information technology, referring 'BS7799'. This management system consists of four parts, which encompass 13 specific control areas. The four parts are information security management (strategic policy, risk analysis, security plan, materialization of security, perception education and work process of security audit), information security industry (related products and original technology of information security), information security technology (authentication, legal, publicity, standardization), infrastructure for information security (accident response, encryption and decryption) [3]. This management system of Korea Internet and Security Agency is suggesting areas which need control such as information security management, information security industry, information security technology, yet there lack practical application cases owing to specific application methodology. It also possesses a possibility of limitation of certain areas' excessive appropriation, due to an extensive assessment.

4 Design of Information Security Management System

4.1 Conceptual Understanding for Design of Information Security Management System

The information security system should be designed based on general application architecture because strategy and level of the system has to be designed in accordance with informatization level of SMBs. For these processes, we have

Table 1 Vulnerability and countermeasure for industrial technology leakage

Vulnerabilities of industrial technology leakage	Industrial technology leakage prevention plan
Insufficient of industrial technology leakage prevention policy and procedure, and etc.	Industrial technology security policy
Lack of recognition of possibility of reducing industrial technology leakage	Uplift of recognition of Industrial technology security
Lack of capability of preventing industrial technology leakage	Education and training about industrial technology security
Defenseless access to industrial technology Information	Industrial technology processing procedure
Lack of investigation and grade classification of industrial technology	Systematization of Industrial technology information
Consciousness of insider’s dissatisfaction with organization (promotion, salary, relocate, and etc.)	Industrial technology task processor management
Lack of control on information leakage via various channel of information circulation	Industrial technology security system
Lack of industrial technology share of access authority and modification management	Obligation of industrial technology compliance
Lack of insuring accountability on application program	Industrial technology disclosure security accident response

conducted the study on evaluation of the level of informatization and preceding research regarding information security management system in this study. In the next phase, we have organized components for basic informatization structure and identification of informatization assets according to the preceding studies analyzed. Then we have defined coverage area of information security management system for SMBs and ranges of information security to protect critical assets and components of informatization. The specific elements of information security management system were organized by eliminating parts that are not appropriate for characteristics of information security of SMBs. The suitability of specific elements was deduced by repetitive survey of professionals and referring to preceding studies.

4.2 Design of SMBs’ Information Security Management System to Prevent Industrial Technology

To design the management system to prevent SMBs’ industrial technology leakage, vulnerabilities were deduced according to analysis results of survey differentiated with general information security and a solution was discussed by professionals (3 scholars, 3 practitioners) with Delphi methods. Delphi method is that the repetitive process of taking advice for statistical analysis from professionals. This method provides professionals with chances to modify their opinions

and to share others' opinion. Currently this method is prevailed in the field of technology forecasting research. It also gives a chance to guarantee reliability by participation of professional group. Table 1 describes the vulnerability and countermeasure for industrial technology leakage.

5 Conclusion

Although Korean SMBs allocate huge budget to information security to construct security systems due to recognition of necessity of industrial technology leakage prevention, there is only a single shot of implementation of the partial certain system. They are not affordable to organize special task team handling comprehensive information security management system with consistency. The constructions of these simple types of information security system cause only single event of investment, when a novel vulnerability emerges. To achieve an objective of investment on information security efficiently and effectively, the organizations' propulsion of information security should be progressed in accordance with the evaluation model of information security level, which manages the level of recognition of information security, the level of information security system construction, and the possibility of application of information security technology comprehensively, in the perspective of managerial level.

We have designed the information security management system for SMBs to prevent an industrial technology leakage, which is differentiated from those of large enterprises, based on survey results of the SMBs' current state of industrial technology leakage in this study. We have analyzed and organized cases regarding current state and vulnerabilities of SMBs' industrial technology leakage, and we designed SMBs' industrial technology leakage prevention management system by applying Dephi method. The validity of designed contents has been verified by applying literature studies to verification process to minimize industrial technology leakages. As a result, three management system areas (support capability, support environment, infrastructure) were developed, and five items of management system (education and training, managerial security, human resources security, physical and technical security), and 22 specific elements of management system (Public Relationship, Professional Education, Policies, Special Task Team, Business Process, Security Audit, Incidents Handling Procedure, Management of Change in Human Resources, Reward System, Management of Restricted Area, Processing Equipment Management, and Management of Retaining Industrial Technology, Access Control System, Alarm Monitoring System, CCTV System, Mail and Messenger Security, Document Security, DB Security, Network Access Control, Content Monitoring, and Filtering, Digital Forensic for protection of industrial technology) were designed.

References

1. Weill P, Vitale M (2002) What IT infrastructure capabilities are needed to implement e-business models? *MIS Q Executive* 1(1):17–34
2. BSI (1999) BS 7799 Part1: information security management—code of practice for information security management
3. Doukidis GI, Lybereas P, Galliers RD (1996) Information systems planning in small business: a stages of growth analysis. *J Syst Softw Arch* 33
4. Eloff MM, von Solms SH (2000) Information security management: an approach to combine process certification and product evaluation. *Comput Secur* 19
5. NIST Technology Administration (1998) An introduction to computer security: the NIST handbook. NIST, USA
6. ISACA (2001) Information security governance, guidance for boards of directors and executive management. IT Governance Institute
7. Levy M, Powell P (1998) SME flexibility and the role of information systems. *Small Bus Econ* 2

Simulation Modeling of TSK Fuzzy Systems for Model Continuity

Hae Young Lee, Jin Myoung Kim, Ingeol Chun,
Won-Tae Kim and Seung-Min Park

Abstract This paper presents an approach to formally model Takagi–Sugeno–Kang (TSK) fuzzy systems without the use of any external components. In order to keep the model continuity, the formal simulation model for a TSK fuzzy system is comprised of three types of reusable sub-models involving primitive operations. Thus, the model can be executed even on limited computational platforms, such as embedded controllers.

Keywords Modeling and simulation · Model continuity · Fuzzy logic · Discrete event system specification · Embedded systems

1 Introduction

Modeling and simulation (M&S) technologies have been widely used in industry to assist in system development [1]. One particular use of these technologies is in the development of embedded controllers since they usually have time constraints [2, 3]. When modelers build simulation models for embedded fuzzy controllers, they typically embed external fuzzy components in their models [4, 5]. These models, however, may not be used throughout all of the design phases since M&S

This work was supported by the IT R&D Program of MKE/KEIT [10035708, “The Development of CPS (Cyber-Physical Systems) Core Technologies for High Confidential Autonomic Control Software”].

H. Y. Lee (✉) · J. M. Kim · I. Chun · W.-T. Kim · S.-M. Park
CPS Research Team, ETRI, Daejeon 305-700, Republic of Korea
e-mail: haelee@ieee.org

environments do not support the use of some external components. Also, the use of external components may make the transformation of simulation models difficult or impossible [6]. Therefore, simulation models should not contain any external components to keep their continuity.

Several research efforts [6–8] have been made to build ‘pure’ simulation models for fuzzy controllers. In [7], Jamshidi et al. proposed an approach to model the Mamdani fuzzy systems [9] based on parallel discrete event system specification (P-DEVS) [10]. The modeling approach proposed by Lee and Kim [8] can reduce the complexity of the Mamdani P-DEVS models. The Mamdani model has a great advantage in terms of expression power, though it involves some complex computation. The standard additive model (SAM) fuzzy systems [11] can be built with P-DEVS models based on the approach proposed in [6]. The main advantage of the SAM is computational efficiency since most parameters can be precomputed. However, simulation modeling of Takagi–Sugeno–Kang (TSK) fuzzy systems [12, 13] has not been addressed yet. Compared to the Mamdani model, the TSK model can reduce the number of rules, especially for complex and high-dimensional problems.

This paper presents an approach to build simulation models for TSK fuzzy systems based on P-DEVS. A P-DEVS model of a TSK fuzzy system is a coupled model consisting of three types of sub-models: an input membership function model, rule model, and defuzzification model. Since the models are all pure simulation models involving only addition and multiplication, they could be executed even on embedded platforms. Consequently, their continuity can be maintained. Compared to the existing approaches for the modeling of fuzzy systems, the proposed approach can model a TSK fuzzy system with a smaller number of sub-models.

2 Background

In this section, we briefly describe the backgrounds of TSK fuzzy systems and P-DEVS.

2.1 TSK Fuzzy Systems

In general, a rule in a TSK model has the following form:

$$\begin{aligned} &\text{IF } x_1 \text{ is } A_{i1} \text{ and } x_2 \text{ is } A_{i2} \text{ and, } \dots, \text{ and } x_k \text{ is } A_{ik} \\ &\text{THEN } y = a_{i0} + a_{i1} \times x_1 + \dots + a_{ik} \times x_k \end{aligned}$$

where x_1, x_2, \dots, x_k are input parameters, $A_{i1}, A_{i2}, \dots, A_{ik}$ are the membership functions of i th rule, $a_{i0}, a_{i1}, \dots, a_{ik}$ are real-valued parameters, and y is the output parameter. The total output, y , of the model is given by Eq. (1), where α_i is the matching degree of the i -th rule.

$$y = \frac{\sum_{i=1}^j \alpha_i (a_{i0} + a_{i1}x_1 + \dots + a_{ik}x_k)}{\sum_{i=1}^j \alpha_i} \quad (1)$$

The great advantage of the TSK model is its representative power. Moreover, due to the explicit functional representation form, it is convenient to identify its parameters using learning algorithms [14].

2.2 Parallel DEVS

The basic formalism of a P-DEVS model is:

$$M = \langle X, Y, S, \delta_{ext}, \delta_{int}, \delta_{con}, \lambda, ta \rangle,$$

where

X is the set of input events,

Y is the set of output events,

S is the set of sequential states,

$\delta_{ext}: Q \times X^b \rightarrow S$ is the external transition function,

where $Q = \{(s, e) \mid s \in S, 0 < e < ta(s)\}$, e is the elapsed time since the last state transition, and X^b is a set of bags over the elements in X ,

$\delta_{int}: S \rightarrow S$ is the internal transition function,

$\delta_{con}: Q \times X^b \rightarrow S$ is the confluent transition function, subject to $\delta_{con}(s, \emptyset) = \delta_{int}(s)$,

$\lambda: S \rightarrow Y^b$ is the output function,

ta is the time advanced function.

3 P-DEVS Modeling of TSK Fuzzy Systems

In the proposed approach, a TSK fuzzy system containing i input membership functions and j rules, with k inputs and a single output is represented as a P-DEVS coupled model with k input ports and a single output port. The coupled model contains $i + j + 1$ P-DEVS atomic models: i input membership function models (IMs), j rule models (RMs) and a single defuzzification model (DM). Figure 1 shows the P-DEVS model of a fuzzy system containing four input membership functions and four rules with two inputs and a single output (i.e., $i = 4, j = 4, k = 2$).

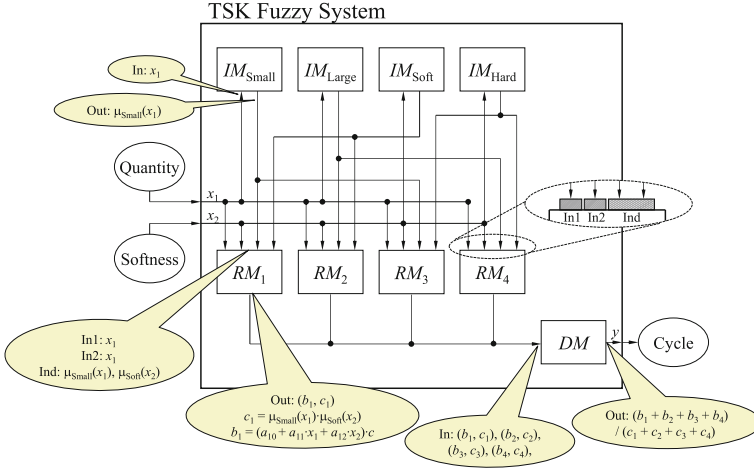


Fig. 1 A model structure for a TSK fuzzy system

3.1 Input Membership Function Models (IMs)

Each input membership function of the fuzzy system is represented as an IM M that is defined as:

$$M = \langle X_M, Y_M, S, \delta_{ext}, \delta_{int}, \delta_{con}, \lambda, ta \rangle,$$

where

$$\begin{aligned} InPorts &= \{ \text{"In"} \}, X_{In} = \mathfrak{R}, \\ OutPorts &= \{ \text{"Out"} \}, Y_{Out} = [0, 1], \\ X_M &= \{ (p, v) \mid p \in InPorts, v \in X_p \}, \\ Y_M &= \{ (p, v) \mid p \in OutPorts, v \in Y_p \}, \\ S &= \{ \text{"passive"}, \text{"active"} \} \times \mathfrak{R}, \\ \delta_{ext} (\text{"passive"}, d, e, (\text{"In"}, x)) &= (\text{"active"}, \mu(x)), \\ \delta_{int} (\text{"active"}, d) &= (\text{"passive"}, d), \\ \delta_{con} (s, ta(s), x) &= \delta_{ext} (\delta_{int}(s), 0, x) \\ \lambda (\text{"active"}, d) &= (\text{"Out"}, d), \\ ta(phase, d) &= 0 \quad \text{if } phase = \text{"active"}; \\ &\infty \quad \text{otherwise.} \end{aligned}$$

Every IM produces a matching degree of the corresponding membership function for each input value. It initially starts with its state = ("passive", d), where d is an arbitrary real value. When the IM for an input membership function I receives a real value x as an input, it transitions its state to ("active", $\mu_I(x)$). Immediately, the IM generates $\mu_I(x)$ as its output and transitions to the passive state.

Once an IM for a membership function type has been implemented, it can be easily reused just by setting the parameters of the membership functions in the same type. Even if any IM for a certain type does not exist, it can be implemented just by redefining δ_{ext} of the existing one. IMs are independent from fuzzy inference models; TSK, SAM, and Mamdani use the same IMs in the approach.

3.2 Rule Models (RMs)

Each if-then rule of the fuzzy system corresponds to an RM. An RM is defined as:

$$M = \langle X_M, Y_M S, \delta_{ext}, \delta_{int}, \delta_{con}, \lambda, ta \rangle,$$

where

$$\begin{aligned} InPorts &= \{ \text{"In1"}, \dots, \text{"Ink"}, \text{"Ind"} \}, X_{In1} = \dots = X_{Ink} = X_{Ind} = \mathfrak{R}, \\ OutPorts &= \{ \text{"Out"} \}, Y_{Out} = \mathfrak{R}^2, \\ X_M &= \{ (p, v) \mid p \in InPorts, v \in X_p \}, \\ Y_M &= \{ (p, v) \mid p \in OutPorts, v \in Y_p \}, \\ S &= \{ \text{"passive"}, \text{"active"} \} \times \mathfrak{R}^{k+3}, \\ \delta_{ext} &(\text{"passive"}, a_0, \dots, a_k, b, c, e, ((\text{"In1"}, x_1), \dots, (\text{"Ink"}, x_k), (\text{"Ind"}, d_1), \dots, \\ &(\text{"Ind"}, d_k)) \\ &= (\text{"active"}, a_0, \dots, a_k, a_0 + \dots + a_k \cdot x_k, \min(d_1, \dots, d_k)) \\ \text{or} \\ &= (\text{"active"}, a_0, \dots, a_k, a_0 + \dots + a_k \cdot x_k, d_1 \times \dots \times d_k), \\ \delta_{int} &(\text{"active"}, a_0, \dots, a_k, b, c) = (\text{"passive"}, a_0, \dots, a_k, b, c), \\ \delta_{con} &(s, ta(s), x) = \delta_{ext}(\delta_{int}(s), 0, x) \\ \lambda &(\text{"active"}, a_0, \dots, a_k, b, c) = (\text{"Out"}, (b \times c, c)), \\ ta(phase, a_0, \dots, a_k, b, c) &= 0 \quad \text{if } phase = \text{"active"}; \\ &\infty \quad \text{otherwise.} \end{aligned}$$

Each RM produces a conclusion of the corresponding rule, based on input values: all input values of the fuzzy system and the membership degrees from the associated IMs. It has k input ports, "In1", ..., "Ink", used to receive the k input values, x_1, \dots, x_k , of the fuzzy system; an additional input port, "Ind", used for k membership degrees, d_1, \dots, d_k , from the associated IMs; and a single output port, "Out." The RM corresponding to rule R starts with the initial state = ("passive", a_0, \dots, a_k, b, c), where a_0, \dots, a_k are the constant values defined in the consequent part (then-part) of R , and b and c are arbitrary real values. When the RM receives x_1, \dots, x_k , through the ports "In1", ..., "Ink", respectively, together with d_1, \dots, d_k through the port "Ind", it stores $b = a_0 + \dots + a_k \cdot x_k$ and $c = \min(d_1, \dots, d_k)$ or $(d_1 \times \dots \times d_k)$. Finally, the RM outputs $(b \times c, c)$ via the output port and transitions to the passive state.

The RM can be reused repeatedly once it has been implemented. The reuse can be done simply through the creation of RM instances and assigning $k + 1$ parameters a_0, \dots, a_k , of each instance.

3.3 Defuzzification Model (DM)

A DM is an application-independent atomic-model that generates the outputs of a fuzzy system. It is formally defined as:

$$M = \langle X_M, Y_M, S, \delta_{ext}, \delta_{int}, \delta_{con}, \lambda, ta \rangle$$

where

$$\begin{aligned} InPorts &= \{ \text{“In”} \}, X_{In} = \mathfrak{R}^2, \\ OutPorts &= \{ \text{“Out”} \}, Y_{Out} = \mathfrak{R}, \\ X_M &= \{ (p, v) \mid p \in InPorts, v \in X_p \}, \\ Y_M &= \{ (p, v) \mid p \in OutPorts, v \in Y_p \}, \\ S &= \{ \text{“passive”}, \text{“active”} \} \times \mathfrak{R}, \\ \delta_{ext} (phase, y, e, (b_1, c_1), \dots, (b_j, c_j)) &= (\text{“active”}, (b_1 + \dots + b_j) / \\ & (c_1 + \dots + c_j)), \\ \delta_{int} (\text{“active”}, y) &= (\text{“passive”}, y), \\ \lambda (\text{“active”}, y) &= (\text{“Out”}, y), \\ ta(phase, y) &= 0 \quad \text{if } phase = \text{“active”}; \\ &\infty \quad \text{otherwise.} \end{aligned}$$

The DM produces a final conclusion of the fuzzy system based on the collection of conclusions of the rules. It starts with the passive state = (“passive”, y), where y is an arbitrary real value. When the DM receives $(b_1, c_1), \dots, (b_j, c_j)$ from all RMs, it transitions its state to (“active”, $(b_1 + \dots + b_j) / (c_1 + \dots + c_j)$). Then, the DM outputs y and transitions its state back to the passive state.

Since any implementation of the DM is application-independent, it is reused for every TSK fuzzy system. Also, it is identical to the DM of SAM described in [6].

3.4 Models Couplings

The coupled model has i input ports and a single output port. Each of the input ports is connected with the associated input ports (e.g., input port “In1” for input x_1 , “In2” for x_2, \dots) of all RMs. The port is also coupled with the input ports of the associated IMs. Consider the following fuzzy if-then rules of a TSK fuzzy system that receives quantity x_1 and softness x_2 :

Table 1 Overhead in four modeling approaches

Complexity	TSK (proposed)	SAM [6]	Mamdani [7]	Mamdani [8]
Sub-models	$i + j + 1$	$i + j + l + 1$	$j \times k + 2j + 2$	$i + j + l + 2$
Communications	$i + 2j \times k + j + 1$	$i + j \times k + j + k + 1$	$2j \times k + 2j + 2$	$i + j \times k + j + l + 2$
Inference	Multiplication + addition	Multiplication	Finding a minimum + clipping of or scaling a MF	
Combining	Addition		Merging MFs	
Defuzzification	Multiplication		Finding the mean of the maximum or the centroid of an area	

Rule 2 : IF x_1 is Large and x_2 is Soft

$$\text{THEN } y = 1 + 2 \cdot x_1 + 2 \cdot x_2$$

Rule 3 : IF x_1 is Small and x_2 is Hard

$$\text{THEN } y = 1 + x_1 + 2 \cdot x_2$$

In the above example, the port that receives x_1 (quantity value) would be connected with the input ports of IM_{Small} and IM_{Large} . Note that each IM has a single input port “In”. The output port “Out” of each IM is coupled with input port “Ind” of each associated RMs. In the example, “Out” of IM_{Small} would be connected with “Ind” of RM_3 , which have a linguistic variable ‘Small’ in the if-part. The output port of every RM is coupled with the input port of the DM. And the output port of the DM is connected with that of the coupled model.

3.5 Overhead Analysis

Table 1 shows an overhead analysis for the approach and the three existing approaches [6–8]. Each fuzzy system consists of i input membership functions, j rules, and l output membership functions, with k inputs and a single output. In the approach, a coupled model for a fuzzy system contains $i + j + 1$ atomic models, while a higher number of sub-models is required to build a coupled model for a fuzzy system in other approaches. The complex couplings among the sub-models in the approach make the communications overhead increase. However, the overhead of the approach is still smaller than that of [7]. Moreover, TSK can describe a highly nonlinear system using a small number of rules [14]. That is, j of TSK could be much smaller than that of SAM or Mamdani. While Mamdani fuzzy systems are widely used, they usually involve complex operations, such as the clipping and merging of membership functions and finding their centroids. Such

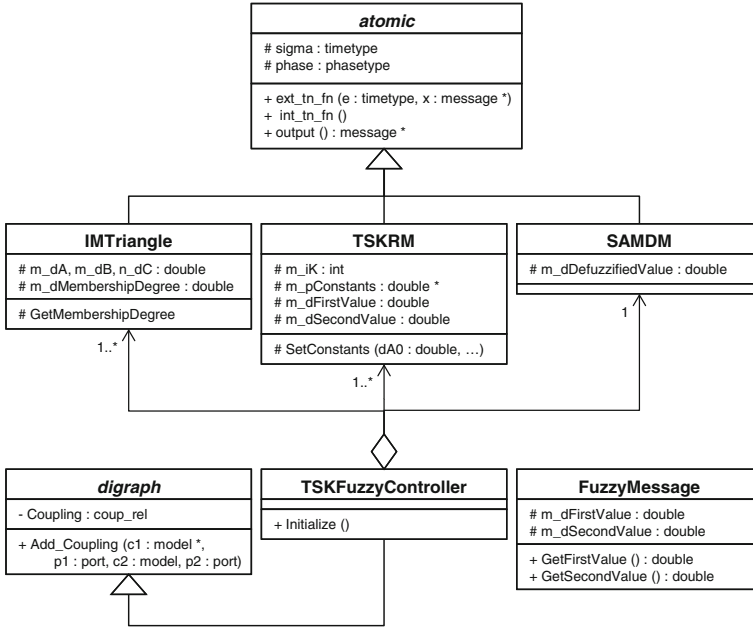


Fig. 2 Simplified UML diagrams of TSK simulation models

complex operations might be too heavy on resource-constrained systems. Similar to [6], the approach can model TSK fuzzy systems, which involve only primitive operations. Thus, the approach will be suitable for the M&S-based engineering of embedded software systems.

4 Implementation Status

The P-DEVS models for TSK systems described in Sect. 3 were implemented in C++ for our simulation environment, the DEVS object C++ (DOC) environment. The IM for the triangular membership function type was implemented as an IMTraingle class. The RM and DM were implemented as TSKRM and SAMDM classes, respectively. As shown in Fig. 2, these classes inherit the atomic class of DOC class, which corresponds to the basic model of P-DEVS. The essential member functions of atomic are ext_tn_fn (the implementation of δ_{ext}), int_tn_fn (δ_{int}), and $output$ (λ). By overriding these functions, the behavior of a subclass is defined. The FuzzyMessage class is used for internal communications between the atomic models of fuzzy systems. In order to facilitate the modeling process, we have also developed a prototype of a visual modeling tool. The simulation modeling of fuzzy systems can be done with ease using the tool. However, they

can also be manually constructed without the use of the tool, thanks to the hierarchical and modular model-composition provided by the P-DEVS environments.

5 Conclusions and Future Work

In this paper, we presented an approach for representing P-DEVS models of TSK fuzzy systems without the use of any external components. Exclusion of external components from simulation models would improve the continuity of the models so that the user can efficiently manage software complexity and maintain consistency throughout the design phase. A P-DEVS model of a fuzzy system is comprised of easy-to-reuse atomic models: IMs, RMs, and a DM. Since each atomic model involves primitive operations, such as addition or multiplication, the model works on target platforms and can be smoothly transformed into other forms of models or languages. A coupled model for a TSK fuzzy system requires a smaller number of sub-models, compared to that of a Mamdani or SAM fuzzy system. Thus, it will be more compatible with embedded platforms. We have implemented P-DEVS models on DOC, for fuzzy systems including TSK, SAM, and Mamdani. To facilitate the modeling of fuzzy systems, a GUI-based modeling tool prototype was developed. We will implement the models for other DEVS environments, such as eCD++ [2].

References

1. Hu X (2004) A simulation-based software development methodology for distributed real-time systems. Doctoral dissertation, The University of Arizona
2. Moallemi M, Gutierrez-Alcaraz JM, Wainer G (2008) ECD++ A DEVS based real-time simulator for embedded systems. In: Proceedings of the spring simulation multiconference, article no. 12
3. Park J, Yoo J (2010) Hardware-aware rate monotonic scheduling algorithm for embedded multimedia systems. ETRI J 32:657–664
4. Garcia AM, Baumgartner B, Schreiber U, Krane M, Knoll A, Bauernschmitt R (2009) Automedic: fuzzy control development platform for a mobile heart-lung machine. IFMBE Proc 25:685–688
5. Muruganandam M, Madheswaran M (2009) Modeling and simulation of modified fuzzy logic controller for various types of DC motor drives. In: Proceedings of international conference on control, automation, communication and energy conservation, pp 1–6
6. Lee HY, Park SM, Cho TH (2010) Simulation modeling of SAM fuzzy logic controllers. IEICE Trans Inf Syst E93-D:1984–1986
7. Jamshidi M, Sheikh-Bahaei S, Kitzinger J, Sridhar P, Beatty S, Xia S, Wang Y, Song T, Dole U, Lie J (2003) V-LAB-A distributed intelligent discrete-event environment for autonomous agents simulation. Intell Autom Soft Comput 9:181–214
8. Lee HY, Kim HJ (2009) Reducing the complexity of DEVS-based mamdani models for enhancing privacy. Proceedings of international symposium on advanced intelligent systems, pp 281–283

9. Mamdani EH (1974) Application of fuzzy algorithms for control of simple dynamic plant. *IEEE Proc* 121:1585–1588
10. Zeigler BP, Kim TG, Praehofer H (2000) *Theory of modeling and simulation*, 2nd edn. Academic Press, New York
11. Kosko B (1997) *Fuzzy engineering*. Prentice Hall, Upper Saddle River
12. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its application to modeling and control. *IEEE Trans Syst Man Cybern* 15:116–132
13. Sugeno M, Kang KT (1988) Structure identification of fuzzy model. *Fuzzy Sets Syst* 28: 15–33
14. Yen J, Langari R (1999) *Fuzzy logic: intelligence control and information*. Prentice Hall, Englewood Cliffs

A New Method of Clustering Search Results Using Frequent Itemsets with Graph Structures

I-Fang Su, Yu-Chi Chung, Chiang Lee and Xuanyou Lin

Abstract The representation of search results from the World Wide Web has received considerable attention in the database research community. Systems have been proposed for clustering search results into meaningful semantic categories for presentation to the end user. This paper presents a novel clustering algorithm, which is based on the concept of frequent itemsets mining over a graph structure, to efficiently generate search result clusters. The performance study reveals that the algorithm was highly efficient and significantly outperformed previous approaches in clustering search results.

Keywords Web clustering engine · Frequent itemsets mining · Hash table · Graph structure

This work is supported by National Science Council of Taiwan (R.O.C.) under Grants NSC99-2218-E-268-001, NSC99-2221-E-006-133, and NSC100-2221-E-309-011.

I.-F. Su (✉)
Department of Information Management, Fotech,
831 Kaohsiung, Taiwan
e-mail: ifangsu@center.fotech.edu.tw

Y.-C. Chung
Department of CSIE, CJCUC, 711 Tainan, Taiwan
e-mail: justim@mail.cjcu.edu.tw

C. Lee · X. Lin
Department of CSIE, NCKU, 701 Tainan, Taiwan
e-mail: leec@mail.ncku.edu.tw

1 Introduction

Search engines are widely used tools for obtaining information from the Web. They usually return a list of results that are ranked in order of relevance to a user's query. The user must start reviewing the list at the top and follow it down checking one result at a time, until the information is found. This representation of search results is good for searching simple tasks, such as finding the home page of a company. However, users often issue semantically incomplete or ambiguous queries. A list representation of the searching results is less effective for these kinds of queries. For example as shown in Fig. 1, a user is trying to find a picture of an apple from the Web. The results are presented in the form of a ranked list; however, since the keyword "apple" may represent a fruit, a newspaper, or a computer manufacturer, the search results are mixed and have different meanings, and the information in the is disorganized and redundant. Generally, users check only the links of the first few pages. If a satisfactory answer is not found in the first few pages, the user will re-issue another keyword for a second round of searching. In this case, the user may need to query numerous times, and read a large amount of irrelevant snippets, to find the results they need.

olve the information disorganization and redundancy in search engine results, a different approach was proposed to group search results into a hierarchy of labeled clusters, in which each cluster contains web pages that are semantically related to each other. By analyzing the hierarchy of the results, the user can have a global view of the different semantic areas of his query, and can explore the areas in which he is interested. Documents can therefore be accessed in logarithmic rather than linear time. For example, in Fig. 2, the results from the above query are categorized into several clusters. Each cluster has a shortcut to the results that have the same related meaning. It allows users to explore their topics of interest without checking numerous irrelevant search results. In this example, users can directly choose the cluster "pictures" and access the results from this cluster. It is obviously more efficient to return results in a clustering engine than in a lengthy ranked list.

Although a clustering engine has many advantages, it also has many challenges. The first challenge is that the clustering engine has to classify input contents that are usually short into several meaningful topics. The next challenge is that the clustering engine has to find readable labels for each topic in an acceptable response time (usually just a few seconds). While classifying the clusters, the engine has to face a third problem of being unable to predetermine the number and size of the clusters, as they may vary with the queries and be inferred from a variable number of search results. As the same result can often be assigned to multiple topics, it is preferable to cater for overlapping clusters. The existing static classification systems, such as DMOZ¹ and Yahoo Directory,² are static web page classifications that generate clusters manually. The categories and labels of manual

¹ <http://www.dmoz.org>

² <http://dir.yahoo.com>

Fig. 1 The search engine results for query “apple”

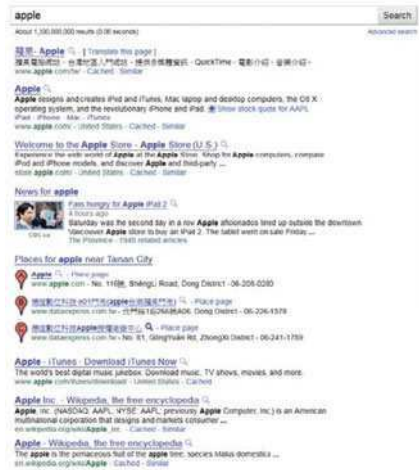


Fig. 2 A clustering engine results for query “apple”



systems are usually readable and easy to understand, since they are generated artificially. In addition, these systems are fast, because there is little dynamic analysis of the significance of the user-issued query. However, manual systems are not efficient for clustering engines, due to the high update rate of web pages. Figure 3 illustrates the number of web pages that were added to the Google search engine³ index in 2010. It shows that the number of web pages increases dramatically over time. The manual categorizing method requires a lot of effort create and maintain clusters. Hence, an automatical clustering system that analyzes contents dynamically is highly demanded.

³ <http://www.google.com>

Fig. 3 The size of world wild web pages which was estimated by *worldwidewebsize.com*



Systems [2, 4, 5, 8] that automatically perform clustering of web search results have become popular in recent years. They are mainly focused on generating expressive clustering labels. However, constructing clustering labels is a time consuming process for these systems. Therefore, this paper presented a novel web clustering engine, named Clustering Algorithm by Frequent Itemsets with Graph structure (CAFIG), which was designed to improve the efficiency of web clustering. This study compared CAFIG to other state-of-the-art search results clustering algorithms using k SSL and F_B measures to evaluate the accuracy and efficiency of clustering construction. The performance study indicated that CAFIG had a significantly lower processing time and a higher clustering accuracy than the previous methods.

The rest of this paper is organized as follows. The related works are discussed in Sect. 2. The details of the proposed method are described in Sects. 3 and 4 discusses the simulation results. Conclusions are finally drawn in Sect. 5.

2 Related Work

As mentioned in the previous section, search result categorizing can be divided into two approaches: web directories and search result clustering. In this paper, we focused on search result clustering algorithms. Hence, a survey of web directories was omitted. In Sect. 2.1, a number of state-of-the-art algorithms are briefly described. Then, a graph-based algorithm is introduced in Sect. 2.2.

2.1 State-of-the-Art SRC Algorithms

We first introduce the three state-of-the-art SRC algorithms, such as KeySRC [2], Lingo [8], and OPTIMSRC [4], in this section, and discuss the problems of these three algorithms.

The main idea of KeySRC is to extract and analyze frequently used phrases from the search results using a tree structure, and then cluster these selected phrases to the end users. The selected phrases are called *keyphrases*. The steps of extracting and analyzing keyphrases are described as follows. First, phrases are constructed by processing the search results using a generalized suffix tree (GST). Then, keyphrases are extracted from the GST using the following rules: Each phrase must (a) not be equal to the query, (b) be contained in at least two search results, (c) contain no more than four words, and (d) contain only adjectives and nouns. As to criterion (d), the word forms are implemented by checking the dictionary. Each keyphrase is then represented as a document vector, and a hierarchical agglomerative clustering of the keyphrase vectors is performed. Finally, a cluster label is determined by the keyphrases of each cluster.

The Lingo [8] method discovers the clusters of the search results by performing singular value decomposition (SVD) in a vector space model (VSM). Frequently occurring phrases are first extracted from the input documents. Next, the existing latent structure of diverse topics in the documents are found using SVD. Finally, group descriptions are matched with the extracted topics and relevant documents are assigned.

Every clustering algorithm discovers clusters from web search results, but OPTIMSRC [4] proposes a different method of merging the outputs of multiple search result clustering algorithms. OPTIMSRC produces clustering by analyzing the outputs.

From the above discussion, problems can be seen in these algorithms. For example, in KeySRC, the accuracy of the clustering is highly related to the accuracy of the dictionary that has to be checked when finding the keyphrase. Since the phrases that are submitted as queries to the system change in popularity over time, they may also have different meanings and word forms over time. If the dictionary is not updated periodically to get the correct word forms of the phrases, the system could get the wrong forms and extract the wrong keyphrases when performing clustering algorithms. The problem with Lingo is that SVD decomposition is computationally quite demanding. A large number of results and phrases will highly affect the efficiency of performing clustering. Although OPTIMSRC has a superior performance compared to Lingo and KeySRC, the response time for OPTIMSRC is significantly higher than that for Lingo and KeySRC. This is because before OPTIMSRC analyzes the outputs of the multiple search result clustering algorithms, it must wait until these algorithms are performed. Thus, OPTIMSRC requires a much longer response time than other clustering systems.

2.2 Graph-Based Algorithm

Most of the search results clustering algorithms represent the clusters using a list or a tree structure. WhatsOnWeb [5] proposes a graph-based user interface for clustering search results. Users may explore the different categories and

relationships of the search results from a graph visualization technique. The vertex in a graph visualization represents a snippet of the search results, and an edge represents the semantic connection between two vertices. The edges are built by sequentially searching the connection between any two vertices. The graph visualization is completed after WhatsOnWeb merges the vertices that have similar semantics into a cluster and refines the edges of the vertices.

Since WhatsOnWeb sequentially searches the semantic connections between vertices, the classification accuracy of this algorithm is significantly high. However, comparing the semantic connection between any two vertices is time consuming and requires a high response time to process a query. Therefore, it is not suitable for processing queries in a real system.

3 Clustering Algorithm by Using Frequent Itemsets with Graph Structure

Our algorithm contains the following phases. The first one is to preprocess the search results. Next, we adopt a mining technique to retrieve the frequent itemsets from the preprocessed results. Then, a graph structure is applied for clustering the semantic itemsets as a group. Finally, the label of each cluster is induced.

3.1 Preprocessing

We assume the input of our algorithm are derived from the Google search engine. In our work, we adopt the mechanism of [7] to perform the operations of stemming, stop words removing, and tokenization. For example in Table 1, the snippets are the search results of the query "zombie". After the preprocessing phase, the snippets are stemming and tokenized into many items, and the stop words of the snippets are removed. The results in Table 2 shows that the items are preprocessed from the previous snippets.

3.2 Frequent Itemsets Mining

The main goal of this phase is to find the frequent itemsets from the preprocessing results. Intuitively, if an item is used repetitively, this item should be a frequent item and it usually is a subject-related keyword. Those itemsets should be retrieved to represent a category. Since there is a long stream of research on frequent itemsets mining, a large number of mining rules [1, 6, 10] have been developed for this purpose. *FP-growth* [6] is the most efficient technique for mining frequent itemsets among these rules. Hence, we adopt the *FP-growth* to mine the frequent itemsets from the preprocessing results.

Table 1 An example of search results for a query “zombie”

Doc. ID	Snippets
D_1	Ziombie PC games is so funny
D_2	Rob zombie: heavy metal singer
D_3	Watch dead films—Walking dead cinema online
D_4	Zombie Wars PC Game
D_5	Rob Zombie Official site
D_6	George A Romero’s dead zombie films
D_7	Are hackers using your PC to spew spam and steal
D_8	Monster Island, a serial horror novel by D. Wellington
D_9	Spam, phish, harass, on the sly.
D_{10}	Horror novel: elements, werewolves, vampires, zombies

Table 2 Snippets after preprocessing

Doc. ID	Items are extracted from snippets
D_1	Zombie PC games fun
D_2	Rob zombie heavy metal singer
D_3	Watch dead film walk dead cinema online
D_4	Zombie wars PC game
D_5	Rob zombie official site
D_6	George romero dead zombie films
D_7	Hacker PC spew spam steal
D_8	Monster Island serial horror novel Wellington
D_9	Spam phish harass sly
D_{10}	Horror novel element werowolvo vampire zombie

FP-growth retrieves the items from the preprocessing results and orders these items in the support descending order. If the number of an item appeared in documents is larger than the minimum support of *FP-growth*, the item is one of the frequent itemsets. Note that the minimum support is a small integer and is set as a system parameter. The minimum support here has been experimentally set to $m \geq 0.1$ where m is the number of search results. We take the results of Table 2 as an example. If the minimum support of *FP-growth* is 2 which means that each retrieved item has been written in at least two documents, the frequent itemsets of Table 2 are shown in Table 3.

3.3 Graph Building

After the frequent itemsets are mined, we use the graph structure to group the similar items as a category. In this section, we first define the vertex of the graph. Then, introduce how to find the relationship between two vertices and determine the level of each vertex in a hierarchical representation.

We take the frequent itemsets as the input and each frequent item is viewed as a vertex, as well as each vertex initially represents a cluster. Take Table 3 as an

Table 3 Frequent itemsets of minimum support of 2

Vertex ID	Frequent Itemsets	Doc. ID
V_1	Zombie	$D_1, D_2, D_4, D_5, D_6, D_{10}$
V_2	PC	D_1, D_4, D_7
V_3	Game	D_1, D_4
V_4	Rob	D_2, D_5
V_5	Dead	D_3, D_6
V_6	Film	D_3, D_6
V_7	PC zombie	D_1, D_4
V_8	PC game	D_1, D_4
V_9	Game zombie	D_1, D_4
V_{10}	Rob zombie	D_2, D_5
V_{11}	Dead film	D_3, D_6
V_{12}	Game PC zombie	D_1, D_4
V_{13}	Spam	D_7, D_9
V_{14}	Horror	D_8, D_{10}
V_{15}	Novel	D_8, D_{10}
V_{16}	Horror novel	D_8, D_{10}

Table 4 Hash table of the vertices

Key	Vertex ID
Zombie	$V_1, V_7, V_9, V_{10}, V_{12}$
PC	V_2, V_7, V_8, V_{12}
Game	V_3, V_8, V_9, V_{12}
Rob	V_4, V_{10}
Dead	V_5, V_{11}
Film	V_6, V_{11}
Spam	V_{13}
Horror	V_{14}, V_{16}
Novel	V_{15}, V_{16}

example, there are sixteen vertices in the graph. We use an edge to represent the connection between two vertices. If two vertices, V_i and V_j , share the same item, there is an edge among V_i and V_j . However, sequentially searching the connection between any two vertices is time consuming. Hence, we design a hash table structure to improve the searching process.

We retrieve the items from the frequent itemsets and set the items as the keys of a hash table. For example in Table 3, the retrieved keys of the hash table are zombie, PC, game, rob, dead, film, spam, horror, and novel. If a vertex contains the item in it, it is hashed to the bucket of the key. For example in Table 4, the vertex V_9 is hashed to the buckets of zombie and game since the frequent itemsets of V_9 is “game zombie”. While the hash table is constructed, we can only check the vertices in the same bucket. For example in Table 4, V_7 only has to check $V_1, V_2, V_8, V_9, V_{10}$, and V_{12} while building the edges of V_7 . Using the

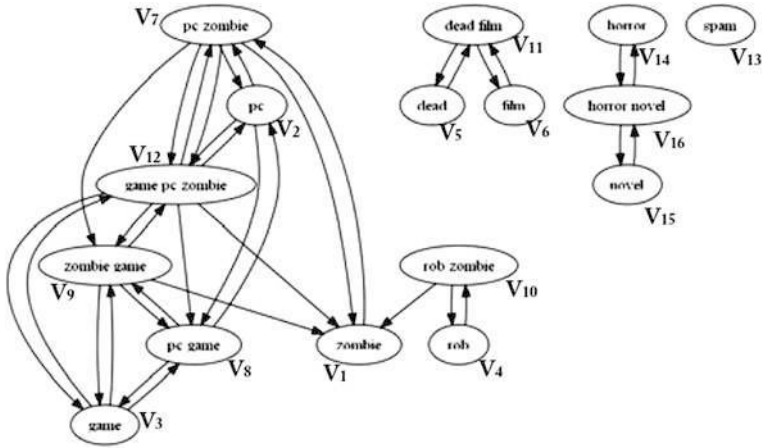
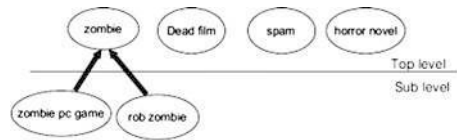


Fig. 4 Graph construction

Fig. 5 The graph after merge



hash table in the search process can highly reduce the computation of connections between vertices from C_2^{16} (=120) to 64 times. The graph is then constructed as shown in Fig. 4.

Next, we further prune the edges that are not highly related to each other. We use *weight* to determine the relationship between two vertices. If two vertices are related to each other, the value of the weight should be high. The weight of two vertices is derived according to Eq. 1. Then we use a threshold hold φ to determine which edge should be prune. In our work, the value of φ is set as a system parameter and it has been experimentally set to 0.4. For example in Fig. 4, the value of *weight* (V_{14}, V_{16}) is higher than 0.4 and the edge of V_{14} to V_{16} is not pruned in this graph.

After the edges are pruned, we merge the semantically related vertices into a same cluster. Figure 5 shows the graph after merging. While the graph is constructed, we give each cluster a human readable label. The label of each cluster is derived from the frequent itemsets of a cluster.

$$weight(V_i, V_j) = \frac{|Iv_i \cap Iv_j|}{|Iv_j|} \tag{1}$$

Table 5 Performance of clustering algorithms on the Ambient test collection expressed as $kSSL$ for several values of k

Algorithms	$k = 1$	$k = 2$	$k = 3$	$k = 4$
CAFIG	14.19	22.67	27.14	31.07
OPTIMSRC	20.56	28.93	34.05	38.94
KeySRC	24.07	32.39	38.19	42.13
Lingo	24.4	30.64	36.57	40.69
WhatsOnWeb	10.54	20.34	25.47	29.12

4 Evaluation

In this section we describe the test collections and metrics used in the experiments and compare our algorithm to keySRC, Lingo, OPTIMSRC and WhatsOnWeb.

Test Collections:

There is no standard test collection for evaluating clustering algorithms. The popular test collections for evaluating clustering algorithms are ODP-239 and AMBIENT [3]. Thus, we use these two collections as our test collections. Due to space limitations, we refer readers to [3] for more details about these two collections.

Clustering Retrieval:

In this section, we evaluate the clustering retrieval effectiveness of our algorithm. We use the same experimental setting as previous experiments. We compare the corresponding $kSSL$ values of our algorithm to the-state-of-art algorithms and WhatsOnWeb. This experiment was limited to the Ambient collection. The performance result, which is shown in Table 5, shows that CAFIG and WhatsOnWeb outperform the other approaches on various k . This is because CAFIG and WhatsOnWeb both precisely retrieve the semantic connections between documents using graph structures. Although WhatsOnWeb performs slightly better than CAFIG, it requires a high response time to process a query. Thus, CAFIG is still a good choice for search results clustering in a real-time system.

Clustering Validation:

We use F_β measure to evaluate how good a clustering method is at recovering known clusters from a gold standard partition. Since many documents in Ambient are not assigned to any category, we only use the ODP-239 collection in this experiment. Due to space limitations, we refer readers to [9] for more details about F_β measure.

Since the OPTIMSRC produces clustering by analyzing the outputs of multiple search result clustering algorithms, we can take OPTIMSRC as an optimal approach. The performance result, which is shown in Table 6, illustrates that our algorithm outperforms KeySRC and Lingo for all evaluation measures and it is close to the result of OPTIMSRC. Although the performance of WhatsOnWeb is also close to OPTIMSRC, the response time of WhatsOnWeb is still a big problem.

Table 6 Performance of clustering algorithms on the ODP-239 test collection expressed as mean F_β measure for several values of β

Algorithms	F_1	F_2	F_5
CAFIG	0.334	0.337	0.349
OPTIMSRC	0.313	0.341	0.38
KeySRC	0.295	0.318	0.341
Lingo	0.273	0.283	0.294
WhatsOnWeb	0.33	0.347	0.37

5 Conclusion

In this paper, we design and implementation an algorithm for efficiently clustering search results. Our design applies the frequent itemsets mining and a graph structure to group the semantical pages into a cluster. The major advantage of this design is that it drastically reduces the processing time and increase the clustering accuracy. Our performance study shows that this design exhibits a superior performance of our algorithm over other approaches.

Currently, we are extending the capability of this design to dealing with the representation of clusters to the end user. Current clustering engines only focus on how to generate clusters, without considering user feedback on the representation of clusters. We try to determine the importance of each cluster and adjust the presentation of clusters accordingly. A feedback clustering algorithm that works for search result clustering is under designed.

References

1. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in large databases. In: VLDB, pp 487–499
2. Bernardini A, Carpineto C, D’Amico M (2009) Full-subtopic retrieval with keyphrase-based search results clustering. In: Web intelligence, pp 206–213
3. Carpineto C, Osinski S, Romano G, Weiss D (2009) A survey of web clustering engines. ACM Comput Surv 41(3):1–38
4. Carpineto C, Romano G (2010) Optimal meta search results clustering. In: SIGIR, pp 170–177
5. Giacomo ED, Didimo W, Grilli L, Liotta G (2007) Graph visualization techniques for web clustering engines. IEEE Trans Vis Comput Graph 13(2):294–304
6. Han J, Pei J, Yin Y (2000) Mining frequent patterns without candidate generation. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data, pp 1–12 ACM
7. Manning CD, Raghavan P, Shtze H (2008) Introduction to information retrieval. Cambridge University Press, New York
8. Osinski S, Stefanowski J, Weiss D (2004) Lingo: search results clustering algorithm based on singular value decomposition. In: Intelligent information systems, pp 359–368
9. Rijsbergen CV (1979) Information retrieval. Butterworth-Heinemann, Newton
10. Zaki MJ (2000) Scalable algorithms for association mining. IEEE Trans Knowl Data Eng 12(3):372–390

A Data Gathering Scheme Using Mobile Sink Dynamic Tree in Wireless Sensor Networks

Kilhung Lee

Abstract This paper suggests a data gathering scheme for wireless sensor networks. A mobile sink gathers data from each sensor node using a dynamic data gathering tree rooted at the mobile sink node. As the sink moves, a tree is formed and changed dynamically as with the position of the sink node. A hop-based scope filter and a transition rate to other branch for the operation and management of the tree are also suggested. Simulation results show that the proposed data gathering scheme has good results in data arrival rate and the end-to-end delay characteristics.

Keywords Data gathering · Wireless sensor network · Mobile sink · Dynamic tree

1 Introduction

A wireless sensor network consists of spatially distributed autonomous sensors to monitor physical or environmental conditions, such as temperature, sound, vibration, pressure, motion or pollutants and to cooperatively pass their data through the network to a main location. The more modern networks are bi-directional, enabling also to control the activity of the sensors. The development of wireless sensor networks was motivated by military applications such as battlefield surveillance; today such networks are used in many industrial and consumer application, such as industrial process monitoring and control, machine

K. Lee (✉)

Department of Computer Science and Engineering,

Seoul National University of Science and Technology, Seoul, Korea

e-mail: khlee@seoultech.ac.kr

health monitoring. These devices can gather information about their surrounding environments once they have been deployed in small or large area [1].

For collecting data from sensor nodes, a sensor network that connecting the sensor nodes and a sink node must be formed. The routing techniques are classified into three categories based on the underlying network structure: flat, hierarchical, and location-based routing. These protocols can be classified into multipath-based, query-based, negotiation-based, QoS-based, and coherent-based depending on the protocol operation [2]. A multihop-based data gathering network can be more efficient than direct connection between sensor nodes and a sink node [3]. Sink nodes can be in a specific location and do not move. But in some cases, it needs that a sink node moves around the sensor network and collects data from sensor nodes directly. This can be more efficient in energy consuming aspects, and more accurate data collection can be possible at some specific interest area.

In this paper, a MSDT (mobile sink dynamic tree), a data gathering scheme for wireless sensor networks, is suggested. A mobile sink gathers data from each sensor node using a dynamic rooted at the mobile sink node. As the sink moves, a tree is formed and changed dynamically as with the location of the sink node.

2 Related Works

For simple and efficient delivery of data from the source to the sink in a large network, a tree-based routing scheme would be more flexible and more practical. MobiRoute [4] is a superset of Berkley MiniRoute [5]. MiniRoute is a routing protocol designed specifically for the all-to-one data transmission of the wireless sensor networks. It takes a distributed distance-vector-based approach: route messages are exchanged periodically among neighbor nodes, and the next hop nodes, a parent, are chosen by evaluating the cost of the routing data through different neighbors. MobiRoute applies a beacon mechanism to trace the state of the neighbors of a sink. MobiRoute increases the rate of route message exchange to speed-up topological changes, but limits the propagation by performing broadcasts until the sink reaching to the anchor node.

A sink node can move while gathering data from sensor network. SMS (sink mobility support) supports the sink mobility of the conventional routing protocols [6]. It does not use flooding method, and does not need to know the geometric location of sensor nodes. This algorithm incurs very small communication overhead. In case of mobile sinks, conventional routing protocols can be drastically improved in terms of both energy and delay. ALURP (adaptive local update-based routing protocol) is a solution with adaptive location updates for mobile sinks to resolve collision and energy consumption incurred by frequent location updates [7]. When a sink moves, it only needs to broadcast its location information within a local area other than among the entire network.

There are some cases where data sink are more than one. TTDD (two-tier data dissemination) model provides scalable and efficient data delivery to multiple

mobile sinks [8]. Each data source in TTDD proactively builds a grid structure which enables mobile sinks to continuously receive data on the move by flooding queries within a local cell only. TTDD's design exploits the fact that sensor nodes are stationary and location-aware to construct and maintain the grid structure with low overhead. TTDD approach can handle multiple mobile sinks effectively with performance comparable with that of stationary sinks. SEAD (scalable energy efficient asynchronous dissemination) protocol seeks minimization of energy consumption in both building the dissemination tree and disseminating data to mobile sinks [9]. SEAD protocol considers the distance and the packet traffic rate along nodes to create near-optimal dissemination trees. The sinks can move without reporting their location to the tree while receiving data updates successfully.

The tree topology is changed as sink moves. DST (dynamic shared tree), an extension of the DDT (distributed dynamic tree), is able to accommodate multiple mobile sinks [10]. DDT is able to identify current location of the sink locally and dynamically transforms the tree shape according to the sink movement. DST performs considerably energy-efficient data with relatively low delay. Rob [11] proposes a method for reconfiguration of a tree-based wireless sensor network with a mobile sink. It does not only reconstruct the routing tree, but also optimizes the parameters of the nodes that were affected to improve the service level. It is efficient and scalable, and able to flexibly trade reconfiguration cost for quality to match the demands of the application.

3 Creation and Dynamic Management of Sink Tree

In sensor networks, each node is deployed at the point where environmental data is required and data is sent to the sink using a data gathering tree. If required, a mobile sink collects data by going the interested area on either a regular or an occasional basis. Each sensor node has a neighbor table and a parent node table. A neighbor table has a set of detected neighbor node. A parent node table is a set of neighbor nodes that sent a data interest. A parent node is a next node to go to the sink node in data path. When a data interest is received for data collection from a sink, a new route to the sink is made and updated. A route element in routing table is made when there is a new interest. It is eliminated when there is not data transfer anymore.

A mobile sink node sends data interest message to neighbor nodes periodically. This data interest message comes from a mobile sink node and floods along the network, and finally reaches to all nodes in a specific area of networks. This specific area is defined as a filter in the message expressed by area description or number of hops to be delivered. A sink tree generated by an interest message is temporal and will be disappeared after an expiration time.

As a sink moves, the neighbor node of the sink node will be changed. So, the tree should be changed, too. The change of the tree is minimized by accepting only a new node or permitting only a shorter path from each sensor nodes to the sink node.

By doing so, we can minimize the exchange of the control data and depress the dissipation of the energy of the sensor node, and minimize the data transfer delay.

3.1 Sink Tree Creation

For collecting data from sensor nodes, the sink node broadcasts an interest message to sensor nodes. An anchor node, which is a neighbor node of the sink, registers to the sink node and gets a branch identifier from the sink node. After successful registration, the anchor node adds the branch identifier to the data interest and increases the hop counts, and broadcasts it again. Whenever a node receives an interest that has a smaller value than that of current hop counts to the sink node, each node registers to the sending nodes as a parent node and makes a parent-child relationship. By repeating this operation, the sink tree is formed and expanded.

There is a filter that defines the scope of an interest message to be delivered. The filter element consists of two fields: one for the hop counts and the other for the expiration of the interest. If the value of hop counts in a message received by a node reaches to the maximum hop counts (*max_hop*), the message is not broadcasted any more. If the current time of the node passes over the timeout time of the interest, the message is ignored.

Once registered to a parent as a child, the node ignores additional interest of the same sink that has more than or equal hop counts compared with the previously received. But the parent table and the routing information are updated. When a branch node receives the same interest from the sink node, the node refrains from sending the interest again. But, when the branch node receives the same interest from sink node after branch waiting timeout (*br_timeout*), the node broadcast the interest message to neighbors. The following is a procedure for an interest message.

```

1  Procedure Rx_Interest(sender, sink, bid, hop, out_time)
2  Increase hop counts and Update Route_Table (sink)
3  If parent == null or hop < hops (sink) then
4    Send Register message to the sender
5    Exit
6  Endif
7  If hop >= max_hop then
8    exit
9  Endif
10 If out_time > current_time then
11   exit
12 Endif
13 If current_time > br_time + br_timeout then
14   Broadcast Interest (this, sink, bid, hops(sink), out_time)
15   br_time = current_time
16 Endif

```

After receiving an interest message, the node updates the neighbor node table and the parent node table. Same interest message can be received from several neighbor nodes with different path. The node selects the path that is shortest and received first. The other sending node will be a candidate parent for that interest. After making a parent–child relationship by registering to the sender, the node broadcasts the interest message to neighbors. If a node receives the same interest from the same parent, the node refrains broadcasting the interest immediately. A child node broadcasts the interest only when the branch timeout time passes if it has sent the interest message before.

3.2 Dynamic Sink Tree Update

A sink node collects data from sensor nodes while it is moving or stopping. As the sink node moves, the shape of the tree is changed. Some anchor nodes connected to sink node lost their connection when it moved away. Some new nodes are added as an anchor node when the sink node is approaching. Now, they can communicate with the sink node directly. The broken connection to the sink node is detected by timeout mechanism after failing the interest from the sink node. Then, the node finds a neighbor node that is connected to the sink tree and makes a parent–child relationship with that node. The new parent node would be one of the candidate parent nodes.

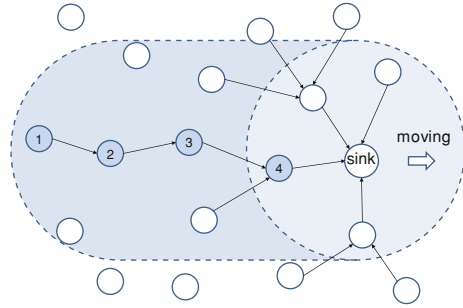
When a node that is not an anchor node receives an interest message from a sink node at first time, the node will be an anchor node. The node broadcasts an interest message after registering and getting a new branch identifier. When a node that is not an anchor node receives an interest message that has shorter hops than before, the node updates routing table and makes a parent–child-relationship with that node. The sink tree is maintained adequately by a timeout mechanism. Each node does not broadcast interest message whenever it receives interest message. The following cases are the time when a node broadcast an interest message.

1. When a node successfully registered to a parent node.
2. When a node receives an interest message after active tree timeout.
3. When there is a change in hops to the sink node.

There are two cases when a node registers to a parent node. One is the case when a node receives an interest message at first time. The other case is when a node receives an interest message that has shorter hops compared with the hops of the previous parent node.

Next, an anchor node broadcasts an interest message to its branch whenever there is a timeout at the branch. The anchor node does not send an interest message when the node receives an interest message from the sink node. The sink node broadcasts an interest message periodically but this message is screened at anchor nodes. Instead, an anchor node broadcasts an interest message whenever a branch timer is timeout. The message sent from an anchor node is restricted in that branch.

Fig. 1 A formation and a shape of sink tree while sink are moving. The maximum length of the tree is confined to length 2 and the node 1, 2, 3 are still connected to a sink tree when there is a data transaction



So, if a node outside of the branch receives this message, the node updates routing information but refrains from any other action. When a node registers to a parent that has better path to the sink, the node sends an interest message to its neighbor. After receiving this message, neighbor nodes update their parent table and may start changing their parent node. But in most cases, there is no action in child nodes when there has been no data transfer from the child nodes.

When a sensor node in a sink tree has a data to send, it sends the data to the parent node of the sink tree. If a parent node receives a data from a child node, the node updates the tree usage information. If a parent node does not receive a data from its child nodes in a period of time, the parent node eliminates such child node from the sink tree. So, a child node can be included in a sink tree when there is a data reception from that child even if the length of the child exceeds the maximum hop counts. Figure 1 shows a formation of a sink tree and also illustrates such an example.

3.3 Traffic Control of a Sink Tree

Each sensor node sends data generated from its own and transfers data that comes from child nodes. If data congests more than the available bandwidth of the node, the waiting delay of the data is increased at that node. This eventually increases the end-to-end delay of the data from sensor nodes to a sink node. When traffic concentrates to a certain node, the traffic that comes to this node must be rerouted to other path. For this, the shape of a sink tree is changed dynamically. For controlling the traffic of a sink tree, the following traffic distribution scheme can be employed.

1. A sink node monitors all incoming traffic and controls the entire traffic by specifying and changing the filter parameters of the interest message.
2. An anchor node monitors all incoming traffic and its own traffic and controls traffic by specifying control parameters of the branch of the tree.

3. A parent node monitors all incoming traffic and its own traffic and controls traffic by recommending to its child for changing the parent from it to other candidate parent node.

For the control of the traffic of each branch, a parameter with the name of “*transition rate*” is devised. This parameter describes the tendency of a node transition to a new branch of the sink tree. When a duplicated interest message received from the other node different from its parent, the node checks the hop counts of the message. When this value is less than that of the previously registered parent, the node changes the parent to a new node. When the hop value is same but has different branch id, the node changed to a new parent with a certain value of the probability. With this transition rate, the sensor node moves to a new parent when it receives with same hop counts and different branch identifier. A sink node specifies this parameter after considering the shape of the tree and the speed of the sink node. An anchor node can modify this value when there is congestion at its branch. A parent node can change this value temporarily when congestion occurs in its node.

4 Simulation Results and Evaluation

For the evaluation of the proposed MSDT (mobile sink dynamic tree) scheme, a standalone simulation program is used. There are a hundred of nodes in simulation network. Each node deployed in $2000\text{ m} \times 2000\text{ m}$ area with 200 meters apart. Each sensor node is fixed after deployment and a mobile sink moves around sensor networks with variable speed between 0 to 100 meters per seconds. In simulation, data is collected within 5 hops from the sink node. For the evaluation of the proposed scheme, data arrival rate and end-to-end data transfer delay is collected and compared.

Figure 2 is the result of data arrival rate of the sink node when it moves around sensor network. The speed varies from 0 to 100 m/s. When the speed is under the 20 m/s, all data are arrived to the sink node correctly. As the speed increases, the arrival rate is decreased. When the speed goes over 50 m/s, the arrival rate goes down under 60%. So, for the stable data gathering operation, the speed of the sink node should remain under 20 m/s. The arrival rate is a little bit larger when the transition rate is high, but the difference is so much.

From Fig. 3, we can see the characteristics of the end-to-end delay from sensor nodes to the sink node. The scope filter is defined as the maximum hop value of 3. So, only nodes near the sink node send data to the sink. When the sink is not moving, the end-to-end delay is about 14 ms. As the speed of the sink increases, the delay also increased, too. But, as the speed of the sink increases more, the delay is decreased at this time. This is because the arrival rate of the message is decreased as shown Fig. 2. Only the messages near the sink could arrive well and the data sent from the long distance from the sink node didn't arrive and not

Fig. 2 The results of data arrival rate when the speed of the sink node varies

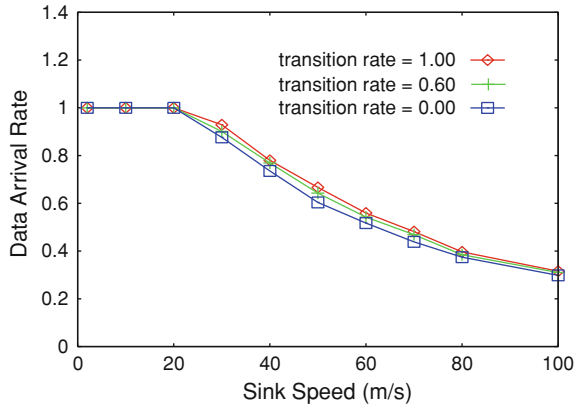
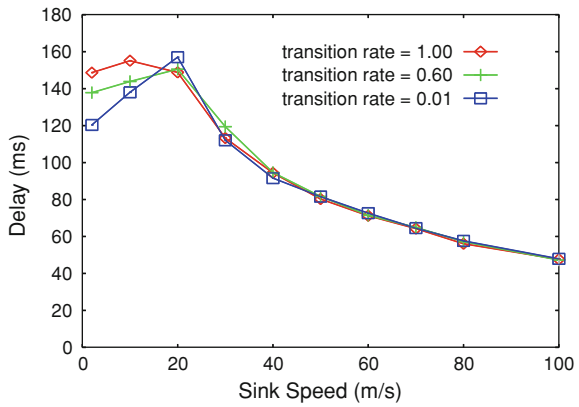


Fig. 3 The results of the end-to-end delay from sensor nodes to the sink node when the moving speed of the sink node varies



counted. As with the variation of the transition rate, the difference can be shown only at low speed. When the transition rate is high, the delay is increased a bit at low speed.

From these results, we can see that the mobile sink dynamic tree operates well in a moderate speed. The arrival rate of the sink node shows good characteristics when the speed of the sink node is not high. The delay is dependent with the speed of the sink node. The delay can be minimized when we increase the transition rate of the branch at low speed, but it shows no difference when the sink moves in high speed.

5 Conclusion

In this paper, we propose a data gathering mechanism for mobile sink in a sensor network. The sink tree is created and changed dynamically as with the movement of the sink. By constructing and adapting the sink tree effectively, we can obtain

some good characteristics in arrival rate of the sensed data collection, end-to-end delay from sensor nodes to the sink node.

The quality of data traffic is varied as with the moving of sink node. In this paper, network traffic is controlled by changing transition rate of a node a bit. But there needs a more elaborate data quality guaranteeing mechanism for the stable and hi quality operation and management. Other issues like energy efficiency should be evaluated simultaneously.

References

1. Carle J, Simplot-Ryl D (2004) Energy-efficient area monitoring for sensor networks. *IEEE Comput* 37:40–46
2. Al-Karaki JN, Kamal AE (2004) Routing techniques in wireless sensor networks: a survey. *IEEE Commun Mag* 11:6–28
3. Lee KH (2010) A time tree medium access control for energy efficiency and collision avoidance in wireless sensor networks. *Sensors* 10(4):2752–2769 MDPI
4. Jun L, Jacques P, Michal P, Matthias G, Jean-Pierre H (2006) Mobiroute: routing towards a mobile sink for improving lifetime in sensor networks. *Lecture Notes in Computer Science (LNCS)*, vol 4026. Springer, Berlin, pp 480–497
5. Woo A, Tong T, Culler D (2003) Taming the underlying challenges of reliable multihop routing in a sensor networks. In: *Proceedings of the first ACM SenSys*
6. Park CS, Lee KW, Kim YS, Ko SJ (2009) A route maintaining algorithm using neighbor table for mobile sinks. *Wirel Netw* 15(4):541–551
7. Wang G, Wang T, Jia W, Guo M, Li J (2009) Adaptive location updates for mobile sinks in wireless sensor networks. *Lect Notes Comput Sci* 47(2):127–145
8. Ye F, Luo H, Chung J, Lu S, Zhang L (2002) A two-tier data dissemination protocol for large-scale wireless sensor networks. In: *Proceedings of the ACM/IEEE international conference on mobile computing and networking (MOBICOM)*. Atlanta, Georgia. pp 23–28
9. Kim HS, Abdelzaher TF, Kwon WH (2003) Minimum energy asynchronous dissemination to mobile sinks in wireless sensor networks. In: *Proceedings of the first ACM international conference on embedded networked sensor systems (ACM Sensys 03)*. pp 193–204
10. Hwang KI, Eom DS (2006) Energy-efficient data dissemination in sensor networks using distributed dynamic tree management. *Lect Notes Comput Sci* 4104:32–45 Springer
11. Rob H, Twan B, Wai Leong Y, Chen-Khong T, Marc G, Henk C (2009) QoS management for wireless sensor networks with a mobile sink. *Lect Notes Comput Sci* 5432:53–68 Springer

An Enhanced Resource Control Scheme for Adaptive QoS over Wireless Networks for Mobile Multimedia Services

Moonsik Kang and Kilhung Lee

Abstract In this paper, an enhanced resource control scheme based on traffic estimation for adaptive QoS control is proposed with the use of the IEEE 802.11e wireless LAN standard for wireless access network. The proposed network model consists of both the core network and a number of wireless access networks including several mobile hosts. The interface between the core network and the access network is designed to include a means of provisioning differentiated service (DS) according to the requirements of a particular flow. Simulation results demonstrate the effectiveness of the proposed enhanced resource control scheme at the aspect of the available bandwidth and throughput.

Keywords Resource Control · Adaptive QoS · Traffic estimation · Wireless access network · Differentiated service (DS)

1 Introduction

With the successful development of the mobile Internet access technology, new demands for multimedia applications over the Internet including both wired and wireless parts have been rapidly increasing [1–3]. In order to cope with the traffic

M. Kang

Department of Electric Engineering, KanungWonju National University,
Kangnung, Kanwon-do, Korea
e-mail: mskang@gwnu.ac.kr

K. Lee (✉)

Department of Computer Science and Engineering,
Seoul National University of Science and Technology, Seoul, Korea
e-mail: khlee@seoultech.ac.kr

requirements, several research topics have aimed at providing users for the required Quality of service (QoS) at different network layers [1, 4]. Here, QoS is the ability to provide different priority to different applications, users, or data flows or to guarantee a certain level of performance to a data flow. For this, we consider the appropriate service model which is required to meet diverse multimedia traffic requirements, such as efficiency, fairness and scalability.

We propose an enhanced resource control scheme considering traffic estimation for adaptive QoS in wireless networks, which is a solution for the required QoS from one end of the network to the other. This may just be a wired connection between the access point and the router, which lies at the border of the core-IP network. The core network is followed by a similar access network at the other end, and finally ends at another wireless user.

It is mentioned that both DS and wireless LAN QoS methodologies try to provide a better service for specific classes of traffic [5–7], and not for the particular end-to-end flows [8]. In this sense, our overall framework may be more specifically considering as a Class of Service (CoS) optimization. Also, we show a solution to optimize the performance of the network for different classes of traffic and a plan to introduce dynamic provisioning based on traffic estimation scheme. The real-time traffic monitoring and estimating is necessary at the ACA (Admission Control Agent) from which now also provisions individual BR (Border Router) based on this information [5].

2 Service Model and Access Network

The primary goal of QoS scheme is providing a prioritized service including dedicated bandwidth, controlled jitter and latency, and improved loss characteristics. Also, it is important to make sure that providing priority for one or more flows does not make the other flows fail. The QoS scheme enables us to provide a better service to certain flows. This is done by either raising the priority of a flow or limiting the priority of another flow. When using congestion management tools, we try to raise the priority of a flow by queuing and servicing queues in different ways. The queue management tool used for congestion avoidance adapts priority by dropping lower-priority flows before servicing higher-priority flows. Policing and shaping scheme provide a priority to a flow by limiting the throughput of other flows. Link efficiency tool limits some large flows that showing a preference for other smaller flows.

The DS model is an appropriate architecture for implementing a scalable service differentiation in the Internet by aggregating traffic classification state [5, 6]. Instead of maintaining state information, the DS applies different PHBs (Per Hop Behaviors) that are specified by DS Code Point (DSCP) in the ToS (Type of Service) field of IP header [3]. Because of aggregation function, the core routers in DS network only maintain minimum state information and yet provide the required QoS. In DS model, DS domains will negotiate a Service Level Agreement (SLA)

when they forward traffic to each other. Using the SLA, a traffic profile is taken for configuring the border routers. As a part of SLA the Traffic Conditioning Agreement (TCA) is translated into a DS specific conditioning TCS (Traffic Conditioning Specification). The TCS is defined as a set of parameters specifying the traffic profile.

The Traffic Conditioning Framework (TCF) includes two parts: traffic classifier and traffic conditioner. The Traffic classifier is used to select packets from incoming packet stream according to predefined rules. Two kinds of classifiers are defined in the DS model. The classifiers may be located at the ingress nodes or at interior nodes in the DS domain. Generally, the classifier located at the ingress node is a MF (Multi-field) classifier. The other is a BA (Behavior Aggregate) classifier located at Interior routers; this classifier is based on DSCP value. The DSCP is a reformatted ToS field of the IP header, which is used to define the class of the packet. This class specifies both forwarding treatment (scheduling) and path selection (routing). Forwarding treatment is a set of rules defining the importance of a class compared to other classes. These rules characterize the relative amount of resources, which should be dedicated for a particular class in the scheduler, and the packet dropping order during congestion. The Traffic conditioner is used to verify whether the offered traffic is in compliance to the agreed profile (subscribed information). Two kinds of routers are identified in DS domain, i.e., border routers and core routers. Border routers exchange packets with other domains and perform traffic conditioning. Border routers are allowed to keep per-flow information, and core routers examine the DSCPs of the packets. Thus, mapping them to different PHBs gives appropriate forwarding treatments to packets.

An RSVP reservation request message contains a flow descriptor. The flow descriptor is composed of two parts: the flow specification (*flowspec*) and the filter specification (*filterspec*). The *flowspec* describes the traffic and the desired QoS. The *filterspec* specifies the parameters to which the *flowspec* needs to be applied. This is done by setting up these parameters in the filter or classifier. The *flowspec* carries the flows traffic specification (*Tspec*) and the requested service specification (*Rspec*). The sender specifies *Tspec* and the receiver send *Rspec*.

Providing QoS is a challenging issue in 802.11 networks due to the limitations of the wireless medium [7]. In the advent of QoS in the IP Core Network, it has become imperative that the wireless access network also provide the required QoS. The end-to-end QoS requires not only QoS support mechanism in the core network, but also in the access networks. The 802.11e proposes an Enhanced Distributed Co-ordination Function (EDCF) for wireless access. EDCF is an extension of the existing DCF scheme with some of the elements of the MAC parameterized per Traffic Category (TC), which works to prioritize traffic on the basis of Access Categories (AC).

Each MAC Frame is tagged with a Traffic Category Identification (TCID). With a mapping of the TCIDs into the ACs, the 8 TCs map directly to the RSVP protocol and other protocol priority levels. The AC3 has highest priority and AC0 has lowest priority. The TCIDs are not strictly in numerical order. This is because the EDCF mechanism has been designed to be compatible with IEEE 802.1D/Q [7].

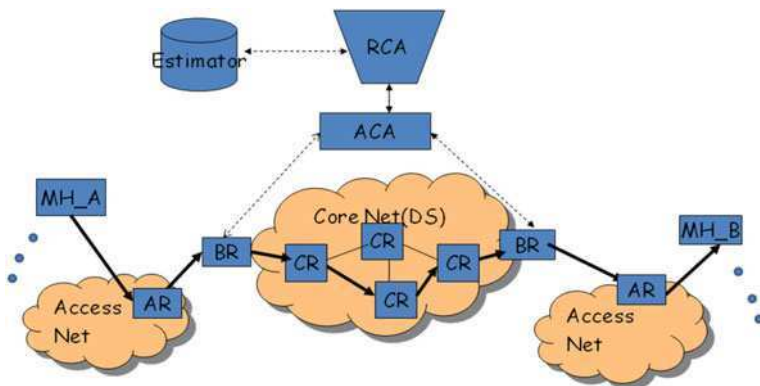


Fig. 1 The proposed resource control network model for adaptive QoS

The TCIDs are the same as the User Priority (UP) tag of the 802.1D/Q Header. The EDCF works based on the TCID values to provide a kind of statistical priority for Traffic.

3 Proposed Resource Control Scheme for Adaptive QoS

3.1 Network Model and Core Network Architecture

Our proposed network model is shown in Fig. 1, of which model will also support legacy 802.11 users. The wireless access point should also be 802.11e enabled. It should be capable of interfacing with the rest of the QoS network. In this regard, we introduce the concept of Access Router (AR). The AR is an AP, which is capable of providing all this functionality. So, AR becomes the end point for PHB communication as also the Service Level Specification (SLS) [3].

In this paper, the AR establishes the requirements based on traffic estimation comparing with the current observed traffic load, and then accordingly asks for service from the network. Thus, the AR must also do some form of traffic monitoring. Another important function of the AR is the marking of packets so that the core DS network may easily recognize it. This is important for the translation of information between the two networks. This is the issue of integrating the 802.11e with the DS framework [7]. The AR and boundary entity of the DS core are the critical elements of this integration. The AR uses the User Priority Tag to mark packets of different type. The integration consists of a translation between these UP tags and the DSCP. The basic thing that needs to be achieved by the inter-networking is actually to translate the 802.11e parameters to DS parameters. The four classes of traffic then map to different TCIDs within the 802.11e framework.

Thus a direct mapping of the DSCP field to the TCID field, and vice versa can be formulated. This gives a simple mechanism for translation.

The boundary entity is co-located at the ingress router to the DS network and is now called a Border Router (BR) as in Fig. 1. The BR has a number of functions and is under the control of the resource control agent (RCA). It maintains the interface with the AR. It is in charge of receiving packets from the AR and marking them with an appropriate DS code point (DSCP). This is not necessarily a direct translation of the UP tag. This is because if incoming traffic is in excess of what is expected; it will be marked simply as BE traffic. In this way the BR performs admission control for incoming traffic.

The RCA can also instruct the BR to drop packets from certain users, or of a certain flow. The BR forwards all incoming traffic information to the RCA. It must provide policing to account for falsification. This can be done by any standard means such as token bucket/leaky bucket policing. The BR may also take part in control layer signaling with other DS routers as well as the RCA. This signaling is to react to emergencies, congestion as well as provisioning. In general, the BR does not perform this role. The BR in fact, can forward RSVP messages to the RCA, who ultimately decides whether or not to grant a certain request. The managing entity is called as the RCA. The RCA is the central management entity. It is in charge of traffic monitoring, dynamically provisioning the DS network based on current load and Time of Day, as well as indirectly controlling the admission control of the BR. This is the end point for RSVP communication. Thus, it is the element in the network, with which users communicate. It may be supported by a database to store SLA information, as well as traffic pattern and monitoring information.

The DS routers within the network are now called core routers. The typical DS core network consists of a number of BRs, a RCA, and a number of CRs (Core Routers). The DS network has a BR at each and every ingress/egress point to the network. For simplicity, at present we restrict our model to two BRs at either end of the DS core. The BR has a number of functions under the control of the RCA. It maintains the interface with the AR. It is in charge of receiving packets from the AR and marking them with an appropriate DSCP.

In this way the BR performs admission control for the incoming traffic. The RCA can also instruct the BR to drop packets from certain users, or of a certain flow. The BR also forwards all incoming traffic information to the RCA. It must also provide policing to account for falsification. This can be done by any standard means such as token bucket. The BR may also take part in control layer signaling with other DS routers as well as the RCA. This signaling is to define emergency conditions, congestion as well as provisioning. The BR can forward RSVP messages to the RCA, who ultimately decides whether or not to grant a certain request. The RCA is the central management entity, which is in charge of traffic monitoring, dynamically provisioning the DS network based on current load and Time of Day, as well as indirectly controlling the admission control of the BR. This is the end point for RSVP communication. Thus, it is the element in the

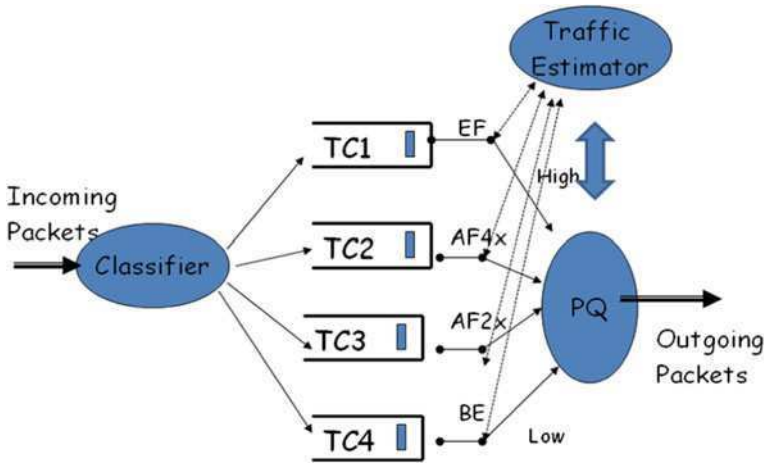


Fig. 2 Bandwidth allocation scheme and traffic classifier

network, with which users communicate. It may be supported by a database to store SLA information, as well as traffic pattern and monitoring information.

3.2 Bandwidth Allocation Scheme Based on Traffic Estimation

In our architecture, the ingress traffic is continuously monitored and estimated. The infrastructure continuously collects data about the traffic. Over time this collection of data, helps to characterize the behavior of the traffic. For example it tells us what kind of traffic dominates at a particular time of day. As an example, we may find that in the morning there is a large amount of data traffic, whereas late in the evening and at night voice traffic tends to predominate. This enables us to compile data about traffic patterns over time. We can then begin to define the required parameters in the core of the network to support this variation of traffic over time.

As shown in Fig. 2, the basic concept within a Core DS router is that of four priority queues—One for Expedited Forwarding (EF), two queues for Assured Forwarding (AF4x, AF2x), and the fourth for Best effort (BE) traffic. As can be seen from the figure different weights (or precedence/priority) for weighting algorithms such as Priority Queuing (PQ) govern each of these queues. It can be seen from the figure that the EF queue, the highest priority one, passes through a single weighing stage, that of the PQ. While AF and BE traffic pass through two levels of weighing stage of PQ. Thus, the EF queue has highest precedence and the least number of weighing stages. As shown in the Fig. 2, the network service maps to different Traffic Classes (TC). Thus, the real time voice traffic is for TC1, which is implemented using EF and so on.

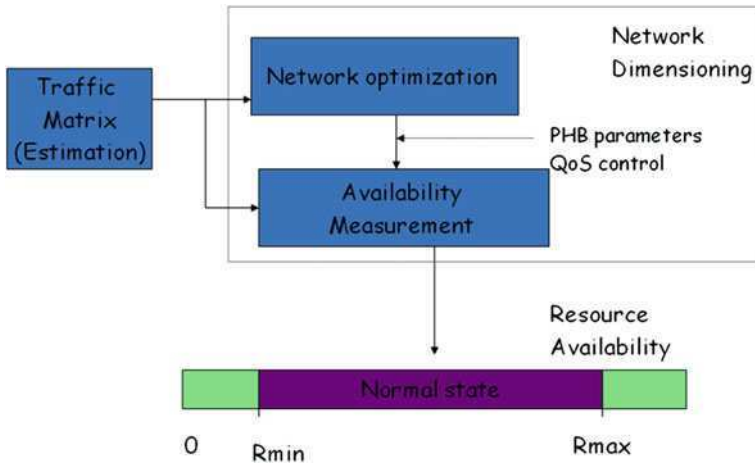


Fig. 3 Resource control and optimization according to traffic matrix (estimation)

The traffic matrix is specified based on traffic patterns according to traffic estimation as shown in Fig. 3. The values within such a pattern matrix are considered to be typical or normal values. The normal values are of course, within a predefined threshold. This is a so called normal state of affairs. In the normal state, we can safely provision the network according to our pre-specified matrix. Continuous monitoring of the incoming traffic enables us to recognize whether at any given time the incoming traffic is within the bounds of the expected traffic. In the presence of sudden changes, the network as an abnormal state notes such a change. In an abnormal state, the network reacts by further changing the weight in discordance with the matrix above. Thus, we can account for such sudden variations. As soon as the network returns to a normal state, we also bring back parameters to the recommended values. This brings us to another interesting implication. Since we are monitoring the network condition continuously it gives us the opportunity to in fact record variations in traffic pattern over longer periods of time. Thus, the weights within the matrix can be defined and re-defined over time as a continuously varying function. Though, this is an additional overhead, the advantages gained from this optimization make up for the initial overhead incurred.

4 Performance Evaluation

In order to evaluate the performance of proposed scheme, several QoS parameters are considered as the followings. First, the delay parameter may refer to either propagation delay or round-trip delay. Propagation delay refers to a short but finite

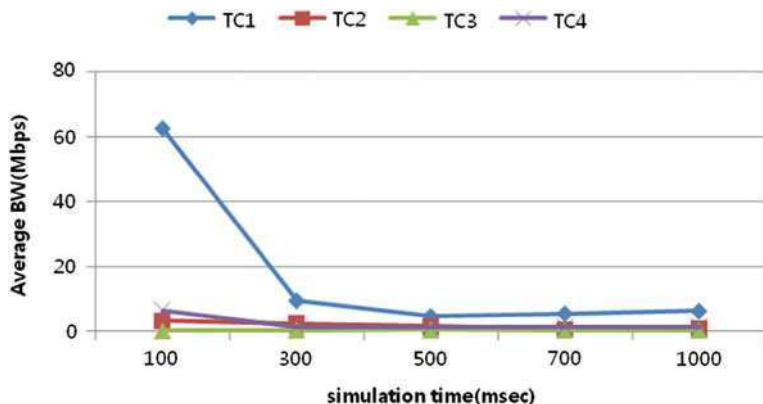


Fig. 4 Average BW for each traffic class according to simulation time

delay for a signal to travel from one end of a transmission medium to the other. Round-trip delay refers to the delay between the first bit of a block being transmitted by the sender and the last bit of its associated ACK being received. Second, the packet loss parameter refers to the ratio between the number of lost packets from source to destination domains and the amount of packets submitted by the application in the source domain. This may be due to either buffer overflow or, in the case of real-time applications, to the end-to-end delay incurred longer than the specified maximum delay requirements.

Finally, throughput can be defined as the bit rate coming out of the last hop of the service scope in the destination domain. This parameter may or may not be considered as an independent parameter, because depending on the definition of the traffic profile it can be calculated as a function of both the transmission rate and the packet loss. The state of the network is determined by the rate at which messages arrive and depart from various queues as well as the set of messages waiting for service. Hence, the state of the network is the collection of all individual nodes and link states. Network traffic consists of messages in the network and can generally not be abstracted by a generic representation. Messages must be represented explicitly as they determine the behavior of nodes and links at a particular point in time. In order to analyze the performance of the proposed scheme, we consider the efficiency measured by the average bandwidth. Here, the autoregressive (AR) model is used for the traffic estimation. Thus the traffic arrival process is selected as Poisson model with mean throughput of 89 Mb/s for the simulation network model as shown in Fig. 1.

Figure 4 shows the average access bandwidth (BW) according to each traffic class. These results show that the BW allocation performance of TC1 (EF traffic) is much improved compared with TC2, TC3, and TC4. This performance resulted from the adaptive resource control scheme considering the flow condition based on the traffic estimation for the high priority packets. In the simulation work, the

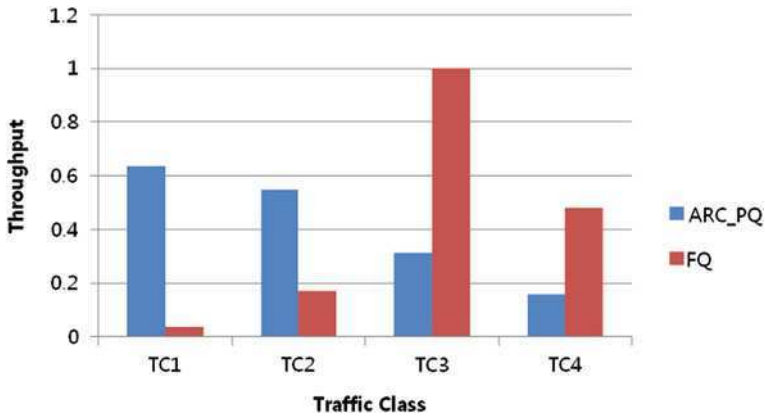


Fig. 5 The end-to-end average throughput

estimation-based resource control adaptive QoS queuing (ARC_PQ) and Fair queuing (FQ) functions are implemented respectively between AR and the core routers (CR) in the core network. The end-to-end access throughput with ARC_PQ and FQ is shown in Fig. 5. This figure shows more than four times difference of end-to-end performance for the TC1 (EF) traffic between ARC_PQ and FQ. When the simulation time is very short and the traffic density is heightened, ARC_PQ performance is much greater than FQ performance.

5 Conclusion

In this paper, the traffic estimation-based resource control model is introduced for the multimedia traffic transmission with the use of the wireless access network. This scheme can provide end-to-end QoS guarantees between mobile users over wireless networks. It classifies all the traffics into different types and then accordingly treats them differently as performed in DS model.

The proposed scheme is analyzed for the multimedia traffic patterns in wireless network model. It is also a scalable solution in the core network since the low-overhead performance nature of behavior aggregation transmission. The performance of the proposed scheme was evaluated from the aspect of access throughput. Simulation results show that the proposed resource control scheme has the better performance of the low delay. For further study, we will devote time to the research of more efficient QoS control schemes to cooperate with wireless sensor and ad-hoc networks including the extension of the service coverage area.

References

1. Bayan AF, Wan T-C (2010) A scalable QoS scheduling architecture for WiMAX multi-hop relay networks. 2010 ICETC
2. Masip-Bruin X, Yannuzzi M et al (2007) The EuQoS system: a solution for QoS routing in heterogeneous networks. In: IEEE communications magazine, February 2007
3. Trimintzios P, Andrikopoulos I et al (2001) A management and control architecture for providing IP differentiated services in MPLS-based networks. In: IEEE communications magazine, May 2001
4. Hanzo L, Tafazolli R (2007) A survey of QoS routing solutions for mobile ad-hoc networks. IEEE Commun 9(2):50–70
5. Zhang F, Macnicol J (2006) Efficient streaming packet video over differentiated service networks. IEEE Trans Multimed 8(5):1033–1044
6. Wang S, Xuan D, Zhao W (2004) Providing absolute differentiated services for real time applications in static priority scheduling networks. IEEE/ACM Trans Netw 12(2):326–339
7. IEEE 802.11 (2001) WG draft supplement to international standard. In: IEEE 802.11e/D2.0, Nov 2001
8. Qin D, Shroff N (2004) A predictive flow control scheme for efficient network utilization and QoS. IEEE/ACM Trans Netw 12(1):161–172

An Analysis of Critical Success Factor of IT based Business Collaboration Network Implementation

Hangbae Chang, Hyukjun Kwon and Jaehwan Lim

Abstract In the diversifying competitive environment, enterprises have cooperated with others in various ways. In the competitive environment where increase in uncertainty of demand and supply, market globalization, diminish in Product Life Cycle (PLC), and rapid changes in technologies and business process take place, the enterprises are able to guarantee their own competitive advantage by collaborating with other enterprises. Under this situation, many enterprises have realized an importance of collaboration with others and changed a traditional transaction relationship into a collaborative relationship. Hence, we would like to analyze a Critical Success Factor (CSF) of IT business collaboration and performance evaluation to develop a method to facilitate IT business collaboration.

Keywords Information technology · IT business collaboration · Inter-organizational system · Critical success factor (CSF)

H. Chang (✉) · J. Lim

Department of Business Administration, Daejin University,
Hogukro 1007, Pocheon-Si, Gyeonggi-Do, Korea
e-mail: hbchang@daejin.ac.kr

J. Lim

e-mail: lim0410@daejin.ac.kr

H. Kwon

Yonsei University, New Millenium Hall, 262 Seongsanno,
Seodaemun-Gu, Seoul 120-749, Korea
e-mail: gloryever@gmail.com

1 Introduction

In the diversifying competitive environment, enterprises have cooperated with others in various ways. In the competitive environment where increase in uncertainty of demand and supply, market globalization, diminish in Product Life Cycle (PLC), and rapid changes in technologies and business process take place, the enterprises are able to guarantee their own competitive advantage by collaborating with other enterprises. Under this situation, many enterprises have realized an importance of collaboration with others and changed a traditional transaction relationship into a collaborative relationship.

The enterprises have implemented an information system for the collaboration and the information system connects business process among enterprises for sharing information. Such business collaboration based on IT is expressed in various ways such as “Information Partnership, Electronic Partnership, and Electronic Integration”. The meaning of business collaboration is located between “Market Relationship and Rank Relationship” and constructing a governance structure [1].

Meanwhile, as a result of investigation, 37.9% of Korean enterprises, which participated in survey, answered that they cooperate with others, and 26.5% of the enterprises utilize IT for business collaboration. As this survey results indicate, the IT collaboration level is low. It is necessary to investigate the cause of low level of IT business collaboration and a way to facilitate IT business collaboration. Hence, we would like to analyze a Critical Success Factor (CSF) of IT business collaboration and performance evaluation.

2 Preliminary Study

2.1 Inter-Organizational IT Business Collaboration and Information System

IT business collaboration means that one or more than one constituent(s) of supply chain cooperate for a creation of competitive advantage through collaborative information sharing, decision making, and profit sharing [2]. This is also related to business process regarding decision making of collaboration and this concept includes joint ownership of collaborative decision and corporate responsibility of results. This is differentiated with ‘Market Relationship’, because collaboration means a cultivation of a prolonged and stable relationship, denying a temporary relationship. In addition, it does not need a direct control, so it is distinguished with a ‘Rank Relationship’. Specifically IT business collaboration includes collaborative a product development, supply forecasting, production planning, circulation and inventory management. The collaboration level is defined as a degree of collaboration of collaborative critical activities [3].

Table 1 Priority of information through IT collaboration system

Priority type of participation	1st	2nd	3rd
Managing enterprise	Delivery information	Production planning information	Resource/inventory information
Participating enterprise	Production planning information	Delivery information	Resource/inventory information

An inter-organizational information system (IT business collaboration system) means the automated information system among more than two enterprises for information sharing. Some of the examples of IT business collaboration system are common application, database, and communication network [4]. It also supports inter-organization IT business collaboration by exchanging structured/unstructured database via network. The inter-organization IT business collaboration system facilitates cooperation among enterprises and manages an inter-organizational conflict by electronic integration [5]. Consequently this function supports an expansion of business range and restructure of business process.

2.2 CSF of Inter-Organizational IT Business Collaboration and Information System

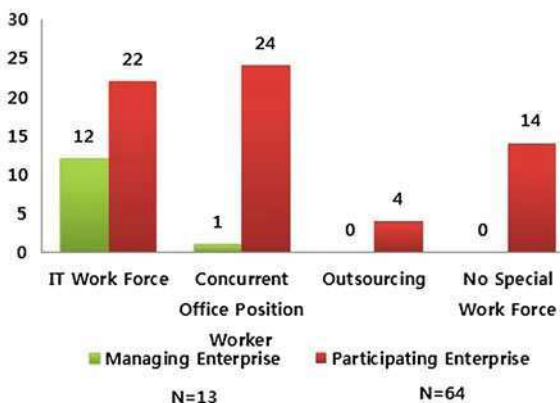
The CSF for an implementation of the inter-organizational IT collaboration system consist of External Environment, Organizational Readiness, Innovation Characteristics, Perceived Benefits, Transaction Characteristics, Resource Dependence, Network Externality, and Culture/Intuitional Forces. Table 1 shows CSF for implementing inter-organizational IT collaboration system [6].

3 CSF of Inter-Organizational IT Business Collaboration and Performance Evaluation of Utilization Level

3.1 Research Design for CSF of Inter-Organizational IT Business Collaboration and Information System

The research subjects are the enterprises, which have participated in ‘Network Construction Project for Inter-organizational Collaboration, Collaborative Network Construction Project for Inter-organizational Collaboration based on IT, ‘Large and Medium Enterprise IT Innovation Project’. Totally 77 enterprises (13 managing enterprises, 64 participating enterprises) participated in this survey, and they are from 11 different types of business (home appliance: 1, display: 6, trade: 3, textile: 1, car: 9, electricity: 23, heavy electric machine: 3, steel: 7, and aerial: 11).

Fig. 1 Status of IT work force



3.2 Analysis of CSF of Inter-Organizational IT Business Collaboration

We have conducted an empirical study on External Environment, Organizational Readiness, Innovation Characteristics, Perceived Benefits, Transaction Characteristics, Resource Dependence, Network Externality, and Culture/Intuition Forces for an analysis of CSF of inter-organizational IT business collaboration.

The study results show that the majority of enterprises have realized a necessity of implementing IT collaboration system and they seem to have sympathy of IT collaboration system which is scheduled to be implemented. But there are still some enterprises, which possess low sympathy, so it is necessary to analyze a relationship between performance of the IT business collaboration project and sympathy. Furthermore, the business processes related to IT were structured at a certain level, but some enterprises' business processes are still needed to be structured. And majority of the CEOs have high interests in IT. Especially CEOs of participating Small and Medium sized Business (SMBs) have higher interests in IT than CEOs of ordinary SMBs. Based on characteristics of SMB, an ability to manage and control business of CEOs of participating SMBs have shown a similar condition with an interest in IT (Fig. 1).

The managing enterprises primarily consist of large enterprise, and as a result of research, they have an appropriate level of IT special work force. But only 21% of participating enterprises have IT special work force or a concurrent officer. Hence the level of IT work force turned out to be CSF of IT business collaboration (Fig. 2).

Based on research, the most of the enterprises, which have participated in the projects, have their own IT system, and some of them have an inter-organizational IT system. According to above fact, IT infrastructure is constructed above the average.

As shown in the Table 1, managing enterprise pose the first priority to delivery information. On the other side, participating enterprise mainly pose the first priority to information regarding production planning (Fig. 3).

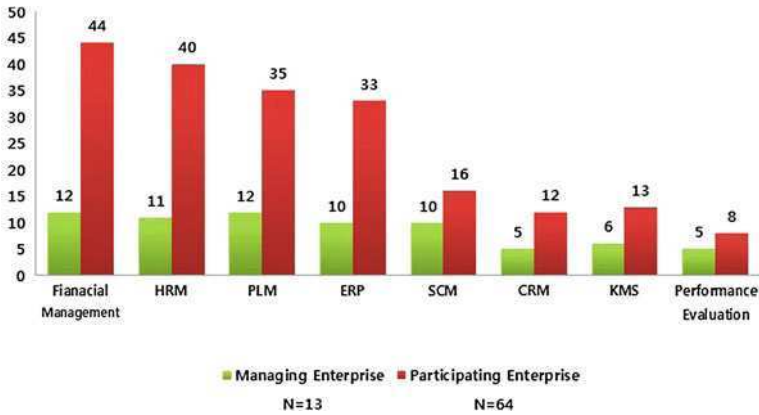


Fig. 2 Status of IT system construction

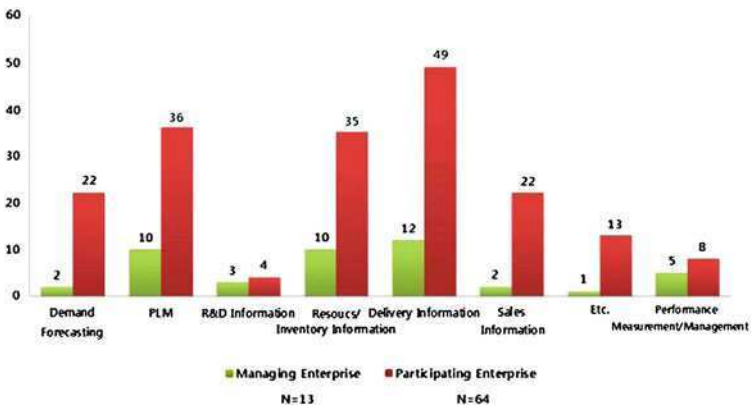


Fig. 3 Range of shared information through IT collaboration system

The information, which became possible to share for managing enterprise, is mainly focused on information about delivery of product, and participating enterprises share information regarding delivery and production planning actively. This tendency is coincident with a priority of IT business collaborating information (Fig. 4).

The information of the certain product, which is shared through the constructed IT system, is primarily related to a critical product of respective enterprise, and in case of participating enterprises, some of them share information of non-critical product (Table 2).

The frequency of transaction shows a similar pattern with the ratio of products manufactured through IT business collaboration system. And the mutual reliability among managing enterprises and participating enterprises was quite high. This tendency attributes to the realization that the risk of information leakage amongst

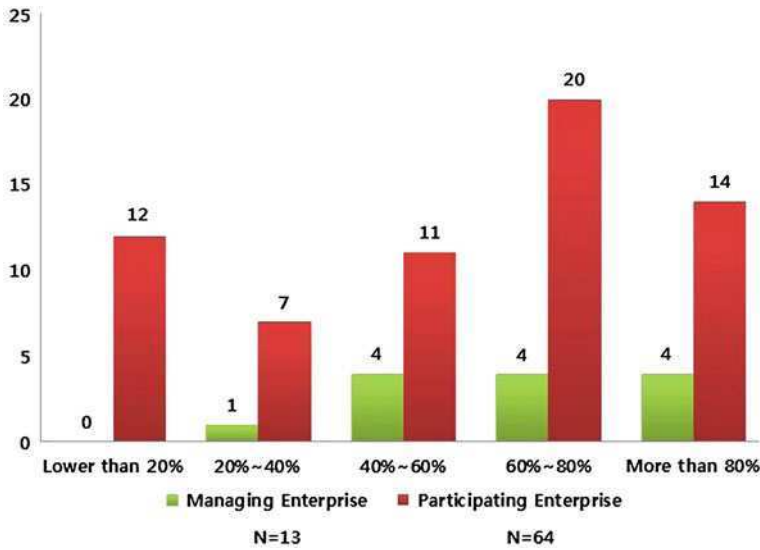


Fig. 4 Range of product manufacture with information sharing through IT collaboration system

Table 2 Analysis of CSF of inter-organizational IT business collaboration

(Unit: number of enterprises)

Area	Item	Very high		High		Average		Low		Very low	
		○	⊙	○	⊙	○	⊙	○	⊙	○	⊙
Frequency of transaction	①	5	9	7	27	1	24	0	1	0	3
Reliability	②	0	7	12	39	1	13	0	3	0	2
Risk of abuse/miss-use of shared information	③	0	0	1	3	2	15	10	37	0	9
	④	0	0	1	5	2	10	7	32	3	17

○ Managing enterprise, ⊙ Participating enterprise

① Frequency of Production through IT Business Collaboration, ② Mutual Reliability, ③ Risk Level of Information Miss-use during IT Business Collaboration, ④ Possibility of Information Leakage during IT Business Collaboration

enterprises might be low. In addition many enterprises believe that the possibility of information leakage during IT business collaboration.

4 Conclusion

‘Collaborative Network Construction Project for Inter-organizational Collaboration based on IT’ has improved a collaborative process and an information integration of entire supply chain by applying IT to Korean industry and it has caused

an improvement of productivity, mutual growth and innovation of enterprises (large and small and medium). In this study, we have conducted an empirical study on CSF of inter-organization IT business collaboration.

As a result of analysis, as a readiness of inter-organization IT business collaboration is high, a performance level is high. Especially, IT sophistication, which includes a level of IT work force and a level of IT system, turned out to be closely related to IT business collaboration performance.

In the future, the analysis of co-relation between IT readiness and IT business collaboration performance and characteristic of collaboration and IT business collaboration performance should be conducted based on a descriptive statistics. Furthermore that analysis results should be utilized as a basic material for constructing a strategy for IT utilization project.

References

1. May P, Ehrlich HC, Steinke T (2006) ZIB structure prediction pipeline: composing a complex biological workflow through web services. In: Nagel WE, Walter WV, Lehner W (eds) Euro-Par 2006, vol 4128, LNCS Springer, Heidelberg, pp 1148–1158
2. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *J Mol Biol* 147:195–197
3. National center for biotechnology information, <http://www.ncbi.nlm.nih.gov>
4. Czajkowski K, Fitzgerald S, Foster I, Kesselman C (2001) Grid information services for distributed resource sharing. In: 10th IEEE international symposium on high performance distributed computing. IEEE Press, New York, pp 181–184
5. Foster I, Kesselman C (1999) *The grid: blueprint for a new computing infrastructure*. Morgan Kaufmann, San Francisco
6. Foster I, Kesselman C, Nick J, Tuecke S (2002) *The physiology of the grid: an open grid services architecture for distributed systems integration*. Technical report, Global Grid Forum

Study of Generating Animated Character Using the Face Pattern Recognition

Seongsoo Cho, Bhanu Shrestha, Bonghwa Hong
and Hwa-Young Jeong

Abstract The similar character of input video image can be generated by combining each components points after extracting the position of each component points using the ratio of face. By selecting the face area, in case of the upper half image, the neck and stained hair become change and accuracy of recognition of selected area becomes low, so in order to compensate the area of inaccuracy, we can find the shape of face, eyes and mouth applying the golden ratio and the character is reflected in the results. Using color informations of the image of the face, the TSL color can be analysed and shown the face recognition ratio of 1.7%. The components in each five input images (left and right eyes, eyebrows and mouth) by selecting the test results showed the performance of 88.3%. By analyzing the characteristics of the elements found in those features automatically generate the appropriate characters, and the rabbit character can be automatically generated from the animation story of 'Rabbit and Tortoise' made by flash tool.

S. Cho (✉) · B. Shrestha

Department of Electronic Engineering, Kwangwoon University,
26 Kwangwoon-gil Nowon-gu, Seoul, 139-701, Korea
e-mail: css@kw.ac.kr

B. Shrestha

e-mail: bnu@kw.ac.kr

B. Hong

Department of Information Communication, Kyunghee Cyber University,
Dongdaemun-gu, Seoul, 130-701, Korea
e-mail: bhhong@khcu.ac.kr

H.-Y. Jeong

Department of General Education, Kyunghee University,
1 Hoegi-dong Dongdaemun-gu, Seoul, 130-701, Korea
e-mail: hyjeong@khu.ac.kr

Using the TSL color model, the face shape and the eyes, eyebrow, nose, mouth, and ears are visualised in a 2D. In the digital cultural content, the character is automatically generated by video input image and oneself can be staged as a hero in the animation which is the development of a new digital cultural contents.

Keywords Face patter recognition · Pattern recognition · Character animation · Auto-character generation

1 Introduction

With the fast development of IT technology and wide use of computers and fast-speed network, many things that have been done in the real world are now performed in the cyber space generated by IT technology [1, 2]. Increased capability, capacity and speed in processing information by computers, improved Tx/Rx of digital information of IT devices, lively exchange of information over international information network like Internet, and integration and multi-functionalization of various types of media accelerate globalization in political and economic terms. In the socio-cultural aspects, they provide the possibility of changing into totally new concepts, such as open society, lifelong study society and global culture. People are surrounded by CCTVs, traffic cameras, phone cameras and web cameras, and tens of thousands of video clips are made every day. Digital contents enable One Source Multi Use (OSMU) in which incidental revenues are made in various forms with a single item. In short, a single content is delivered through various forms of media. Recently, a material produces various contents with various media such as comics, novel, movie and game. Study on human face on computer extends its scope of application from extraction of human face or features of human face and recognition of expression via analysis of general image or video image to various fields such as teleconference and human-computer interaction.

The technology of extraction and tracing of a face is one of the technologies being studied for many application systems, such as locomobile robot, monitoring system and human-computer interface. So far, it is a challenging task to extract and trace facial features in real time. With the development of image processing technique thanks to increasing performance of computer, this technology is one of the subjects being investigated most actively [3]. Active Contour Model (ACM) and the data from Edge are used to as the base technology to extract facial area and features. In extracting facial area, the change of facial area is used as the external energy for ACM, reducing the effect of deterioration of lighting and picture quality of low-resolution pictures [4]. Since this technique requires manual definition of start and end of change of facial expression, however, it requires long production time and it is difficult to generate an animation with realistic facial expression.

2 Related Studies

2.1 RGB Color Model

Many color models have been developed for image processing, and various models are under development. All these models require conversion of Red, Green, Blue (RGB) color mode in order that they should be used for special purposes. It means that additional operations are required to those for image processing, and as a result, the amount of real time operation required for each frame grows exponentially. Therefore, this study aimed at reducing the amount of operation by restricting use of other color models but using the RGB model only. In this study, color data were identified with the ratio of RGB. To reduce the search area, facial tones were defined based on the RGB model. Because the skin color in a picture typically depends on the lighting, skin colors were defined with the color ratio of RGB. The following formula (1, 2) is used to define the area of skin colors.

$$R_{(i,j)} + l > UCCM_{ax}(G_{(i,j)}, B_{(i,j)}) \quad (1)$$

$$\begin{aligned} (R_{(i,j)} > UCCM_{ax}(G_{(i,j)}, B_{(i,j)})) \\ (R_{(i,j)} > m) \end{aligned} \quad (2)$$

In the above formula, R, G and B represent Red, Green and Blue in the pixel (i, j) , respectively, and i, j, m are constants. Normally, skin color has more red than others. Therefore, in this study, red was detected to reduce the search area first.

2.2 Blocking process

The blocking process evaluates the block (normally, square) area in an image, and realign the data. Because the process reflects the data of a specified area, it removes small images (noise). It rounds the border line of an image, and fills up small blanks between pixels [5–7]. Formula (3) describes the blocking process.

$$\sum_{k=i-s}^{i+s} \sum_{k=j-s}^{j+s} \sum (i,j) > a \quad (3)$$

In Formula (3), (i, j) is the (i, j) th pixel, and s is the block size and a is a constant. By adjusting a , the level of blocking is changed. The bigger the value, the amount of loss of the image grows. Figure 1a illustrates the result of blocking. In Fig. 1b, small noises are all removed, and the blank caused due to eyes and eyebrows are filled up. The face is not in a lump. Using this face, labeling can be performed.

Fig. 1 Results of blocking
a Extraction of skin color region
b Method of eliminating a small image (noise) using blocking

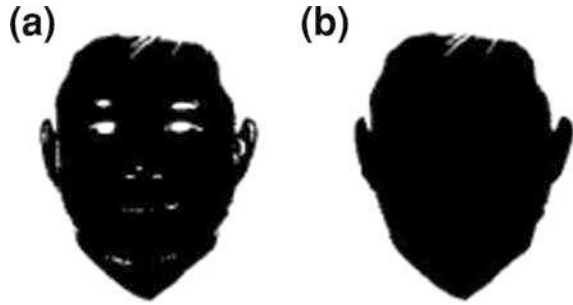
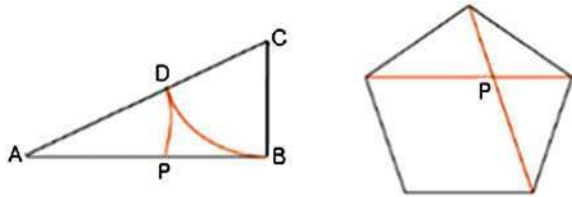


Fig. 2 Example of sampling of golden ratio



2.3 Face golden ratio

Facial areas are found in the above process, but you cannot be sure if all of them are faces. Especially for a bust shot, the distorted image of neck and dyed hair lowers the accuracy of recognition. Therefore, in order to supplement the inaccurate area, the golden ratio is used. When a line segment is split into two by a point, the ratio between the two parts which makes the square of the length of one part to be equivalent to the length of the other part multiplied by the entire length of the line segment is called a Golden Ratio, which is about 1:1.61803.

In the above example in Fig. 2, in the triangle $\triangle ABC$, assume that $AB = 2BC$. On AC , take the point of CD which is equivalent to BC . Then, on AB , take the point AP which is equivalent to AD . The point P is the golden ratio point of AB . The golden ratio was first found by ancient Greeks, and was named as it provides the most well-balanced ratio. Human face also is considered to be most beautiful when it is at the golden ratio. This golden ratio is applied to paintings and comics [8]. The golden ratio is expressed in Formula (4).

$$(AP)^2 = BP \times AB \quad (4)$$

2.4 Extraction of human face

A new color space was created to guarantee accurate extraction results without any additional operation and regardless of input data. The most widely-used CbCr, HS

and TS were broken down to Cb, Cr, H, S, T and tS (Note: ‘tS’ was used in order to distinguish this component from ‘S’ of the HIS color model.), and then, these components were combined to make a new color space. Formula (5) is an equation of a line that passes two points (Max. and Min. of H and T). After converting each pixel of the input image into the H-T color space, in order to extract the skin color only from the input image, the distance between the straight line in Formula (5) and the input image was measured with Formulas (5) and (6).

$$F(h) = \frac{T_{min} - T_{max}}{H_{min} - H_{max}}(h - H_{min}) + T_{max} (H_{min} \leq h \leq H_{max}) \quad (5)$$

$$d(x, y) = |f(x) - c(x, y)| \quad (6)$$

As expressed in the above formulas, the distance between the straight line and the pixel of the input image was measured, and the pixel was considered as the skin area if the value is smaller than the critical value.

3 Result of Experiment

3.1 Result of the Study

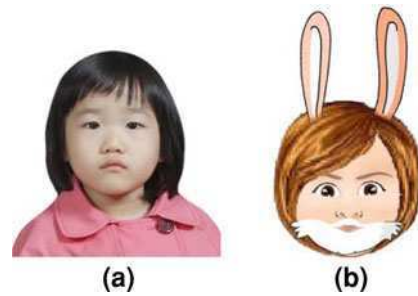
The images used in this experiment were randomly selected from Internet. It is difficult to distinguish sex of the characters on the photos with the image processing only. Therefore, sex distinction was entered manually, and the images of the characters were selected accordingly.

Because the boundary of components of an input image is not determined with one or two pixels only, it is impossible to separate the face area or component in the input image by pixel. Therefore, by maintaining the original video of extracted area, the experiment is evaluated that the existence of desired position of component characters can be extracted. Because there is a restriction that the face area should be located in the center of the input image, out of 60 input images, face recognition through analysis of TSL color was successful for 55 images and failed for 5 images, resulting in the success ratio of 91.7%. For the test of 5 components (left/right eyes, left/right eyebrows and a mouth), recognition was successful for 53 images and failed for 7 images, resulting in the success ratio of 88.3%. The result is 10% increase in face area extraction when compared with the ACM-based method, and 3.4% increase in face component extraction when compared with Kenny Edge-based method.

3.2 Results of Automatic Character Creation

When compared with live films, various characters with clearer and more exaggerated looks and characteristics appear in animated films. Because characters

Fig. 3 The components points recognized character generation results of the original image. **a** Input image **b** Generation of character by detecting the main points



have distinctive roles and personalities, animated files are appropriate for collecting character images of variety of personality. Especially in case of 2D animation, because most characters are expressed with clear lines and simple shapes, unnecessary face components are eliminated. With the processes above mentioned, it was possible to find the face area, and the relative position and size of eyes and mouth in a face area. By analyzing characteristics of the detected components, a character that meets the characteristics was automatically created. A rabbit character from the animation “the Rabbit and the Tortoise” made with Flash was created. Figure 3 shows the result of detection of the components of an input image, and the character created. This character was created with the image that has similar components.

In the character created in Fig. 3, the distances between the feature vectors used to represent a face must be combined as well as other geometric features captured from the animation “the Rabbit and the Tortoise”. The feature vectors acquired from the test images and the images in the database are used to recognize a face. The measurement of similarity between the vectors of the shortest distance is used to determine the uniqueness of the face. Based on the conversion of a real time input image to a character, it is possible to create a new character-composed animation and to extract the face recognition information. Overall key frame animation flow can be acquired by substituting animation frames between each other, rather than using a single image.

4 Conclusion

The 3 min and 25 s 2D Flash animation featuring ‘Rabbit and Tortoise’ was made with main key frames where input image is converted into the Rabbit character. Users input face images to create the desired character of the animation. This technology can be used as the basic module in a wide range from Internet-based technology, e-book to 3D simulation games. The proposed technology will enhance the existing 2D effect to the 3D effect, providing more realistic and effective animation effect. The current technology cannot automatically create or create movement of objects, but under certain rules, it may give movement of

certain form with certain level of change. The resulted character may be used on Internet and in personal devices. It is expected that more diverse characters are implemented if feather analysis is performed in various ways and more original character images are added. One of the subjects of the future studies is to enable users to make 3D animations by modeling images as the like. For this purpose, it is required to develop the 3D animation encoding technology to enhance the memory and transmission time and the processing speed. Sound function must be developed to support the multimedia function.

References

1. Parke FI (1982) Parameterized models for facial animation. *IEEE Comput Graphics Appl* 2(9):61–68
2. Lee Y, Terzopoulos D, Waters K (1995) Realistic modeling for facial animation. In: *Proceedings of SIGGRAPH95 computer graphics*, pp 55–62
3. Fong T, nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. *Robot Autonom Syst* 42
4. Hsu R-L, Abdel-Mottaleb M, Jain AK (2002) Face detection in color images. *IEEE Trans Patt Anal Mach Intell* 24(5):696–706
5. Lee L-W, Wang J-F, Lee J-Y, Shie j-D (1993) Dynamic search-window adjustment and interlaced search for block-matching algorithm. *IEEE Trans Circuits Sys Video Tech* 3(1):85–87
6. Feng J, Lo K-T, Mehrpour H, Karbowski AE (1995) Adaptive block matching motion estimation algorithm for video coding. *Electro Lett* 32(8):1542–1543
7. Liu B, Zaccarin A (1993) New fast algorithms for the estimation of block motion vectors. *IEEE Trans Circuits Sys Video Tech* 3(2):148–157
8. <http://blog.naver.com/imbc21c?Redirect=Log&logNo=30049372282>

Enhancing Performance of Mobile Node Authentication with Practical Security

Kyusuk Han and Taeshik Shon

1 Introduction

In this paper, we enhance our untraceable mobile sensor node protocol [1] that reduce computation overhead, and also satisfies practical security strength.

1.1 Untraceable Mobile Node Authentication

In this section, we briefly describe procedures of [1] as in Fig. 1. There are a base station BS , a sink S_1 , a neighbor sink S_2 , and a mobile node N in the network. We define the neighbor sink as the sink that is in the 1 hop communication range. S_1 periodically broadcasts HELLO in Phase 0. When S_2 receives HELLO, S_2 initiates the neighbor relationship if S_1 is a newly discovered sink. After the pairwise key between S_1 and S_2 has been exchanged in Phase 1, S_1 and S_2 exchange the authentication key that is used to verify the authenticated user in Phase 2. Phase 1 and Phase 2 are only required during establishing the static sensor network. We let the establishing the static sensor network follows the any previous protocol such as [2].

When N firstly joins the network, N may be connected to S_1 in the network as in Fig. 1. After receiving HELLO of S_1 , N initiates the initial authentication with S_1 in Phase 3. After N is authenticated S_1 , N only needs the re-authentication in Phase

K. Han

Information and Communications University, Daejeon, Republic of Korea

T. Shon (✉)

Division of Information and Computer Engineering, Ajou University,

Gyeonggido, Suwon 443-749, Republic of Korea

e-mail: Taeshik.shon@gmail.com

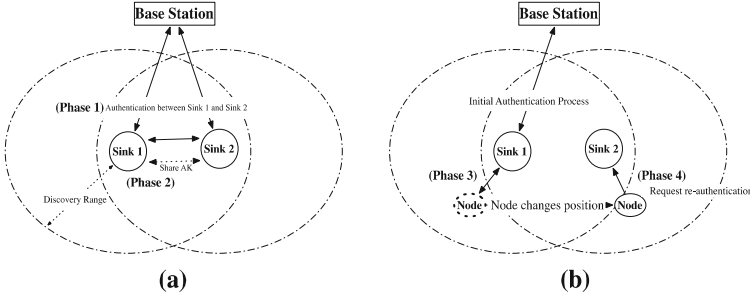


Fig. 1 Protocol overview: in receiving HELLO of Sink 2 (S_2), **a** Sink 1 (S_1) mutually authenticates Sink 2 (*Phase 1*), and share the authentication key (*Phase 2*). **b** Node is initially authenticated by Sink 1 (*Phase 3*), and requests re-authentication to Sink 2

4 when N continuously moves and request the authentication again. The authentication process in Phase 3 is only necessary when the re-authentication fails due to the certain case that the neighbor sink is not available.

1.2 Authentication Ticket

In previous work, we defined ‘Authentication Ticket’ that is used for the node re-authentication. When a node requests authentication to a sink, the sink generates the authentication ticket and sends it to the node. The authentication key that is given to neighbor sinks can verify the authentication ticket. Using the authentication ticket, the node movement is untraceable. Verification of the authentication ticket is available to neighbor sinks of the sink that issued the ticket. We adopted the idea of ‘cluster key’ in [3] that shared to neighbor sinks. The main difference is that the cluster key in [3] is used for broadcast communication in the cluster, while the key in our protocol is used for verifying the authentication ticket. Thus, we rename the key as ‘authentication key’ due to the different use in the protocol.

2 Previous Protocol

Previous protocol consists of five phases as follows: **Phase 0** The common neighbor discovery, **Phase 1** Neighbor sink relationship set up, **Phase 2** Neighbor group authentication key share, **Phase 3** Initial node authentication, and **Phase 4** Node re-authentication.

The notations used in the protocol are defined in Table 1. Key IK_N is the integrity key derived from K_N , where $IK_N = KDF(K_N)$. KDF is a one-way key derivation function. We can also use a hash function for KDF .

Table 1 Notations

Term	Description
BS	Base station
$h\{m\}$	Hash arbitrary message m
TS	Time stamp
IK_N	IK derived from K_N
IK_S	IK derived from K_S
SIK	IK derived from SK
AIK_S	IK derived from AK_S
NIK	IK derived from NK
$E_t\{m\}$	Encrypt arbitrary message m using t
$MAC_t(m)$	Message authentication code using t
K_N	Pre-shared key between N and BS
K_S	Pre-shared key between S and BS
SK	Shared session key between sinks
AK_S	Group authentication key of sink
NK	Shared session key between S and N
IK	Integrity key

2.1 Phase 0: Neighbor Discovery

A sink S_1 periodically generates a random nonce R_0 . S_1 also generates $u_0 = E_{K_{S_1}}\{R_0||TS_0\}$ and $v_0 = MAC_{IK_{S_1}}(S_1||HELLO||u_0)$, where TS_0 is time stamp. u_0 and v_0 are included in the HELLO message. Then S_1 broadcasts u_0 and v_0 as follows:

$$S_1 \rightarrow \text{Broadcast} : S_1||HELLO||u_0||v_0$$

Phase 0 is the periodical common procedure. When a sink receives HELLO, the sink initiates Phase 1 or Phase 2. When a node receives HELLO, the node initiates Phase 3 or Phase 4.

2.2 Phase 1: Neighbor Sink Relationship Set Up

Assume another sink S_2 receives HELLO message. S_2 checks the sender of HELLO whether S_1 is known or not. If S_2 already knows S_1 , S_2 discards the message. Otherwise, S_2 requests the setting up the neighbor relationship as follows:

P-1.a. S_2 randomly selects R_1 and generates $u_1 = E_{K_{S_2}}\{R_1||u_0\}$, and $v_1 = MAC_{IK_{S_2}}(S_2||BS||S_1||u_1||v_0)$.

$$S_2 \rightarrow BS : S_2||BS||S_1||u_1||v_1||v_0$$

P-1.b. After verifying v_1 , BS decrypts u_1 and retrieves R_1 and u_0 . Then, BS verifies v_0 and decrypts u_0 . Finally, BS retrieves R_0 and TS_0 . BS generates and sends u_4 , v_4 , and v_3 to S_2 where, $u_3 = E_{K_{S_1}}\{R_1||h(TS_0)\}$, $v_3 = MAC_{IK_{S_1}}(BS||S_1||u_3)$, $u_4 = E_{K_2}\{R_1||u_3\}$ and $v_4 = MAC_{IK_2}(BS||S_2||R_1||u_4||v_3)$.

$$BS \rightarrow S_2 : BS||S_2||S_1||u_4||v_4||v_3$$

2.3 Phase 2: Neighbor Group Authentication Key Share

Phase 2 can be operated solely or after Phase 1 is completed. In Phase 2, S_1 initiates following procedures.

P-2.a. S_1 randomly selects two nonce $ASEED_{S_1}$ and R_1 . Then S_1 generates $u_1 = E_{K_{S_1 S_2}}\{ASEED_{S_1}||R_1\}$ and $v_1 = MAC_{IK_{S_1 S_2}}(S_1||S_2||u_1)$

$$\setminus[\{\{S\}_{-1}\}\setminus\{\{S\}_{-2}\} : \{\{S\}_{-1}\}||\{\{S\}_{-2}\}||\{\{u\}_{-1}\}||\{\{v\}_{-1}\}\setminus].$$

2.4 Phase 3: Initial Node Authentication

When N receives HELLO that S_1 broadcasts in Phase 0 and is not yet authenticated by any sink, N proceeds followings.

P-3.a. N randomly selects R_1 and generates $u_1 = E_{K_N}\{R_1||u_0||v_0\}$ and $v_1 = MAC_{IK_N}(N_1||S_1||u_1)$.

$$N \rightarrow S_1 : N||S_1||u_1||v_1$$

P-3.b. S_1 generates $v_2 = MAC_{IK_{S_1}}(S_1||BS||N||u_1||v_1)$.

$$S_1 \rightarrow BS : S_1||BS||N||u_1||v_1||v_2$$

2.5 Phase 4: Node Re-authentication

When N receives HELLO that S_2 broadcasts in Phase 0 and is previously authenticated by a sink, N proceeds followings.

P-4.a. N generates $v_1 = MAC_{NK_N}(N||S_2||t||w||v_0)$.

$$N \rightarrow S_2 : N||S_2||t||w||v_1$$

P-4.b. S_2 verifies w and decrypts t . S_2 retrieves R_1 , NK_N and TS . Using NK_N , S_2 verifies v_1 . Then S_2 generates $NK' = KDF(R_1||R_0)$, also generates $t' = E_{AK_{S_2}}$

$\{R_1||NK'_N\}$ and $w' = MAC_{AIK_{S_2}}(N||t')$. S_2 generates $v_2 = h(NK'_N||R_0)$ and $u_3 = E_{NK_N}\{R_0||v_2||t'||w'\}$, $v_3 = MAC_{NK_N}(S_2||N||u_3)$.

$$S_2 \rightarrow N : S_2||N||u_3||v_3$$

3 Analysis

In revised protocol, sink doesnot broadcast encrypted random nonce. Such reduction could weaken theoretical security strength of proposed protocol, yet still secure in practical environments. Also, proposed protocol reduced additional computation and communication overheads by reducing R_0 .

4 Conclusion

In this article we proposed a modified model of [1] for improvement of efficiency. Although such reduction may need more clear security analysis, proposed protocol still support practical security strength.

References

1. Han K, Kim K, Shon T (2010) Untraceable mobile node authentication in WSN. Sensors 10(5):4410–4429
2. Ibriq J, Mahgoub I (2007) A hierarchical key establishment scheme for wireless sensor networks. In: Proceedings of 21st international conference on advanced networking and applications (AINA'07), pp 210–219
3. Zhu S, Setia S, Jajodia S (2006) LEAP+: efficient security mechanisms for large-scale distributed sensor networks. ACM Trans Sen Netw 2:500–528

A Study on Turbo Coded OFDM System with SLM for PAPR Reduction

Mashhur Sattorov, Sang-Soo Yeo and Heau-Jo Kang

Abstract Orthogonal frequency division multiplexing (OFDM) technique is a promising technique to offer high data rate and reliable communications over fading channels. The main implementation disadvantage of OFDM is the possibility of high peak to average power ratio (PAPR). This paper presents a novel technique to reduce the PAPR using turbo coding and selective mapping (SLM). We show that the probability of the PAPR of OFDM signal with 512 subcarriers in a 16-QAM channel can be reduced by 25% in a Turbo selector. The rest of the paper, simulation results are superimposed by using Matlab interface as well.

Keywords OFDM · PAPR · Turbo coding · SLM

Sang-Soo Yeo—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 (2011-0014394).

M. Sattorov · S.-S. Yeo · H.-J. Kang (✉)
Division of Computer Engineering, Mokwon University,
Deajeon, Republic of Korea
e-mail: hjkang@mokwon.ac.kr

M. Sattorov
e-mail: mashhurs@yahoo.com

S.-S. Yeo
e-mail: sangsooyeo@gmail.com

1 Introduction

Orthogonal frequency division multiplexing (OFDM), is a multicarrier communication technique, where a single data stream is transmitted over a number of lower rate subcarriers. OFDM has become tangible reality, it has been employed for wire-line communications and also has been employed in wireless local area network (WLAN) e.g. IEEE 802.11. Other applications of OFDM are digital audio broadcasting (DAB) and digital video broadcasting (DVB).

Unfortunately, OFDM has the drawback of a potentially high peak to average power ratio (PAPR). Since a multicarrier signal consists of a number of independent modulated subcarriers that can cause a large PAPR when the subcarriers are added up coherently.

To reduce the PAPR different techniques were proposed. These techniques can be categorized into the following, clipping and filtering [1], coding [2], phasing [3], scrambling [4], interleaving [5], and companding [6].

In this paper we propose and examine a technique for reducing the probability of a high PAPR, based on part on a method proposed in [7, 8]. This technique is a variation of selective mapping (SLM) [7], in which a set of independent sequences are generated by some means from the original signal, and then the sequence with the lowest PAPR is transmitted. To generate these sequences we use turbo encoder. Using turbo coding will offer two advantages, significant PAPR reduction and astonishing bit error rate (BER) performance.

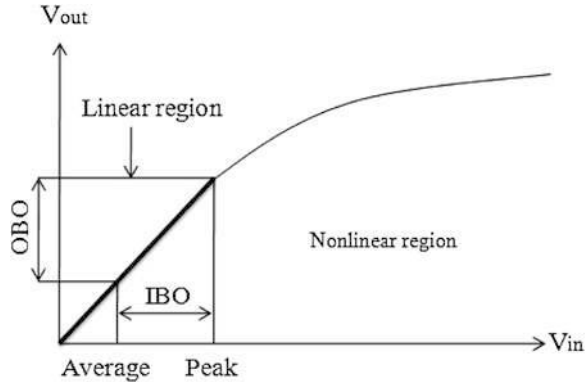
The rest of the paper is organized as follows: The problem of high PAPR of OFDM signal is briefly defined in [Sect. 2](#). [Section 3](#) introduces the proposed technique. Some simulation results are superimposed in [Sect. 4](#). Finally, the conclusions are drawn in [Sect. 5](#).

2 PAPR Problem Definition

When transmitted through a nonlinear device, such as a high-power amplifier (HPA) or a digital to analog converter (DAC), a high peak signal generates out-of-band energy and in-band distortion. These degradations may affect the system performance severely. The nonlinear behaviour of an HPA can be characterized by amplitude modulation/amplitude modulation (AM/AM) and amplitude modulation/phase modulation (AM/PM) responses. [Figure 1](#) shows a typical AM/AM response for an HPA, with the associated input and output back-off regions (IBO and OBO, respectively).

To avoid such undesirable nonlinear effects, a waveform with high peak power must be transmitted in the linear region of the HPA by decreasing the average power of the input signal. This is called (input) backoff (IBO) and results in a proportional output backoff (OBO). High backoff reduces the power efficiency of the HPA and may limit the battery life for mobile applications. In addition to

Fig. 1 A typical power amplifier response



inefficiency in terms of power, the coverage range is reduced, and the cost of the HPA is higher than would be mandated by the average power requirements. The input backoff is defined as:

$$IBO = 10 \log_{10} \frac{P_{insat}}{\overline{P}_{in}} \tag{1}$$

Where P_{insat} is the saturation power, above which is the nonlinear region, and \overline{P}_{in} is the average input power. The amount of backoff is usually greater than or equal to the PAR of the signal. The power efficiency of an HPA can be increased by reducing the PAR of the transmitted signal. Clearly, it would be desirable to have the average and peak values are as close together as possible in order to maximize the efficiency of the power amplifier. In addition to the large burden placed on the HPA, a high PAR requires high resolution for both the transmitter’s DAC and the receiver’s ADC, since the dynamic range of the signal is proportional to the PAR. High-resolution D/A and A/D conversion places an additional complexity, cost, and power burden on the system.

3 SLM Using Turbo Coding

The probability that, the PAPR of the OFDM signal exceeds a certain threshold γ is given by

$$\Pr\{PAPR > \gamma\} = 1 - (1 - e^{-\gamma})^N \tag{2}$$

$N = 2.8 * N$, when the system uses a large number of N , the [9]. In SLM it is assumed that, U statistically independent alternative sequences, which represent the same information, are generated by some suitable means. The sequence with the lowest PAPR is selected for transmission. The probability that, the lowest PAPR γ_l exceeds a certain threshold γ is given by

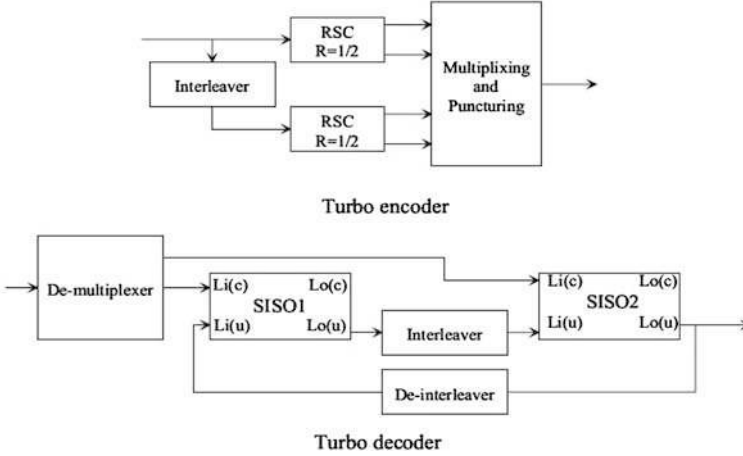


Fig. 2 Turbo system

$$\Pr\{\gamma_l > \gamma\} = (\Pr\{PAPR > \gamma\})^U. \quad (3)$$

To generate these sequences linear feedback shift register (LFSR) is used [10]. A LFSR is used to transform the data before it is mapped to the orthogonal channels. Different sequences are generated by inserting different bits labels at the beginning of the data. This results $U = 2^m$ different sequences, where m is the length of the inserted bits. Turbo codes [11] are parallel concatenated convolutional codes in which the information bits are first encoded by a recursive systematic convolutional (RSC) code and then, after passing the information bits through an interleaver, are encoded by a second RSC code. Turbo decoder is used to recover the transmitted signal at the receiver side. The Turbo decoder consists of two soft input soft output (SISO) modules [12], an interleaver and de-interleaver. Figure 2 shows a turbo system, turbo encoder and decoder.

In this paper, instead of using LFSR, we use turbo encoder to generate different sequences and the sequence with the lowest PAPR is selected for transmission. The different sequences are generated by inserting different bits labels at the beginning of the data. The Fig. 3 can be an example for PCCC Turbo Code system that uses 16-QAM Modulation with code rate of 1/3 and in a simulation results section, the PAPR of this system is simulated.

Figure 4 shows the transmitter of an OFDM system, where the turbo coding and SLM are used for PAPR reduction. For each bits labels b_i , $i = 1, 2, \dots, U$, where b_i a sequence of m bits, the turbo encoder will generate a sequence x_i , $i = 1, 2, \dots, U$. The sequence that has the lowest PAPR will be selected for transmission. At the receiver side, the receiver does not need any side information, and the bits labels are discarded after decoding.

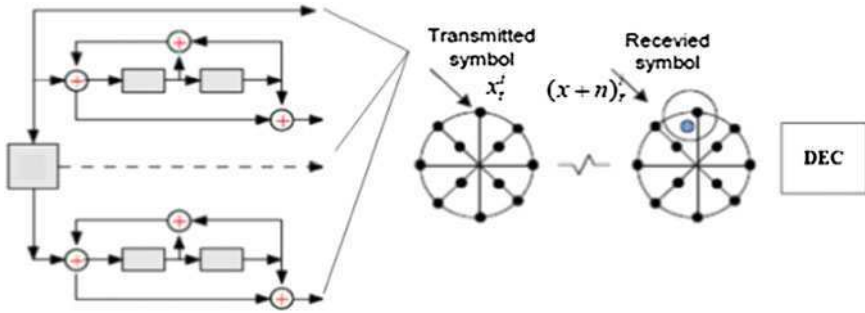


Fig. 3 A rate 1/3 parallel concatenated convolutional code turbo coding in a 16-QAM channel

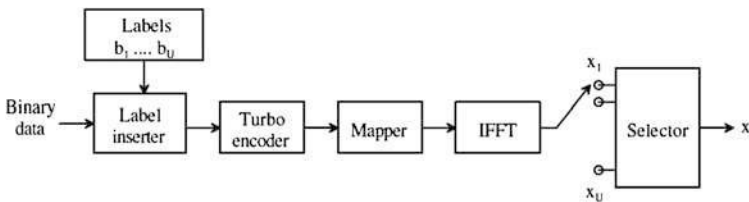


Fig. 4 System model

4 Simulation Results and Discussions

The PAPR reduction and BER performances of the proposed scheme are examined by computer simulation. In the simulation we consider an OFDM signal with $N = 512$ subcarriers, quadrature amplitude modulation (16-QAM) mapping. Puncturing is used to increase the overall code rate to $R = 1/2$. And also AWGN channel is assumed. At the receiver side Soft output Vitebri Algorithm (SOVA) is used to implement the SISO modules.

Figure 5 shows the probability of PAPR at given threshold. It gives us a great feature that selector can be in range of $0.86 * PAPR_{\gamma} \leq PAPR_{\gamma_i} \leq 0.25PAPR_{\gamma}$.

Parameters for simulation: (Fig. 6)

Frame size = 100

Code generator:

1	1	1
1	0	1

Punctured, code rate = $1/2$

Iteration number = 3

Terminate frame errors = 5

E_b/N_0 (dB) = 2.00

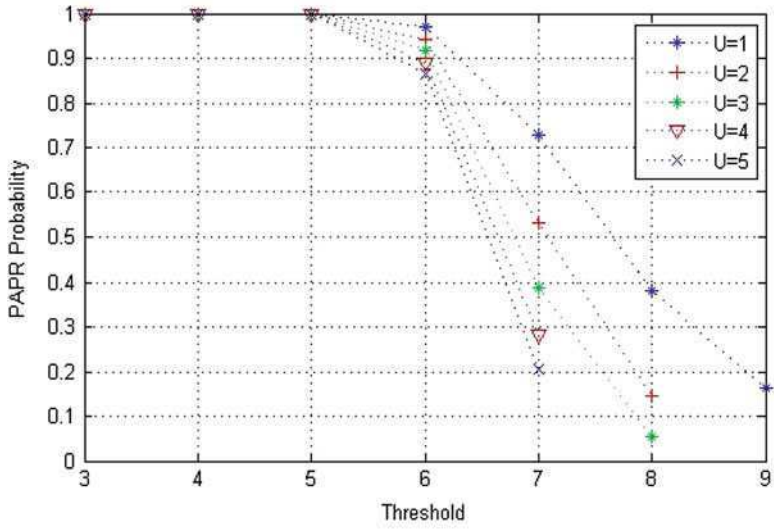


Fig. 5 PAPR probability in selector

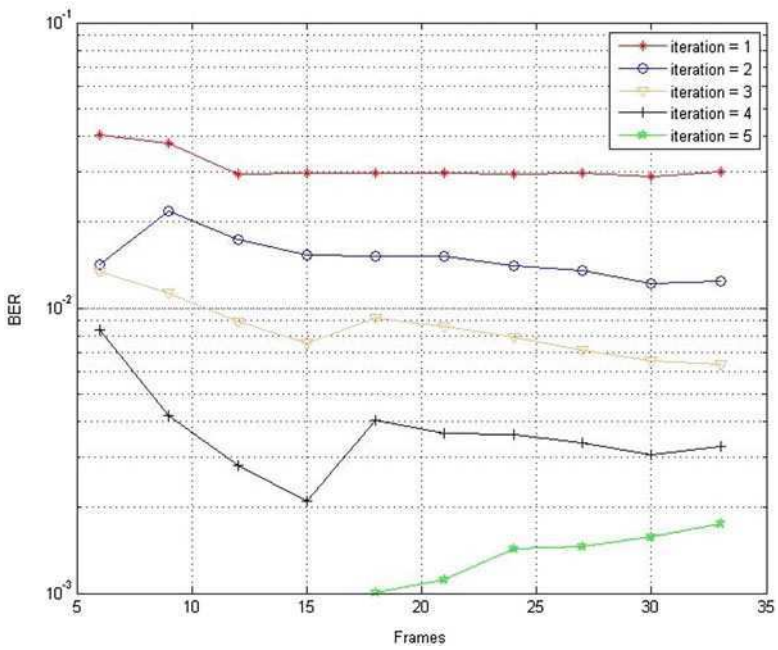


Fig. 6 BER performance

In our turbo system, the frame BER gradual decrease condition can be seen when the iteration values rise. For instance, in generated frame 41 the frame BER is less than transmitted symbol BER. This condition could be easily seen in Fig. 6 that provides to draw BER performance. From this illustration, it can be noticed that our simulation maintains better result at four iterations. However, even though five iteration is the reason for BER increasing, its performance is better than others.

From simulations and theories given in this manuscript, we can say that as everything has a pros and cons, Turbo system proposed to reduce PAPR is also not an exception. The advantage sides of system are PAPR and BER reduction technique. Opposite to these, a time consuming and hard implementable features are main disadvantages of Turbo coding technique.

5 Conclusion

We have shown that Turbo coding and SLM can be combined to reduce the PAPR of OFDM signal with quite moderate additional complexity with our experiments. The advantage of the proposed scheme is that the Turbo encoder is used for two purposes, error correction and PAPR reduction. Even though it raises the hardware complexity of the system, it can be considered as an effective model for certain communication environments.

References

1. Li X, Cimini LJ Jr (1998) Effects of clipping and filtering on the performance of OFDM. *IEEE Commun Lett* 2:131–133
2. Jones AE, Wilkinson TA, Barton SK (1994) Block coding scheme for reduction of peak to mean envelope power ratio of multicarrier transmission schemes. *Electr Lett* 25:2098–2099
3. Tarokh V, Jafakhani H (2000) On the computation and reduction of peak to average power ratio in multicarrier communications. In: *Proceedings of the IEEE 53rd vehicular technology conference 2*, pp 37–44
4. Van Eetvelt P, Wade G, Tomlinson M (2000) Peak to average power reduction for OFDM schemes by selective scrambling. *Electr Lett* 32:1963–1994
5. Jayalath ADS, Tellambura C (2000) The use of interleaving to reduce the peak to average power ratio of an OFDM signal. In: *Global Telecommunication Conference, IEEE*, 1, pp 82–86
6. Huang X, Lu J, Chuang J, Zheng J (2001) Companding transform for the reduction of peak to average power ratio of OFDM signal. *IEEE Trans Commun* 48:835–839
7. Bäuml R, Fischer R, Huber J (1996) Reducing the peak to average power ratio of multicarrier modulation by selected mapping. *Electr Lett* 32:2056–2057
8. Carson N, Gulliver TA (2002) PAPR reduction of OFDM using selected mapping, modified RA codes and clipping. *Proc IEEE VTC* 2:1070–1073
9. Richard N, Ramjee P (2000) *OFDM for wireless multimedia communications*, Boston, London, pp 120–122

10. Breiling M, Müller-Weinfurter S, Huber J (2000) Peak-power reduction in OFDM without explicit side information. In: Proceedings of the 5th International OFDM Workshop, Germany, pp 28.1–28.4
11. Berrou C, Glavieux A, Thitimajshima P (1993) Near Shannon limit error correcting coding and decoding: turbo codes. In: Proceedings of the ICC'93, Geneva, pp 1064–1070
12. Benedetto S, Montorsi G, Divsalar D, Pollara F (1996) A soft-input soft-output maximum a posteriori (MAP) module to decode parallel and serial concatenated codes. TDA progress report, Jet Propulsion Lab, Pasadena, pp 42–127

A Context Information Management System for Context-Aware Services in Smart Home Environments

Jong Hyuk Park

Abstract In recent years, context-aware systems that exploit a variety of context created in the smart home environment to provide information and services to users have been studied. Existing context-aware systems determine the environment's state (context) based on isolated context data from individual devices or sensors. In addition, there has been little attempt to infer the inhabitants' needs or preferences from the sensed context. This paper examines the characteristics of context recognition based on the context information created in the smart home environment and suggests a way to provide context-aware services using the inhabitants' behavior patterns. The context information management system proposed in this paper analyzes context information to identify patterns, conflicts and faults in a smart home environment.

Keywords Behavior pattern · Context-aware · Context information · Context-aware service · Smart home network

1 Introduction

In the smart home networking environment, recognizing and using different devices in the network is a critical technology. Early stage research on smart home networking focused on providing compatibility between heterogeneous middleware to connect a variety of entities in the network. This research enabled users to

J. H. Park (✉)
Department of Computer Science and Engineering,
Seoul National University of Science and Technology,
172 Gongreung 2-dong,
Nowon-gu, Seoul 139-743, Korea
e-mail: parkjonghyuk1@hotmail.com

manage various devices and applications (services) in smart home environments in a convenient and flexible manner [1]. Nowadays, more advanced smart home applications such as situation information management, adaptive self-configuration management, and automated failure management are being studied. One of the essential technologies needed to build such advanced systems is context-awareness, perceiving the current state of the environment based on context data produced in the smart home network. Context awareness that infers the user's preferences or intentions to provide services best suited to the user has been studied in many different fields [2, 3]. In the smart home environment, the context is recognized based on data produced by a variety of household devices and sensors. In general, devices and sensors in the smart home environment have their own way to represent context data. This impedes producing comprehensive context information through the integration of context data from different sources. Previous works in [2–4] perceive the context of the environment based on isolated context information sensed by individual devices or sensors. For the acquisition of integrated context, a flexible and scalable framework that supports the addition of new devices or services is needed. In addition, previous research on context-awareness in the smart home environment was environment-oriented rather than user-oriented [4]. That is, previous works focused on context related to physical environment and relatively neglected human factors related context. It is important to study people-centered context information in order to provide services that add more convenience for smart home users, such as situation information management, adaptive self-configuration management and automated fault management. This paper proposes a context information management system that addresses the shortcomings of the conventional context-aware applications in the smart home environment.

The rest of this paper is organized as follows. [Section 2](#) presents related work and discusses how context in the smart home environment has been handled in previous work. [Section 3](#) describes uniform context generation in smart home environments. In [Sect. 4](#), the proposed context information management system is presented. [Section 5](#) states the experiments performed to evaluate the proposed system in a simulated environment. Finally, conclusions and future research directions are given in [Sect. 6](#).

2 Related Work

With the advance of context-awareness technologies in the smart home environment, several systems that infer the user's needs and environmental state from context data and provide information or services adapted to the user's current situation have been proposed. Currently, methods that provide intelligent services based on the user's personal profile, temporal context, and the user's intentions and emotional state are being investigated. A smart home environment called the 'adaptive house' was implemented by a research group at Colorado

University [2]. The adaptive house observes the occupancy patterns and desires of the inhabitants using the sensors installed in the house and adapts itself to the lifestyle of the inhabitants. The Aware Home constructed at Georgia Institute of Technology [3] has the ability to sense and understand contextual information about the house, surrounding environments and inhabitants. In this work, a practical smart home model was suggested by clearly specifying residential information to be recognized. In previous work, frameworks that provide a means to use context in smart home settings were studied. In these frameworks, the context information was not considered in an integrated manner, resulting in poor flexibility and scalability (i.e., it is difficult to add new devices or services). To provide more relevant services to the smart home user, context-aware technologies that recognize the current environmental state and the inhabitants' needs are utilized. While the configuration of the smart home network is continually changing in connection with technological advances, previous context-aware solutions are not scalable. This gives rise to difficulties in adding or updating smart home devices or sensors [5]. The overall context of the environment can be recognized by integrating context information from different devices and sensors. The problem is that various household devices (TV, refrigerator, etc.) and sensors (temperature, humid, etc.) in the smart home network have their own way to create context data. Thus, the overall context information cannot be obtained by simply integrating context data from different sources. A mechanism to integrate context information in different forms is needed. There are regularities (patterns) in the inhabitants' behavior. These patterns can be exploited to provide intelligent, adaptive smart home services that meet the inhabitants' needs. In addition, a certain pattern in the inhabitants' behavior might cause failures in the smart home environment. When a device failure or a service conflict occurs, the related user pattern can be filed as a cause for future reference.

3 Context Generation in the Smart Home Environment

In this paper, context is defined as "information that can be used to characterize the situation of an entity in the smart home environment." Context information from devices and sensors in the smart home network is represented in a certain way. When devices use different methods to represent their current state or context in the environment, it is hard to integrate context information from different sources. To address this problem, this paper employs the 5W1H mechanism providing a uniform context representation, as shown in Table 1.

In the current smart home environment, the amount of context information created by the devices, sensors, and applications is not sufficient to be represented in the 5W1H. Thus, their context is represented in 4W1H, excluding the property 'Why'. Uniform, comprehensive context can be obtained by integrating 4W1H context from different sources.

Table 1 Context representation

Classifier	Description
Where	Location where the action occurs
Who	The actor (user or application) that activates the action
What	The action that the actor (user or application) performs
When	The time when the action occurs
How	A description about the action (What)
Why	The actor's behavior pattern associated with the action occurred

4 Context Information Management System (CIMS)

The Context Information Management System (CIMS) proposed in this paper collects and manages information like user patterns and conflicts/faults based on the contextual data created in the smart home environment. Rule-based scripts are adopted to increase flexibility and scalability. The system recognizes the state (context) of the environment by integrating different contextual data from different devices and sensors. The system predicts the inhabitants' needs using patterns in the inhabitants' behavior. Figure 1 shows the overall architecture of the proposed system, consisting of a number of modules. The major modules are Context Integrator, Context Manager, Pattern Manager and Situation Manager.

The Context Integrator collects contextual data from different devices and sensors and integrates the collected data to create complete 4W1H context. As described earlier, devices and sensors in the smart home environment produce context information in diverse forms. Integrating context data from different sources to perceive the overall environmental state requires a significant programming effort because devices' source code must be updated whenever a new device or a new integration rule is added. To provide flexibility and scalability, the Context Integrator uses rule-based scripts for context integration. Whenever new context is created in the smart home environment, the Context Integrator verifies that this new context can be incorporated into the existing context. If the condition for integration is satisfied, the new context is integrated according to the integration rules.

Table 2 shows an example of context integration rules. Rule1 integrates context related to the user's action. Rule 2 integrates context related to the application's action. The Context Manager analyzes the 4W1H context created by the Context Integrator to identify any conflicts or faults. If a conflict or fault is detected, this information is stored in the database. Otherwise, the context is passed to the Pattern Manager and Situation Manager. The Pattern Manager retrieves the context stored in the database and finds patterns for each user or application (service). The identified pattern is made up of single context or integrated context of the data from several context sources. If new context is similar to any existing pattern, the matching pattern is given to the 'Why' property of the context. This facilitates finding the user's or application's device usage pattern. The Situation Manager

Fig. 1 CIMS overview

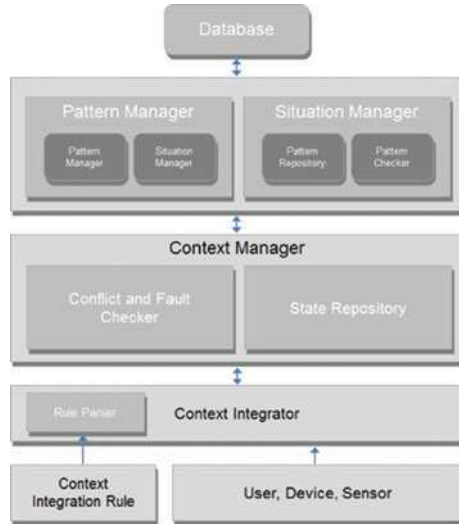


Table 2 Example of integration rules

Rule1	If(clock.time and User.action and radio.power) then Integrate (clock, User1, radio) Integrate (clock, radio) Integrate (clock, User1)
Rule2	If (clock.time and light.power and Service.action) then Integrate (radio, User2)

analyzes house occupancy patterns using the user’s patterns. It is composed of two submodules: one for collecting the context and the other for performing pattern comparison. The Situation Manager monitors the context newly created in the smart home environment and checks its similarity to the patterns in the database. The similarity comparison results are stored in the database.

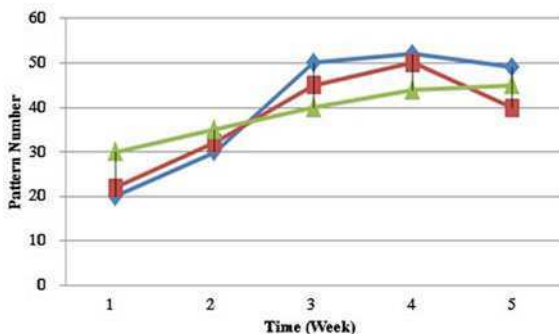
5 Experiment

A simulator was built to demonstrate that the proposed CIMS manages a variety of context in the smart home environment and uses the context to find patterns and conflicts/faults. Multiple users and devices were populated in a way that the simulator has an environment similar to the real-world smart home environment. Table 3 shows the entities in the simulator.

In the experiments, diverse user behavior patterns were created to test the proposed CIMS using the simulator. For example, one pattern used in the experiments is described as follows: “The user enters Room 1 and turns on the light. After watching the television, the user turns off the light and moves to the living room.”

Table 3 Entities in the simulator

Classifier	Entity type
Space	Room 1, room 2, living room, boiler room
Device	Air conditioner, boiler, gas range, electric lights, TV, router
Sensor	Temperature sensor, humid sensor, user location sensor (motion detector)

Fig. 2 Number of patterns for the user

The created patterns were implemented in the batch file in terms of batch items. In the experiments, the number of generated patterns was monitored by varying the number of batch items for the user in the batch file (Fig. 2).

6 Conclusion

Context-aware systems that anticipate the user's needs and provide services suited to the user needs play a key role in increasing the applicability and efficiency of the smart home. Despite the fact that a considerable amount of research has been performed on context awareness in smart home environments, the context information produced by different devices and sensors in the environment is not considered in an integrated manner. That is, previous works rely on isolated context data created by individual devices in recognizing the overall state (context) of the environment. In addition, context data from various devices and sensors is typically in diverse forms of representation, which makes conventional context-aware systems weak in terms of flexibility and scalability.

This paper proposes a way to incorporate information about the context to provide context-aware services to the smart home inhabitants. To improve flexibility and scalability, a context representation (the 5W1H mechanism) that facilitates the integration of the context from a variety of devices is suggested. The CIMS proposed in this paper identifies various context data available in the smart home environment and defines the integration rules by considering the relationship between context sources. The CIMS uses the defined rules to produce the

integrated 4W1H context. The produced context is used to identify patterns and conflicts/faults, which, in turn, are used to produce information or services adapted to the current state of the environment.

In the future, a way to make the CIMS interact with other smart home applications will be examined and exploiting context information for fault prevention, rather than fault detection provided in the current version of CIMS, will be studied.

References

1. Kim M, Kim S (2006) A Scenario-based user-oriented integrated architecture for supporting interOperability among heterogeneous home network middlewares. ICCA2006, pp 669–678
2. Mozer MC (1999) An intelligent environment must be adaptive. IEEE Intell Syst Appl 14:11–13
3. Dey AK, Salber D, Abowd GD (1999) A Context-based infrastructure for smart environments. In: Proceedings of the 1st international workshop on managing interactions in smart environments (MANSE '99), pp 114–128, Dec 1999
4. Hong JI et al (2001) An infrastructure approach to context-aware computing. Human-Computer Interaction (HCI) Journal 16
5. Schilit B, Adams N, Want R (1994) Context-aware computing applications. IEEE workshop on mobile computing systems and applications, Santa Cruz, CA, US

Enhanced Security Scheme for Preventing Smart Phone Lost Through Remote Control

Jae Yong Lee, Ki Jung Yi, Ji Soo Park
and Jong Hyuk Park

Abstract The smartphone market is growing at a rapid pace. The security issues of smartphones are increasing, and preventive measures for smartphone loss are needed. The portability of smartphones has increased the risk of loss incidents, and the dangers of loss are increased by personal information leakage and subsequent damages. In this paper, we propose an enhanced security scheme through remote control for the prevention of smartphone loss through remote control to minimize such damages. The proposed scheme provides remote synchronization, blocking access to personal information, location tracking, remote camera control, and event log transferring. In addition, the scheme can prevent personal information leakages and increase the possibility of re-acquisition.

Keywords Android · Smart phone security · Preventing smart phone lost

J. Y. Lee · K. J. Yi · J. S. Park · J. H. Park (✉)
Seoul National University of Science and Technology (SeoulTech),
172 Gongreung 2-dong, Nowon-gu, Seoul 139-743, Korea
e-mail: jhpark1@seoultech.ac.kr

J. Y. Lee
e-mail: whenever@seoultech.ac.kr

K. J. Yi
e-mail: kaksy@seoultech.ac.kr

J. S. Park
e-mail: jisoo08@seoultech.ac.kr

1 Introduction

With the development of telecommunication technology, smartphone market shares have increased in the mobile market. Compared to the feature phone, the smartphone is advanced and equipped with a general-purpose OS. It combines the personal information manager of a PDA with the communication function of a feature phone. Recently released smartphones have touchscreen and QWERTY keyboards to provide convenience to the user, and can be equipped with GPSs, accelerometers, and high-resolution cameras [1].

Because of the extensive functionality of these smartphones, they store more of the user's personal information more than a feature phone. For Google's Android OS, it stores users' Google account at activation time, and information about contacts and schedules is stored. In addition, because it saves a lot of personal information in the SMS list and call list and a lot of data on the SD_card, the amount of leaked information due to attacks by malicious code will be grow faster than it will with the feature phone. These smart phones are always prone to security threats because the data network is always connected. Gartner predicts that security risks will rise, as smart phones get smarter, the smartphone user has careless habits in using their smart phone in spite of increasing security threats. The internet security company Trusteer investigated the 'phishing sites' that collect personal information illegally and found out that smart phone users are three times more likely than users of desktop computers to offer up confidential login details to a phishing sites, and that they are quicker to respond to phishing scams [2, 3].

In this paper, we inquire into various security threats to smart phones, and describe an enhanced security scheme through remote control to prevent personal information leakage and smartphone loss. The organization of this chapter is as follows: in Sect. 2, we will describe related works for smartphone security; in Sect. 3, we will propose an enhanced security scheme through remote control; finally, in Sect. 4, we will summarize and conclude this paper.

2 Related Works

In this section, we discuss important issues of personal information on smart phones, remote control, security threats due to the loss or theft of a smartphone, existing security solutions, and their issues.

Personal information on smartphones: "Law about information network use promotion and information privacy" defines personal information as "personal information is information that can identify a individual by name, social security number, such as code, letter, voice, video". Therefore most of the information that is stored in smartphones such as "phone numbers, contacts, text message list,

schedules, pictures, videos, call logs, internet history” can be defined as personal information [4].

Remote control: Remote control is often thought of as the functionality of managing servers remotely. Most IT managers use such remote control methods as Telnet, and FTP to remotely manage servers. In addition, the necessity of managing PCs on other networks has emerged. But there are problems posed by firewalls when connecting to PCs on other networks. Thus, an early remote control technique was to restrict access to authorized IPs. More recently, remote control has become possible via secure communication channels even through firewalls [5].

Security threat due to the loss or theft of a smart phone: Most of smartphone users store their personal information in their phones because of its functional ability and portability. Accordingly, the risk of personal information leakage and financial damages due to smartphone security incidents has been increasing. Predictable security threats are as follows [6, 7]:

- History leakage: Leakage of video/picture viewing logs, search logs, and password logs, that are stored in the smartphone.
- Leakage of stored information: Leakage of data such as pictures, videos, music, documents, the user’s mail account, stored cookies, and session information.
- Leakage and fabrication of call, and message information: Leakage of call log and message transfer log which can also lead to secondary damage by fabrication.
- Unauthorized charges: Causing financial damage by sending SMSs, attempting calls, or making unauthorized payment request.

Security solution trends: Most solutions for managing smartphone loss are based on existing USB security solutions. If devices registered on an administrative system the server can remotely control registered smartphones to prevent leakage of significant personal information. These security solutions for managing smartphone loss have been provided by several carriers. One of these, “Mobile Protection Service” provides a function that restricts the use of unauthorized USIMs, “Mobile Device Management Security Solution” provide ‘factory reset’, ‘camera block’, ‘print screen block’ functions. And “Information Keeper Service” can enable users to lock, adjust, manage, and back up or delete the information in a smartphone device through remote control. It also provides a function that can send a text message to a preset phone number when a different USIM is equipped. “Find My iPhone” for iPhone provides functions for ‘location tracking’, ‘sending message’, ‘change password’, and ‘delete all data’ [10].

Issues with the existing solutions: Most existing solutions for managing smart phone loss only restrict access through the touchscreen via ‘remote locking’. Therefore, it is still possible to access the phone through its USB port or even through serial cables, Bluetooth, or the infrared port. Data backup or removal services even back up users’ data to service providers’ private servers or back up data to smartphones’ SD_cards that have already been stolen. It decreases the

possibility of data recovery or has no use if the smartphone is not re-acquired. Also, the availability of GPS-based location tracking decreases indoors because of reception attenuation. With “Find My iPhone”, location tracking is not available in Korea because it is contrary to “Laws about Protection and Use of Location Information”.

3 Enhanced Security Scheme Through Remote Control

Even if users restrict touchscreen access by remote locking, it is still possible to access the smartphone through Bluetooth, USB, or 3G network. To make up for this, the enhanced security scheme through remote control blocks all routes that can access the user’s personal information in the smartphone. It facilitates location tracking while indoors by not just using the GPS, but also using the proximity to 3G cell tower, Wi-Fi access points(Aps) to establish a location. To improve the possibility of reacquiring a lost smartphone, it remotely operates the built-in camera, and sends the pictures to the user’s email address. If someone else uses the smartphone, it records an event log and sends it to the users’ email address.

In this chapter, we look at countermeasures to mitigate the loss of an Android-based smart phone (Fig. 1).

Remote synchronization: Android-based smartphones provide a synchronization function with a Google account. Because these synchronizations operate passively, when a smart phone is lost or stolen, there is no way to get a user’s information that was not already synchronized. To make up for this, there is a method to synchronize remotely on the web, so the user can get his information (Fig. 2).

Blocking access to personal information: If a smart phone detects an attempt to access a user’s personal information it will cut it off by blocking not only touchscreen access but also that of USB and Bluetooth (Fig. 3).

Location information transmission: Users can get their smartphone’s location information remotely on the web. The GPS-based location tracking has better accuracy, but while indoors, the accuracy is decreased. Therefore, the system enables location tracking by 3G cell-tower and Wi-Fi APs (Fig. 4).

Remote camera control: Users can control the built-in camera remotely on the web and pictures taken are sent to the user’s email address. It increases the possibility of reacquiring the smartphone (Fig. 5).

Event log transfer: Android-based smartphones can record logs about all the processes running on the system. When a smartphone is lost or stolen, these recorded logs are filtered and processed into an identifiable shape, and sent to the user’s email address. This way, the user can see the state of his smartphone (Fig. 6).

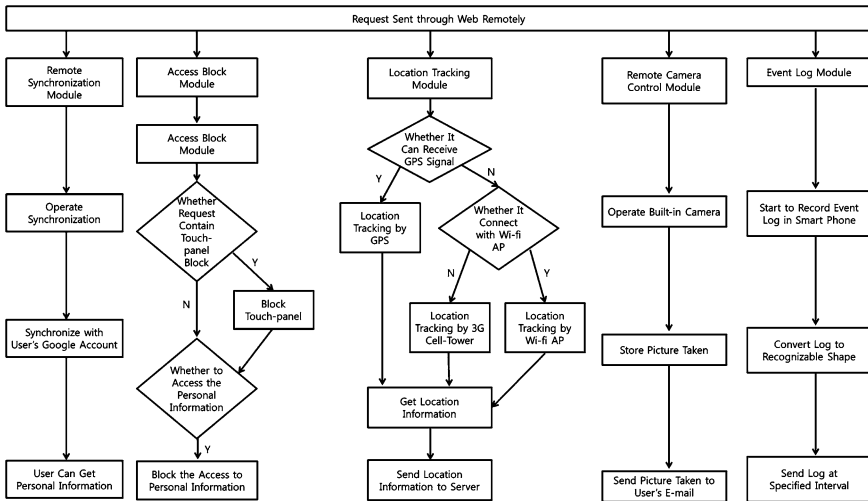


Fig. 1 Enhanced security scheme through remote control diagram

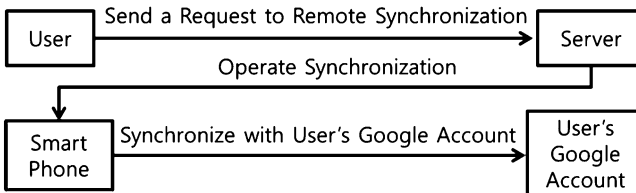


Fig. 2 Remote synchronization diagram

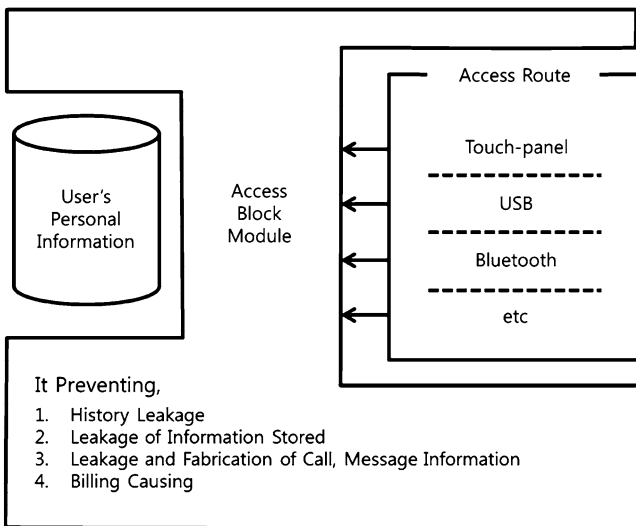


Fig. 3 Blocking access to personal information diagram

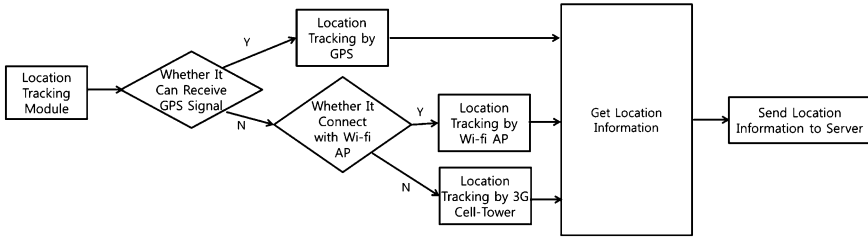


Fig. 4 Location information transmission diagram

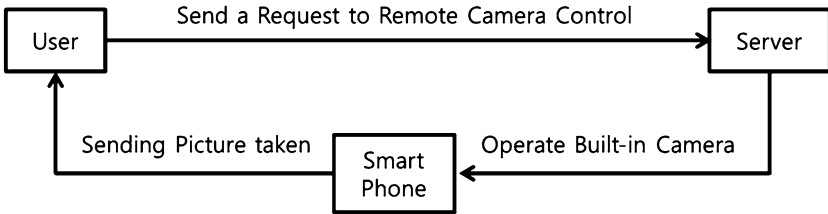


Fig. 5 Remote camera control diagram

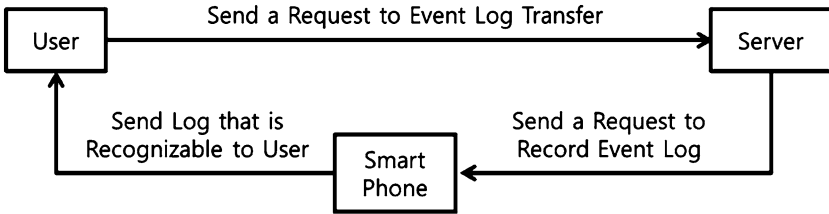


Fig. 6 Event log transfer diagram

4 Conclusion

In this paper, an enhanced security scheme through remote control is described. It focuses on preventing smartphone loss, blocking the leakage of the user’s personal information, and increases the possibility of reacquiring a lost smartphone.

These methods enable the user to actively cope with the security threats of a lost or stolen smartphone. They help to obtain personal information that was not previously synchronized and block unwanted access to the user’s information. Users can identify the event log of a lost smart phone, track its location information, or operate the built-in camera remotely. These can increase the possibility of reacquiring the smartphone and help minimize the financial damages.

For future research measures, the system could be made more lightweight to comply with the limited specifications of smartphones, and be adjusted to suit a variety of Android-based platforms. Management techniques and vulnerability analysis and the undergoing of certification processes by the user should be done in parallel, because these methods are liable to be abused by others.

Acknowledgments “This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)” (NIPA-2011-C1090-1131-0004)

References

1. Wikipedia, Smart Phone. <http://ko.wikipedia.org/wiki/%EC%8A%A4%EB%A7%88%ED%8A%B8%ED%8F%B0>
2. Gartner (2008) Security risks rise as smart phones get smarter, Computerworld
3. Trusteer (2011) Cell phone users are gullible, report says, PCWorld
4. Noh H, Kim J, Kim K (2008) Private information protection method supporting forensic in cell phone, engineering, information and communications univ
5. Support R (2005) Development of remote control products
6. Ju S (2010) Smart phone security threat and countermeasure in convergence environments, A3 security
7. Kang D, Han J, Lee Y, Cho Y, Han SW, Kim J, Kim H (2010) Smart phone threats and security technology, electronics and telecommunications trends 25-3
8. Trusted Computing Group (2007) TCG specification architecture overview. Specification revision 1.4. <http://www.trustedcomputinggroup.org>
9. Korea Internet & Security Agency (2010) A study of smartphone application market-oriented information security countermeasures
10. Nam G, Park S, Kang H, Gil J (2010) Smart phone security technology and solution trends, national IT industry promotion agency

SSP-MCloud: A Study on Security Service Protocol for Smartphone Centric Mobile Cloud Computing

Ji Soo Park, Ki Jung Yi and Jong Hyuk Park

Abstract Users can overcome the capacity and performance limitations of smartphone by using mobile cloud computing services. However, security threats accompany the use of the cloud computing with smartphone, both in preexisting cloud computing security issues and smartphone security issues, thus highlighting the need for research on security technologies. This paper will discuss the security threats and response technologies associated with cloud computing, smartphone and the security technology required for mobile cloud computing service using smartphone.

Keywords Smartphone · Cloud computing · Security

1 Introduction

Various services are provided in the new cloud computing paradigm, such as simple data storage services, OA(Office Automation) or development environment services. Users include daily users who log on to portal sites and employees who use internal cloud computing systems. The devices that are able to connect to cloud computing are changing from PCs to smartphone to Tablet PCs, which

J. S. Park · K. J. Yi · J. H. Park (✉)
Seoul National University of Science and Technology (SeoulTech),
172 Gongreung 2-dong, Nowon-gu, Seoul, 139-743, Korea
e-mail: jhpark1@seoultech.ac.kr

J. S. Park
e-mail: jisoo08@seoultech.ac.kr

K. J. Yi
e-mail: kaksy@seoultech.ac.kr

allows various terminals to use this service. Users can save data from their PCs, and store and access it using various mobile devices such as smartphone and Tablet PCs. Thus, users tend to prefer the IaaS (Infrastructure as a Service) for sharing and synchronizing various devices that can access cloud computing. Furthermore, by utilizing SaaS (Software as a Service), users can have word processing capability without purchasing OA. All cloud computing services are web-based. This means that cloud computing has the same vulnerabilities as the web, in addition to the vulnerabilities of the smartphone [1, 2].

This paper will discuss various security threats and countermeasures associated with cloud computing and smartphone as well as technologies for cloud computing security that are based on the mobile environment.

This article is organized as follows: “[Spam Host Detection Using Ant Colony Optimization](#)” will discuss related work on this topic; “[Location Estimation of Satellite Radio Interferer Using Cross Ambiguity Function Map for Protection of Satellite Resources](#)” will explore security service protocol for smartphone centric mobile cloud computing and “[Korean Voice Recognition System Development](#)” will present the conclusions of this study.

2 Related Works

2.1 Cloud Computing Security Technology

Cloud computing can be classified based on the scope of usage under the following categories: public cloud, private cloud and hybrid cloud. Public clouds offer services to a wide scope of people without any user limitations, whereas private cloud are only accessible to a limited number of users, for instance, within a company. A hybrid cloud is a mix between the public and private sphere, and is used to overcome the shortcomings of each one [1, 2].

Although there were pre-existing security threats in web-based cloud computing, new mobile terminals have created their own set of threats. Among these are smartphone data protection, ID management, cloud computing standards and problems involving service-to-mass.

In an attempt to solve ID management problems using SSO, the cloud service existed ‘outside’ the firewall, and the ID management was conducted through the router within the firewall. This allowed users to manage their IDs without going through authentication each time they used the service. Other major security issues include data integrity, availability and recovery, virtual machine protection, network security and attack modeling and simulation (M&S) [3–5].

Since cloud computing provides various resources necessary for computing, different security technologies are required for each type of resource. User authentication and access control technology is required for the platform. Service providers use authentication to confirm that the user is valid and access control to

restrict the user's access to services they are allowed to access. For storage services, personal information is protected by encryption or limiting the search of user data. To this end, searchable encryption systems and privacy preserving data mining (PPDM) technologies are implemented. For network services, security is enhanced by SSL/TLS using security protocol IPsec, certificate, application firewall and defense mechanisms that block increasing DDoS attacks [2].

2.2 Smartphone Security Technologies

Unlike contemporary PCs, smartphone can access a network via various routes and stays connected to the network at all times. Such characteristics present the greatest threats to smartphone users. Furthermore, there is the possibility of losing the phone, due to its mobility. The privacy issues associated with losing a smartphone are serious because the user's account information, work data and other critical information are stored in the device's internal storage. Moreover, the probability of malicious codes affecting the smartphone is much higher because of the phone's performance limitations. This requires a different defense technology to protect against malicious codes. Already, many users are experiencing problems caused by malicious codes, including device malfunctions, discharge, unfair accounting, information breaches and spreading to other devices via cross platforms [6].

The standardized method of verifying smartphone terminal reliability and integrity is currently under development. The same is true for server-based security management by remote control. In order to prevent the spread of malicious codes, applications are verified before registration. Application verification is different for each smartphone OS (Operating System). Google Android OS has relatively weaker criteria compared to other OS. Other technologies prevent the outflow of personal information by monitoring the control of sensitive data, which is a minimum policy suitable for the smartphone environment. In order to maximize resource efficiency, they also provide access control over user data, and automation of security services based on the purpose, location and network [6, 7].

3 Security Service Protocol for Smartphone Centric Mobile Cloud Computing

3.1 Smartphone Cloud Computing Security Threats

Smartphone cloud computing security threats involve the vulnerability of smartphone, wireless networks and cloud computing services. Major smartphone include, malicious codes that cause accounting problems, denial of service, information leakage, device malfunction and the loss and theft of devices.

Furthermore, if the smartphone loses access to the cloud computing service, then the stored data and resources within the cloud are exposed to threats. For wireless networks, information leakages by sniffing, network profiling, jamming and session hijacking are major potential threats. Finally, vulnerabilities exist within the cloud computing system, such as information leakages due to mismanagement, denial of services, access control through applications with defaults and other authentication threats [8].

3.2 Attack Scenario and Response Options

In order to respond to security threats in smartphone centric mobile cloud computing systems, a plan for each component (smartphone, network and cloud computing) is required.

- If data or applications with malicious codes are downloaded by a user, the user Cloud Computing account and data is extracted and unfair accounting occurs.
Response option: Only download verified data and block applications with abnormal activities.
- Cloud computing services are abused or account information is leaked if a smartphone using cloud computing is lost or stolen.
Response option: Utilize the internal lock system and enhance user authentication by establishing a secondary authentication process. Moreover, backup and initialize critical information using encryption and remote control.
- Information is leaked and a terminal malfunctions through various routes such as 3G, Wi-Fi, Bluetooth and USB.
Response option: Control access points based on usage frequency and cut any abnormal access.
- A user does not use encryption enabled wireless AP, and personal information within the smartphone, such as internet history and account information, is extracted.
Response option: Enable encryption and refrain from using a public wireless AP.
- Information is leaked from a broadcasted SSID and unauthorized personnel gain access.
Response option: Disable the broadcast of SSID and utilize an enhanced key authentication algorithm.
- Abnormal access occurs and web/network vulnerabilities are abused. Information is leaked and a denial of service attack occurs.
Response option: Implement network encryption such as SSL/TLS and IPsec to control access to mistrusted sites.
- Bypass authentication, authorization and access control, which are abusing the defects of an application.
Response option: Check for application defects when uploading the service, and if any defects are found, block the upload.

- A denial of service attack occurs and information is leaked, which abuses the internal defects of the cloud computing system.
Response option: Analyze vulnerabilities within the system and continue monitoring to identify abnormal patterns.

3.3 Proposed SSP-MCloud

In the proposed SSP-MCloud, protocol was comprised of two phases Smartphone Verification and Cloud Computing Verification. Smartphone sends to device status verification server to control the use of Cloud Computing services. Verification Server verifies a defect of Cloud Computing system and application.

3.3.1 System Parameters

The System parameters used in this scheme were as follows.

- *: (**SPD**: Smartphone Device, **SCVS**: Smartphone centric Mobile Cloud Computing Verification Server, **CCS**: Cloud Computing Service, **OD**: Other Device)
- **DDL**: Download Data Link
- **MCL**: Malicious Code List
- **LS**: Lost Status
- **NS**: Network Status
- **AID**: Application Information Data
- **SID**: System Information Data
- **V()**: Verify Function
- **S()**: Search Function
- **M()**: Status Modification Function

3.3.2 Phases for SSP-MCloud

- Phase 1. Smartphone Verification (Fig. 1).
Step 1. **SPD** → **SVCS**: SPD sends DDL to SVCS.
Step 2. **SCVS**: **S(MCL, DDL)**, SCVS searches MCL for same DDL.
Step 3. **SCVS** → **SPD**: SCVS responds accept or deny to SPD.
Step 4. **SPD** → **SVCS**: SPD sends NS to SVCS.
Step 5. **SCVS**: **V(NS)**, SCVS verifies network status.
Step 6. **SCVS** → **SPD**: SCVS responds accept or deny to SPD.
Step 7. **OD** → **SCVS**: OD requests to Smartphone lost at SCVS.
Step 8. **SCVS**: **M(LS)**, SCVS modifies lost status.

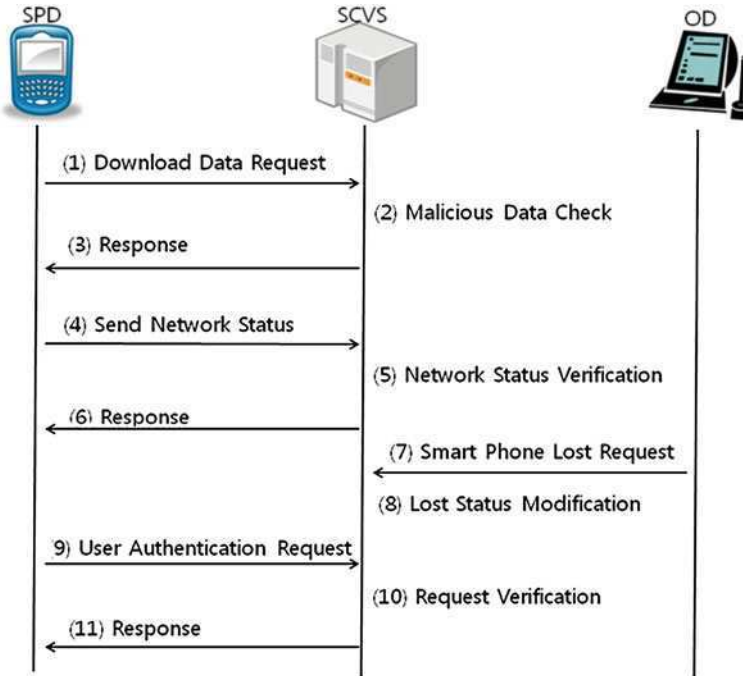


Fig. 1 Smartphone verification phase

- Step 9. **SPD** → **SCVS**: SPD request to user authentication at SCVS.
- Step 10. **SCVS**: $V(\text{Request}, \text{LS})$, SCVS verifies lost status.
- Step 11. **SCVS** → **SPD**: SCVS responds accept or deny to SPD.

- Phase 2. Cloud Computing Verification (Fig. 2).

- Step 1. **CCS** → **SCVS**: CCS sends AID to SCVS.
- Step 2. **SCVS**: $V(\text{AID})$, SCVS verifies Cloud Computing application.
- Step 3. **SCVS** → **CCS**: SCVS respond to CCS about application defect.
- Step 4. **SPD** → **CSS**: SPD request to Cloud Computing Service at CSS.
- Step 5. **CSS** → **SPD**: CSS respond to SPD about Cloud Computing Service.
- Step 6. **CSS** → **SCVS**: CSS sends SID to SCVS.
- Step 7. **SCVS**: $V(\text{SID})$, SCVS verifies Cloud Computing System.
- Step 8. **SCVS** → **CCS**: SCVS respond to CCS about System defect.
- Step 9. **SPD** → **CSS**: SPD request to Cloud Computing Service at CSS.
- Step 10. **CSS** → **SPD**: CSS respond to SPD about Cloud Computing Service.

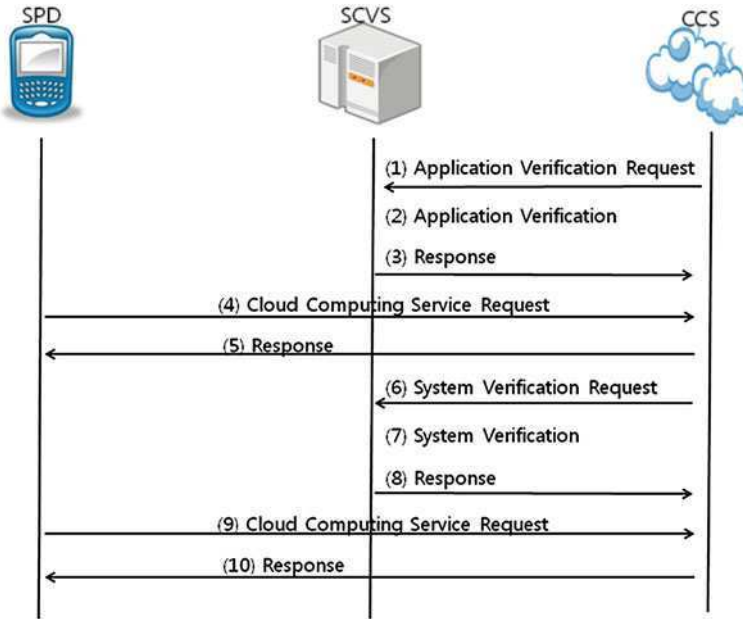


Fig. 2 Cloud computing verification phase

4 Conclusion

This paper discussed security threats and response options for cloud computing, smartphone, and smartphone cloud computing. Cloud computing is no longer based on the PC environment alone, but is relevant for mobile-based smartphone as well.

Smartphone cloud computing allows users unlimited resources through their smartphone and overcomes the limits of smartphone resources and capabilities. Despite the added benefits, threats to data and service leakages exist, as well as abuse caused by smartphone and cloud computing security vulnerabilities. This paper discusses various response options and technologies that counter smartphone cloud computing security threats.

In order to enhance the security of smartphone cloud computing, continuous research is needed to understand the various types of smartphone and cloud computing security threats. Moreover, continued research on new threats, and creating a response plan to such threats, is necessary to create a safer and more secure user environment for smartphone cloud computing.

Acknowledgments “This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency)” (NIPA-2011-C1090-1131-0004)

References

1. Min OG, Kim HY, Nam GH (2009) Trends in technology of cloud computing. *Teletronics Telecommun Trends* 24(4):1–13
2. Un SK, Jho NS, Kim YH, Choi DS (2009) Cloud computing security technology. *Teletronics Telecommun Trends* 24(4)
3. Five Cloud security issues to emerge in 2011 (2010) <http://www.idg.co.kr/newscenter/common/newCommonView.do?newsId=63667>
4. In reality, Cloud Security: Four Case (2010) <http://www.idg.co.kr/newscenter/common/newCommonView.do?newsId=62005>
5. Lim CS (2009) Cloud computing security technology. *J KIISC* 19(3)
6. Kang DH, Han JH, Lee YK, Cho YS, Han SW, Kim JN, Cho HS (2010) Smartphone threats and security technology. *Teletronics Telecommun Trends* 25(3)
7. Kim K, Kang DH (2009) Smartphone security technology in an open mobile environment. *J KIISC* 19(5)
8. Jang EY, Kim HJ, Park CS, Kim JY, Lee J (2011) The study on a threat countermeasure of mobile cloud services. *J KIISC* 21(1)

Self-Adaptive Strategy for Zero-Sum Game

Keonsoo Lee, Seungmin Rho and Minkoo Kim

Abstract Strategy is one of the most important factors to win a game. Especially in zero-sum game, where a loser is necessary to make a winner, the player who has better strategy can be the winner. A fixed or solid strategy cannot be the better strategy, because game is like dancing with partner and responding the partner's behavior is important. In order to win, the strategy should be dynamically adapted to the situation of the game according to the opponent's action and at the same time, the strategy should provide the suitable action with performance limitation such as time and space. In this paper, we propose a method of dynamically modifying the strategy to the drift of the game. This method classifies the game situation and selects the best action in that situation by evaluating all the possible options.

Keywords Evaluation function · Rule based strategy · Min–Max algorithm

K. Lee · M. Kim
Graduation School of Information and Communication,
Ajou University, Suwon, Korea
e-mail: lks7256@ajou.ac.kr

M. Kim
e-mail: minkoo@ajou.ac.kr

S. Rho (✉)
School of Electrical Engineering, Korea University, Seoul, Korea
e-mail: smrho@korea.ac.kr

1 Introduction

As the Sun Tzu said in his book “The Art of War”, it is the best not to fight when there is no chance of defeating the opponent [1]. In games, especially zero-sum games, every participant has a chance to win and want to achieve the winning. But what the zero-sum means is that all the participants cannot be the winner at the same time. If there is a winner, then there should be a loser. One of the most important elements for winning is a strategy. The basic strategy in zero-sum game is making actions which maximize my advantage and at the same time minimize the opponent’s advantage in every step until the game is over. If these actions are properly made, a victory can be guaranteed for any games. But finding such actions is not an easy task to achieve.

One reason for this difficulty is the limitation of time and space. In most games, a player is asked to finish the turn in a proper time. And to make a decision in that time, the player cannot consider all the possible options which can lead to the victory. For example, let me assume a game which has 40 turns to the end and in each turn, the player can choose one action from the eight possible options. In order to find the path to the victory, players need to travel the 8^{40} nodes which can be generated according to the players’ choices. As the game is getting more complex, the size of searching space is getting larger. Therefore, it will be impossible to consider the whole environment of the game to decide the current action with a given limited resources such as time and space for computation. The player should choose the plausible action for the victory with incomplete information. The way of finding the plausibility of each option will be the answer to this requirement.

The other reason is that the game is drifting. Until the end of the game, we cannot be sure who will be the winner. This is how the Rocky wins in the movie. There is always a chance of winning the losing game and losing the winning game. Even if a player does anything s/he wants to do, it cannot be guarantee the victory. It is just a necessary condition for winning. To disturb the opponent is required additionally. For this, a player, who wants to win, should keep eyes on what the opponent does and change the strategy to the opponent’s reaction.

Therefore, it is the most important factor for victory to rapidly select the proper action in a given situation. In order to achieve such task, it needs to take two subtasks. The first subtask is to recognizing the current situation. With this knowledge, we can decide whether the new strategy needs to be adapted or the existing strategy needs to be preserved. The second subtask is generating the suitable strategy for each situation. With these two subtasks, the player can decide the proper action with right strategy in a recognized situation. In this paper, we propose the method of finding such action. It consists of two parts, each for the subtask. One is classifying the current game situation. The other is modifying the strategy to the newly classified situation. In next article, the employed techniques in the proposed method will be described.

2 Proposed Method

2.1 Self-Adaptive Strategy

In using the minimax algorithm, the evaluation function works as the key role. As this evaluation function decides the best action is best in a given situation, we can say that the perfection of this function will decide the winner of the game. In order to make a good evaluation function, two features need to be considered. One is that a single evaluation function should not be used during the entire game. For example, chess can be divided into three phases; opening, middle game, and closing. In each phase, a different and proper strategy needs to be used. As a strategy for opening may be useless for middle game or closing, the strategy of the game should be modified according to the transition of the game. And to be a proper strategy, the strategy should be changed flexibly according to the opponent's movement. In the chess's opening phase, the best action for the opponent's king's Indian will be the king's Indian defense, even if the player prepared to make a Sicilian defense move. King's Indian and Sicilian opening are well known opening strategy in chess play. In this paper, we propose the way of modifying the strategy which is represented as evaluation function in minimax algorithm to provide the flexibility.

An evaluation function decides the best action in a given situation and this decision is affected by multiple features. If n features affect this selection, the form of this function will be like this. Each factor of this formula means how the feature affects the action's appropriateness. For example, if the features to be considered are king's safety and queen's activity, the action which has the biggest value of summing king's safety (f_1) and queen's activity (f_2) will be selected.

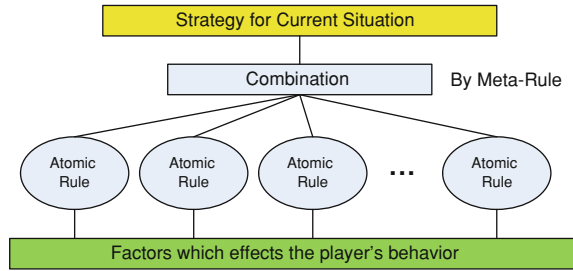
$$\text{Action's appropriateness} = \sum_{i=1}^n f_i \quad (1)$$

Each factor represents the rule of how to calculate the feature's affecting value. For example, the f_1 can be the rule which is that the king's safety is increased if the king is surrounded by same colored pieces. These rules are generated from the knowledge engineering of that game domain. If the rules are well made, the strategy can be better. And each rule has different priority. This means if several rules are fired at the same time, or conflicts among rules are made, more important rule should make more affection to the resulted action. With this importance, the above formula can be modified to this. Each rule is multiplied by its importance and the sum of all the factors makes the decision of next action. And these weight values are normalized to avoid the excessive leverage of a specific rule.

$$\text{Action's appropriateness} = \sum_{i=1}^n w_i * f_i \quad (2)$$

The weight of a rule defines the importance of that feature and this weight should be changed when the drift of a game is changed. For example pawns in

Fig. 1 Structure of strategy hierarchy from atomic action to composited action set



chess are not that important in the early phase. But if a pawn advances all the way to the opposite side, its importance will be the same as that of queen. Therefore, this set of weights should be modified flexibly to correspond to the current game state. And the modified calculation of each action's appropriateness becomes the strategy in the given game state. Figure 1 shows the basic structure of evaluation function from the atomic rules to the strategy.

2.2 Modification of Strategy

The modification has two components. One is to make a set of weight values for game states. And the other is defining the game states. With these two components, when the game situation is inserted, the situation is classified to the predefined state and the state's weight value set is applied to the current strategy as shown in Fig. 2.

In order to define the states of a game, we use time and topology as basic parameters for game state types. This step can be different according to the game domain. The time parameter has three values; opening, middle game, and closing. Opening means the phase from the start of the game to the completion of the formation. Generally, it takes 20% of the average turns for the game. Closing means the phase of making the drift of s game immobile. In this phase, the chance of victory is fixed and any move cannot turn the game around. Generally, when the minimax algorithm's foreseeing step reaches the end of the game, it can be regarded as closing. The between opening and closing will be the middle game. The topology parameter has various values [2]. So we recommend using x-means clustering method [3]. For the game's replay data, it divided into the time phase and each phase is clustered.

When the game's states are made from this process, each state's weight value set needs to be made. The features in this set should be made beforehand. Defining features to be considered in the strategy is the result from the knowledge engineering of that game domain. As this passes the bounds of this paper's topic, we assume that the proper atomic rules are generated. With this assumption, the weight values are made by the reinforcement learning [4, 5]. From the game's

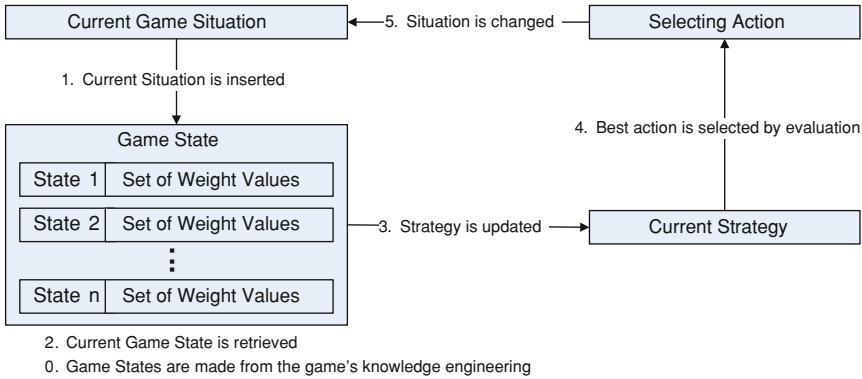


Fig. 2 Process of strategy modification

replay data, the best action in a given game state can be extracted. Simply the winner’s action in that situation will be the best action. In order to make more accurate training data set, the best action is defined according to the game state’s transition flow. If the action changes the game state to the favorable state for the player, it can be classified as the good action. If the action changes the game state to the unfavorable state for the player, it can be classified as the bad action. To use this method, each game state should be classified as favorable or unfavorable. This classification can be made based on the statistics. If the state is included more in the won game’s replay, it is classified as favorable. If it is included more in the lost game’s replay, it is classified as unfavorable.

If this pre-process is finished, this strategy can be used for new game. As shown in Fig. 2, firstly, the game’s situation data is inserted. Then, this data is classified to the defined game state. The matched state’s value set are applied to the current strategy and with this updated strategy, the possible actions are evaluated. The best action is selected and this action changes the game situation and this changed game situation is inserted when the player’s next turn starts.

3 Implementation

3.1 Othello as Game Domain

In order to test the proposed method, we select Othello game as the target game domain. Othello is a classic board game with two players [6, 7]. There are 64 places in the board and each player places his/her piece in the board. The piece can be placed only if the piece can flip the opponent’s pieces. One player’s piece between the opponent’s pieces is flipped to the opponent’s piece. When no empty

Table 1 The atomic rules for Othello game

ID	Description
Rule 1	The position which can flip the opponent's piece is good
Rule 2	The position which can reduce the opponent's possible positions is good
Rule 3	The position which is far from the border is better
Rule 4	The position between the opponent's pieces is good

place is left or both players have no place where their piece can be placed, the game ends and the player, who has more pieces in the board, wins. We make this Othello application in C# and the various strategies can be applied by changing the position selecting class. In order to make the test data set, 10 users played 300 games. Four players are novices, four players are in average level, and two players are very good at this game.

3.2 Atomic Rules and Meta Rules

First of all, the atomic rules are made by the help of the test users. The atomic rules are shown in Table 1. The states of game are defined by the topology of all the pieces in the board. The topology is defined by the rates of current possible positions over all the empty positions, current my pieces over opponent's pieces, and the possibly flipped opponent's pieces over my pieces. Time factor is not considered to make this game's states because the left empty positions, which is used, for topology can indicate the consumed time for the game. Each state's weight values are calculated from the 300 game replays. The inserted game situation is classified to the pre-defined game states by calculating the distance between each cluster's center position and the current situation. The state which has the shortest distance is selected. And the state's weight values are applied to the current strategy which consists of the rules shown in Table 1.

The result of this implementation and execution shows that the trained strategy wins the novices but hardly wins the two experts. But any single rule in Table 1 cannot win against the novices. And if the weight values of the strategy are fixed, the victory rate is lower than the tested result. The obvious conclusion of this simulation is that the good strategy is made from the good domain knowledge engineering. If we find the better atomic rules than those in Table 1, the result seems to be better. However, the significant point is that even the poor rules can win the game if they are correctly used in a given situation. As the test shows, the poor rule which cannot win can make victory by combined with other rules and the rate of victory can be increased by tuning their weights. With more appropriate rules, this proposed method of self-adaptive strategy can be more useful.

4 Conclusion

Strategy is the most important key to win a game. And this strategy needs to be flexible to handle the dynamically changing drifts of the game. For chess, the game is divided into three steps, opening, middle game, and closing. A single fixed strategy cannot be used these three steps. In this paper, we propose a method of modifying the strategy according to the situation of game to provide the flexibility. This method classifies the situation into the pre-defined types and employs the type's weight value set to modify the strategy. In order to maximize the advantage of this method, the atomic rules, which present the effect of each attribute to the game, and the pre-defined types, which tell when the strategy needs to be changed and how the strategy should be changed, should be provided correctly. This information is different according to the game domain. In this paper, Othello is selected as the game domain and implementation shows winning ratio of a little better than 50%. But considering that the used atomic rules, which hardly win if each single rule is employed, this ratio made by this proposed method can be good enough. With proper domain knowledge engineering, this method can increase the winning ratio by providing the dynamically changing strategy according to the game situation.

References

1. Sun T (2005) Translated by Lionel Giles [Translation first published 1910]: the art of war by Sun Tzu—Special Edition. El Paso Norte Press. ISBN 0-9760726-9-6
2. Xiang L, Guo Y, Lan T (2007) Topological cluster: a generalized view for density-based spatial clustering, international conference on management science and engineering pp 422–428
3. Pelleg D, Moore AW (2000) Extending K-means with efficient estimation of the number of clusters, ICML 2000 proceedings of the seventeenth international conference on machine learning. ISBN:1-55860-707-2
4. Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: a survey. *J Artif Intell Res* 4:237–285
5. Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. MIT Press, Cambridge, ISBN 0-262-19398-1
6. Othello game. <http://en.wikipedia.org/wiki/Reversi>
7. US Othello Association. <http://www.usothello.org/joomla/>
8. Russell SJ, Norvig P (2003) Artificial intelligence: a modern approach, 2nd edn. Prentice Hall, Upper Saddle River, pp 163–171. ISBN 0-13-790395-2
9. Victor A (1994) Searching for solutions in games and artificial intelligence. PhD Thesis, University of Limburg, Maastricht, The Netherlands. ISBN 9090074880
10. Heineman GT, Gary P, Stanley S (2008) Chapter 7: path finding in AI. Algorithms in a Nutshell. O'Reilly Media, Sebastopol, pp 217–223. ISBN 978-0-596-51624-6
11. Reilly DL, Cooper LN, Elbaum C (1982) A neural model for category learning. *Biol Cybern* 45:35–41. doi:[10.1007/BF00387211](https://doi.org/10.1007/BF00387211)

Effect of Light Therapy of Blue LEDs Irradiation on Sprague Dawley Rat

Taegon Kim, Yongpil Park, Yangsun Lee and Minwoo Cheon

Abstract In order to examine the healing effect of the blue light emitting diode (LED) irradiation on skin wounds, a round slice of wound 1 cm in diameter was cut on the back of the laboratory animal. Animals treated with blue light emitting diode (LED) irradiation ($p < 0.05$) healed at a faster rate than non-irradiated controls. The blue light emitting diode (LED) irradiated groups also had more collagen, according to Masson's trichrome staining for collagen analysis. In conclusion, blue light emitting diode (LED) irradiation had a beneficial effect on wound healing and could probably replace low level laser therapy.

Keywords Light therapy · Light emitting diode · Photobiomodulation · Photostimulation · Epithelialisation · Excision · Wound healing

T. Kim

Department of Electrical and Electronic Engineering,
Dongshin University Graduate School, 252 Daeho-dong, Naju,
Jeonnam 520-714, Republic of Korea

Y. Park · M. Cheon (✉)

Department of Biomedical Science, Dongshin University,
252 Daeho-dong, Naju, Jeonnam 520-714, Republic of Korea
e-mail: mwcheon@dsu.ac.kr

Y. Lee

Department of Information Communication Engineering,
Chosun University, 375 Seosuk-dong, Dong-gu,
Gwang-ju 501-753, Republic of Korea

1 Introduction

Currently, lasers are one of the most popular light sources in use for medical treatment. They can be divided into high power lasers and low power lasers depending on radiation level; the former is applied to surgery without bleeding or to instantly burn cells by using higher energy level [1–3] and poses a lower risk of side effects or inflammatory indications. It is also used for skin peeling, hair removal and artificial depigmentation. The latter penetrates into cells, effectively stimulating cellular tissues and activating cellular function [4]. Lasers with specific wavelengths such as He–Ne and GaAlAs lasers induce proliferation of fibroblasts depending on the wound area or wavelength [5] and are effective in pain relief [6], anti-inflammation [7] and wound healing [8–13]. Vinck et al. [14], reported that both LED irradiation and low level laser treatment induced high cellular proliferation in specific cells and that LED irradiation had a higher rate of proliferation than low level lasers. Many studies on low power lasers are being done in cell culture or through animal tests and most report different findings, making it difficult to verify their true effects. There are shifts in trends of studies from laser and LED that are expensive and generate heat problem to LED that are economically effective and safe.

This study verified the effect of blue LEDs irradiation on wound healing by using LEDs irradiation system.

2 Materials and Methods

2.1 LEDs Irradiator

The LEDs irradiator was designed to apply LEDs beams of wavelengths to light therapy, thus activating cell proliferation and wound healing. The device consists of five parts; power supply, key switch, control panel, LED driver, and LEDs module.

TLC5941 IC, a component of LED driver, was used to control multiple LEDs simultaneously including monicolor, multicolor, LED display and display back-light. The element used for our device controls 96 LEDs simultaneously and runs at 3.0–5.5 V and can control static current up to 90 mA. It can precisely adjust brightness in 4096 steps by using 12 bit gray scale PWM and brightness variance of LED driver by using 64 step static current sink (dot correction), respectively. The TLC5941 is composed of 16 static current output channels from 0 to 15 and can independently adjust output current by adjusting brightness variance in LEDs connected to channels. Eight TLC5941 ICs in total were used to control four LEDs. Three output channels were used for one TLC 5941 and beam output from one module of 96 LEDs was controlled with two TLC 5941 s.

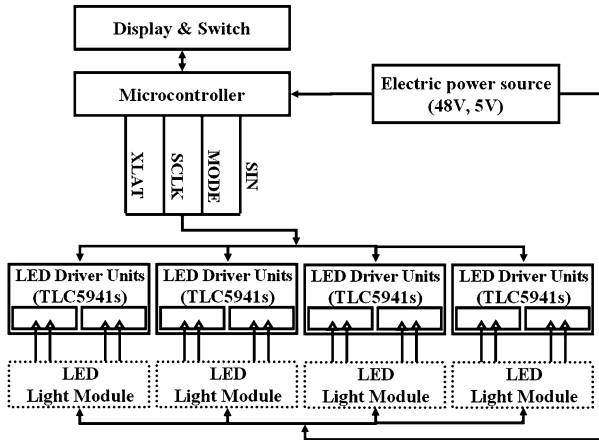


Fig. 1 Schematic diagram of LEDs irradiator. The micro controller used to control the LED irradiator consisted of a power supply, LCD display and switch. It was designed to control multiple LEDs that consisted of TLC 5941 IC through micro controller and serial interface and to make LED irradiation applicable by using light module that included 96 LEDs

One LED module consisted of 96 LED so that LEDs irradiation could be applied to cells and wounds and it was configured to supply current and voltage necessary for operation and control of LEDs. The LEDs module was connected to controller and D-sub connector via cable in order to make beam radiation available. For LEDs module, five Φ LEDs were used and it was designed so that LEDs irradiation could be evenly applied to well position of 96 well plate which was widely used for cell culture and bioassay. Fig. 1 shows the schematic diagram of LEDs irradiator.

2.2 Laboratory Animal Model

This study used eight week old male adult 250–300 g Sprague–Dawley Rat (ILAR CODE: NTacSam:SD) and minimized stressful factors while checking injury and disease and reducing environmental changes during test period. Breeding chamber was automatically adjusted at the temperature of 21–23°C and test chamber that was subject to LEDs irradiation was kept constant at 20°C, who were freely given animal feed and water. Both blue LEDs irradiation group and non-irradiation group were anesthetized with sevoflurane, and hairs on the back of test animals were removed. To obtain wounds even in size, a forceps type of tool was made, affixed to a 1 cm diameter circular blade and its case, and excisional wounds were taken out after removing the epidermal and dermal layers.

Fig. 2 Image of wound of laboratory animal. Its epidermal and dermal layers were removed by using our independently developed blade

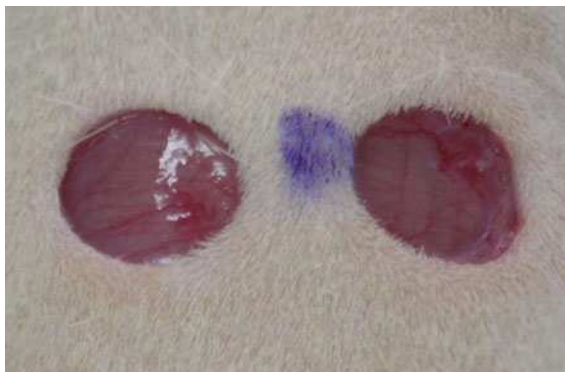


Table 1 The conditions of blue LEDs irradiated group

	Blue LEDs irradiation group
Light intensity	3.5 mW/cm ²
Irradiation time	1 h/day for 9 days
Wave type	Continuous wave

Wounds were made on both the right and left sides of the animals. The right side wound, 1 cm in diameter, which was pierced through the circular blade, was correctly taken out but the left side was deformed. So, the right side wound was used for the test. Fig. 2 shows the image of laboratory animal that had full the thickness excisional wound.

The LEDs irradiator was used to identify the effect of blue LEDs irradiation on wound healing. The test was carried out for non-irradiation group, blue LEDs irradiation group (n of each group = 7). Test animals were relieved for 24 h after wounds had been excised and then the blue LEDs irradiation group was given irradiation therapy over nine days one hour per day. Table 1 describes the experiment conditions for blue LEDs irradiated group.

2.3 Reagents

For immunohistological examination, mouse monoclonal antibody, pan-Cytokeratin (C-11), mouse monoclonal antibody Actin(C-2) and mouse monoclonal antibody PCNA (PC10) manufactured by Santa Cruz Biotechnology Co. Ltd (Santa Cruz, CA, USA) and secondary anti-body and color former, LSAB Kit and Diaminobenzoic acid (DAB) kit manufactured by Dako Co. Ltd.(Glostrup, Denmark), respectively were used. Contrast staining was conducted by using Mayer's hematoxylin and inclusion was treated with Universal Mount developed by Research Genetics (Huntsville, AL, USA).

Table 2 Defect size (mean \pm S.D.)

	Mean	S.D.	P value
Non-irradiation group (n = 7)	2,428 \pm	623	
Blue LED irradiation group (n = 7)	1,594 \pm	497	*

*P < 0.05

2.4 Tissues Samples

Tissues for observation were fixed with 10% neutral formalin and sliced in 4–5 μ m thickness by paraffin embedding and double stained with hematoxylin-eosin and treated with immunohistological staining against pan-cytokeratin and PCNA. Massons' trichrome staining was carried out to identify the regeneration of tissues in the defect area.

3 Results and Discussions

3.1 Defect Size Analysis

In order to examine the effect of blue LEDs irradiation on wound healing, the respective sizes of defects in the blue LEDs irradiation group and non-irradiation group were measured by using a MagnaFire digital camera system (Optronics, Goleta, CA, USA) and the long meter of defect was measured through Visus Image Analysis System (Image & Microscope Technology, Daejeon, Korea). Immunohistochemical staining was conducted for cytokeratin in order to precisely measure the defect size. Measurements were made for defect sizes after blue LEDs irradiation and mean \pm S.D was obtained and shown in Table 2.

Comparison was made between the blue LEDs irradiation group and the non-irradiation group and the defect size in the blue LEDs irradiation group decreased more than in the non-irradiation group (P < 0.05). It describes images of representative defect size in the non-irradiation group and the blue LEDs irradiation group. Figure 3 describes images of representative defect size in the non-irradiation group and the blue LEDs irradiation group.

3.2 Proliferating Cell Nuclear Antigen

To examine how fast tissues are regenerated, immunohistochemical staining was conducted for expression of proliferating cell nuclear antigen (PCNA), focusing on the defect area under repair. It was determined that pan-Cytokeratin was positive when cytoplasm was stained with brown color and PCNA when the nucleus was

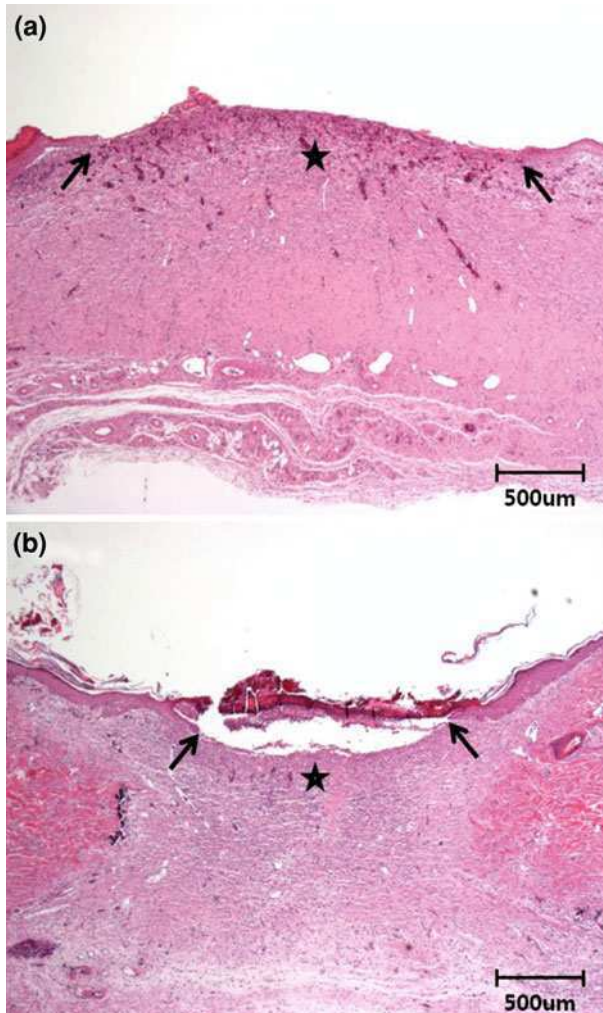


Fig. 3 Cases of defect areas by group. **a** shows 2,331µm tissue close to average 2,428 µm wound size for non-irradiation group, **b** 1,366 µm tissue close to average 1,594 µm wound size for the blue light irradiation group

Table 3 Proliferating cell nuclear antigen (mean \pm S.D.)

	Mean	S.D.	P value
Non-irradiation group (n = 7)	77 \pm	7.5	
Blue LEDs irradiation group (n = 7)	85 \pm	8.7	N.S.

*P < 0.05

N. S Non Significance

stained. The most strongly stained 4 or 5 points of PCNA were selected for photographing and observed by using computer measurement program and the percentage of cells that were stained positive was calculated for comparison with total cells. Table 3 shows the average percentage of PCNA by group.

As result of examining PCNA, there was no observed statistical significance in PCNA expression for both LEDs irradiation group and non-irradiation group. It is expected that this may be because each group went through the wound repair process from cell proliferation stage to collagen.

4 Conclusion

The study was carried out in vivo experiment by using the LEDs irradiation system to investigate the effects of the blue LEDs irradiation on the wound healing as a preliminary study aimed at the application of the blue light to the wound healing of human skin injury.

The defect size of the wound in the experimental animal was measured and it was found that blue LEDs irradiation group showed decreasing rate of defect size quicker than non-irradiation group. For the blue LEDs irradiation group defect sizes decreased in a statistically significantly manner ($p < 0.05$). Immunohistochemical staining was carried out for expression of PCNA and it was observed that cells were conspicuously proliferating around neighboring wound tissues. As there was some difference in wound repair among groups, no statistical significance could be observed for expression of PNCA between blue LEDs irradiation group and non-irradiation group.

References

1. Mosman T (1983) Rapid colorimetric assay for cellular growth and survival : application to proliferation and cytotoxicity assay. *J Immunol Method* 65:55–63
2. Polanyi TG, Bredemeier HC, Davis TW (1970) A CO₂ laser for surgical research. *Med Biol Eng* 8:541–548
3. Yahr WZ, Strully KT (1966) Blood vessel anastomosis by laser and other biomedical applications. *J Assoc Adv Med Instrum* 1:28–31
4. Beauvoit B, Kitai T, Chance B (1994) Correlation between the light scattering and the mitochondrial content of normal tissues and transplantable rodent tumors. *Biophys* 67:2501–2510
5. Bisht D, Gupta SC, Misra V, Mital VP, Sharma P (1994) Effect of low intensity laser radiation on healing of open skin wounds in rats. *Indian J Med Res* 100:43–46
6. Sakurai Y, Yamaguchi M, Abiko Y (2000) Inhibitory effect of low-level laser irradiation on LPS-stimulated prostaglandin E2 production and cyclooxygenase-2 in human gingival fibroblasts. *Eur J Oral Sci* 108:29–34
7. Whelan HT, Smits RLJ, Buchman EV, Whelan NT, Turner SG, Margolis DA, Cevenini V, Stinson H, Ignatius R, Martin T, Cwiklinski J, Philippi AF, Graf WR, Hodgson BGL, Kane

- M, Chen G, Caviness J (2001) Effect of NASA light-emitting diode irradiation on wound healing. *J Clin Laser Med Surg* 19:305–314
8. Alena RAP, Medrado L, Pugliese S, Regina S, Andrade ZA (2003) Influence of low laser therapy on wound healing its biological action upon myofibroblasts. *Laser Surg Med* 32:239–244
 9. In de Braekt MMH, Van Alphen FAM, Kujipers-Jagtman AM, Maltha JC (1991) Effect of low level laser therapy on wound healing after palatal surgery in Beagle dogs. *Lasers Surg Med* 11:462–470
 10. Jiro I, Kenji K, Kuo I, Masakazu K, Chie S, Shigeru K (2000) Progress in retinal and eye research 19:113
 11. Kana JS, Hutschenreiter G, Haina D, Waidelich W (1981) Effect of low-power density laser radiation on healing of open skin wounds in rats. *Arch Surg* 116:293–296
 12. Mester E, Spry T, Sender N, Tita J (1971) Effect of laser ray on wound healing. *Amer J Surg* 122:523–535
 13. Wong-Riley MT, Bai X, Buchamann E, Whelan HT (2001) Light-emitting diode treatment reverses the effect of TTX on cytochrome oxidase in neurons. *Neuroreport* 12:3033–3037
 14. Vinck EM, Cagnie BJ, Cornelissen MJ, Declercq HA, Cambier DC (2003) Increased fibroblast proliferation induced by light emitting diode and low power laser irradiation. *Laser Med Sci* 18:95–99
 15. Tada H, Shiho O, Kuroshima KI, Koyama M, Tsukamoto K (1986) An improved colorimetric assay for interleukin 2. *J Immunol Methods* 93:157–165

Fast Cancer Classification Based on Mass Spectrometry Analysis in Robust Stationary Wavelet Domain

Phuong Pham, Li Yu, Minh Nguyen and Nha Nguyen

Abstract Mass spectrometry (MS) is a technology recently used for high dimensionality detection of proteins in proteomics. However, due to the high resolution and noise of MS data (MALDI-TOF), almost existing MS analysis algorithms are not robust with noise and run slowly. Developing new ones is necessary to analyze such data. In this paper, we propose a novel feature extraction method considering the inherent noise of mass spectra. The proposed method combines stationary wavelet transformation (SWT) and bivariate shrinkage estimator for MS feature extraction and denoising. Then, statistical feature testing is applied to denoised wavelet coefficients to select significant features used for biomarker identification. To evaluate the effectiveness of proposed method, a double cross-validation support vector machine classifier, which has high generalizability, and a fast Modest AdaBoost classifier, which improves significantly experimental runtime, are applied for cancer classification based on selected features by proposed method. Several experiments are carried out to evaluate the performance of our proposed methods. The results show that our proposed method can be an effective tool for analyzing MS data.

Keywords Feature extraction · Mass spectrometry · SWT · Bivariate shrinkage · SVM · Boosting

P. Pham (✉) · L. Yu

Department of Electronics and Information Engineering, Huazhong University of Science and Technology, Wuhan, China
e-mail: phuongxuanpham@yahoo.com

M. Nguyen

Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA

N. Nguyen

School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

1 Introduction

The development of technology in recent years has stimulated many researches using proteomic mass spectrometry to recognize biomarkers for determining many types of cancer, distinguishing between tumor and normal tissues or between malignant and benign ones. Moreover, proteomic patterns also help to monitor disease progression as well as response to treatment. Protein serum profiles using time-of-flight (TOF) mass spectrometry technologies, such as Surface enhanced laser desorption/ionization—TOF (SELDI-TOF) and Matrix-assisted laser desorption/ionization—TOF (MALDI-TOF) for cancer classification have been developed by several researchers [1–4]. De Noo et al. [5] in 2006 assessed the feasibility of mass spectrometry based protein profiling for the discrimination of colorectal cancer (CRC) patients from healthy individuals. They pre-processed the spectra and applied a linear discriminant analysis with double cross-validation to classify protein profiles. Then, Alexandrov et al. [6] presented an improved procedure of biomarker extraction on the basis of the ideas proposed in Schleif et al. [7]: (i) discrete wavelet transformation of the spectra, (ii) feature (wavelet coefficients) selection by statistical testing, and (iii) double cross-validation SVM classification.

In this paper, we propose to use stationary wavelet transform (SWT) instead of DWT in feature extraction step of Alexandrov et al. [6]. Since SWT is shift-invariant transform, it transforms input data and restores perfectly the translation invariance, which is a desired property not obtained by DWT. Furthermore, a bivariate shrinkage function is combined with SWT in order to threshold noisy wavelet coefficients. Then biomarker selection is deployed on selected wavelet coefficients after the statistical feature testing. For cancer classification, the double cross-validation SVM classifier used in [5, 6] has high generalizability. However, the experimental runtime reaches tens of hours for each experiment. Therefore, we propose another procedure using bootstrapping resampling and boosting classifier, SWTBF-MBoost, which gives much faster runtime. We implement several experiments and compare our results with the results of [5, 6, 8] to evaluate the effectiveness of proposed methods. We also perform experiments using double cross-validation SVM classifier, SWTBF-SVM, in order to prove the superior of our proposed feature extraction to methods in [5, 6] and compare with the SWTBF-MBoost in terms of performance as well as runtime.

2 Methodology

2.1 Robust Stationary Wavelet Domain

SWT is derived from discrete wavelet transform. The DWT decomposes an input signal using successive low and high pass filters into subbands, followed by decimators. The DWT is a concise representation since the number of coefficients

over all subbands is equal to the length of input signal. The SWT performs the same transformation as DWT except the decimation steps. Appropriate low and high pass filters are applied to the input data at each level to produce two coefficient sequences at the next level. Without decimation steps, two new sequences each have the same length as the original data sequence. The filters, instead, are upsampled at each level of decomposition by zero-padding. This leads to a redundant representation of the original data. For example, with an input signal decomposed at level L , the redundant ratio is $(L+1):1$. However, this property, also called translation-invariant, has considerable statistics and classification potentials. For more information, see [9].

Wavelet thresholding is an approach which is usually used to remove noise and recover the true signal. Denoising procedure often follows three steps, using shrinkage techniques [9]. First, noisy data are transformed in wavelet domain. Next, hard or soft thresholding is applied to suppress noise portion from noisy wavelet coefficients. Wavelet coefficients falling below a threshold are set to 0 in hard thresholding. In soft thresholding, those falling below a threshold are set to 0 and others are shrunk toward 0 by subtracting the threshold from those coefficients. Finally, resulting wavelet coefficients are converted to original domain. Thresholding methods such as hard, universal, and un-universal thresholding were considered [9, 10]. However, these methods are not exploited the dependency of wavelet coefficients. Therefore, for denoising inherent noise of mass spectrometry signal, bivariate shrinkage estimator [11, 12], which takes into account the statistical dependencies among wavelet coefficients, is used. Bivariate shrinkage function can be obtained as below.

In wavelet domain, the noisy wavelet coefficient y can be calculated as

$$y = w + n \quad (1)$$

where w is the original coefficient and n is coefficient of noise, which is assumed as independent and identically distributed (i.i.d) Gaussian noise. To estimate w from the noisy observation y , the maximum a posteriori (MAP) estimator can be used. Let w_2 represent the parent of w_1 (w_2 is the wavelet coefficient at the same position as w_1 but at the next coarse level), y_1 and y_2 are noisy coefficients according to w_1 and w_2 respectively, n_1 and n_2 are noise coefficients. Then, as proposed in [12], the MAP estimator of wavelet coefficient w_1 is derived to be

$$\hat{w}_1 = \frac{\left(\sqrt{y_1^2 + y_2^2} - \frac{\sqrt{3}\sigma_n}{\sigma}\right)_+}{\sqrt{y_1^2 + y_2^2}} \cdot y_1 \quad (2)$$

Here $(g)_+$ is defined as

$$(g)_+ = \begin{cases} 0 & \text{if } g < 0 \\ g & \text{otherwise} \end{cases}$$

Here, we can estimate the noise variance from noisy wavelet coefficients by a robust median estimator at the finest scale wavelet coefficients [13]

$$\sigma_n = \frac{\text{median}(|y_i|)}{0.6745} \quad (3)$$

In this paper, the wavelet coefficients are considered as features which are further used for statistical feature selection and classification. We propose an approach of coefficient extraction, denoted as SWTB (SWT with bivariate shrinkage estimation), including detail coefficients of each level together with approximation coefficients of the maximum level. The SWTB coefficients are obtained as follows.

Suppose h'_j and g'_j are reconstruction filters at decomposition level j . Here, h'_j and h_j , g'_j and g_j have to satisfy the perfect reconstruction criterion by using quadrature mirror filters. The SWTB detail coefficients at level j , d'_j , can be calculated from SWT detail coefficients d_j, d_{j+1} as Eq. 2

$$d'_j = \mathbf{bishrink}(d_j, d_{j+1}) \quad (4)$$

and the SWTB approximation coefficients at level j , a'_j , are

$$a'_j = h'_{j+1} * a'_{j+1} + g'_{j+1} * \mathbf{bishrink}(d_{j+1}, d_{j+2}) \quad (5)$$

Suppose the maximum decomposition level is J and apply recursive rule in the Eq. 5, we can have

$$\begin{aligned} a'_j &= h'_{j+1} * h'_{j+2} * \dots * h'_{J-1} * (h'_J * a'_J + g'_J * d'_J) \\ &\quad + h'_{j+1} * h'_{j+2} * \dots * g'_{J-1} * (d_{J-1}, d_J) \\ &\quad + \dots + h'_{j+1} * g'_{j+2} * (d_{j+2}, d_{j+3}) \\ &\quad + g'_{j+1} * (d_{j+1}, d_{j+2}) \end{aligned} \quad (6)$$

Note that $a'_j = a_j$, $d'_j = d_j$ and $h'_j * a_j + g'_j * d_j = a_{j-1}$, so we can get

$$\begin{aligned} a'_j &= h'_{j+1} * h'_{j+2} * \dots * h'_{J-1} * a_{J-1} \\ &\quad + h'_{j+1} * h'_{j+2} * \dots * g'_{J-1} * \mathbf{bishrink}(d_{J-1}, d_J) \\ &\quad + \dots + h'_{j+1} * g'_{j+2} * \mathbf{bishrink}(d_{j+2}, d_{j+3}) \\ &\quad + g'_{j+1} * \mathbf{bishrink}(d_{j+1}, d_{j+2}) \end{aligned} \quad (7)$$

Here, if we define

$$\begin{aligned} p_{j+1,J} &= h'_{j+1} * h'_{j+2} * \dots * h'_{J-1} \\ p_{j+1,J-k} &= h'_{j+1} * h'_{j+2} * \dots * g'_{J-k} \\ p_{j+1,j+1} &= g'_{j+1} \end{aligned} \quad (8)$$

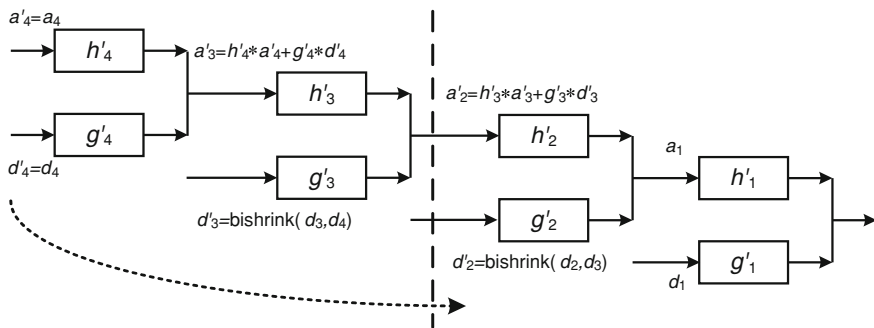


Fig. 1 SWT coefficients at level 2 with maximum decomposition level $J = 4$

then Eq. 7 can be written as

$$a'_j = p_{j+1,J} * a_{J-1} + \sum_{k=j+1}^{J-1} p_{j+1,k} * \mathbf{bishrink}(d_k, d_{k+1}) \tag{9}$$

With the combination as we propose above, the computation complexity reduces $O(\frac{N}{2} \log_2 N) + O(N \log_2 2^L)$ when comparing to general wavelet thresholding method, which decomposes noisy signal at maximum level, estimates true coefficients, reconstructs denoised signal, and then decomposes denoised signal at level L (Fig. 1).

2.2 Statistical Feature Selection

The feature extraction, in general, leads to a high-dimensional feature space. In case of SWT feature extraction, decomposition step creates a huge amount of coefficients that makes feature selection methods difficult to obtain a good selection for subsequent processing step. Therefore, we apply non-parametric statistical testing to extract the discriminative features between two classes of wavelet coefficients as in [4, 6, 7]. Two types of statistical testing are conducted, consisting of the two-sample Kolmogorov–Smirnov (KS) test and Mann–Whitney U (MW) test (or Wilcoxon rand sum test). Besides, multiple testing corrections which adjust p -values derived from multiple statistical tests to correct for the occurrence of false positives are carried out as used in [6], including Bonferroni (Bonf), Benjamini–Hochberg (BH), and Benjamini–Yekutieli (BY) adjustment. The statistical testing reduces considerably feature space dimensionality.

2.3 Biomarker Selection

The objective of biomarker selection is to find mass spectra that can be used for distinguishing cancer patterns from healthy patterns. In this paper, the biomarkers for each class are obtained from mass spectra which are represented by discriminative SWTB coefficients of each class. In details, after the statistical feature testing, only selected SWTB coefficients of each class are used to reconstruct the m/z ranges represented by these coefficients. Then, the average of the obtained spectral components of all spectra are computed for each class. The significant peaks are selected by the differences between the average reconstructed spectra of two classes.

2.4 Proposed Procedures for Mass Spectrometry Classification

To classify mass spectrometry data, we propose two procedures, one, denoted as SWTBF-SVM, is based on the above feature extraction method and ideas of [6], and the other, denoted as SWTBF-MBoost, uses boosting classifier for runtime improvement.

The classification procedures can be summarized as follows:

Step 1: *SWT with bivariate shrinkage estimator denoising*

- a. *Decompose mass spectrometry data using the SWT to maximum level.*
- b. *Calculate noise variance by using Eq. 3.*
- c. *Estimate denoised SWTB detail coefficients at each level (from 1 to L) by using Eq. 2.*
- d. *Estimate denoised SWTB approximation coefficients using Eq. 9.*

Step 2: *Perform statistical test on two groups of data and adjust p-values of the test to obtain discriminative coefficients.*

Step 3: *Classify obtained data by using double k-fold cross validation SVM or using bootstrapping resampling and Modest AdaBoost classifier*

There are some advantages of using SWTB coefficients for classification. First, the combination of SWT and bivariate shrinkage function is responsible for not only separating noise from signal but also estimating true components (wavelet coefficients) from noisy ones. This assists the statistical feature selection (Step 2) in finding significant differences of two classes. Second, shift invariant property of the SWT provides a great deal of spatial information, which makes it superior to the DWT in denoising and classification. The redundant representation property of the SWT reduces many of the artifacts created around the discontinuities of the input signal [9], which inherently exist in mass spectra, and provides more significant features for the better statistical feature selection and classification.

For classification in Step 3, we use a double k -fold cross-validation SVM to classify the discriminative coefficients obtained from previous steps. This SVM is a C-SVM type with a gaussian kernel and SVM hyperparameters C and γ are chosen by using the double CV paradigm as proposed in [14]. The double CV scheme includes an inner training loop using double k -fold CV to find the best C and γ based on grid search and an outer validation loop using leave-one-out cross-validation. For more complete coverage, see [6] and references therein.

Classification methods based on double k -fold cross-validation SVM, however, face a big problem relating to runtime since they take a long time to perform grid search steps. This makes them difficult to be applied in practice. Therefore, we propose to use the Modest AdaBoost classifier [15–17] which improves significantly experimental runtime but still gives good classification results. To assess the prediction performance of the classifier, bootstrapping resampling is used for repetitively selecting training and testing subsamples of data. Each subsample is a random sample, with replacement from the full dataset.

3 Experiments and Discussion

3.1 MALDI-TOF Serum Protein Profiles

The first dataset is the mass spectrometry based protein profile of colorectal cancer (CRC) patients and healthy individuals as mentioned in [6]. The preprocessed data consisted of 64 cancer and 48 control spectra of length 16331 m/z covering a domain of 960-11163 Da. The second dataset obtained from serum samples of HCC cases, cirrhosis cases, and healthy controls was collected from Egypt in the period from 2000 to 2002 as concerned in [8] and analyzed using an Ultraflex MALDI TOF/TOF mass analyzer (Bruker Daltonics). The preprocessed data consisted of 201 spectra (78 HCC, 51 cirrhosis and 72 normal) of the length 23846 m/z over the mass range 0.9–10 kDa.

3.2 Biomarker

Average biomarkers of CRC and control spectra using SWTB feature extraction with MW, BH statistical testing are presented in Fig. 2a. The biomarkers obtained by our method are more detailed than those of APPDWT in [6]. In the results of [6], the large peak at 1467 Da seem to be combined with the small peak at 1451 Da but it is separated clearly in our results. Besides, peaks at 1520 Da and 1538 Da of control biomarkers are difficult to recognize as in [6]. To compare visually the discrimination of average cancer and control biomarkers, the difference between class-discriminating parts of spectra (cancer minus control) are

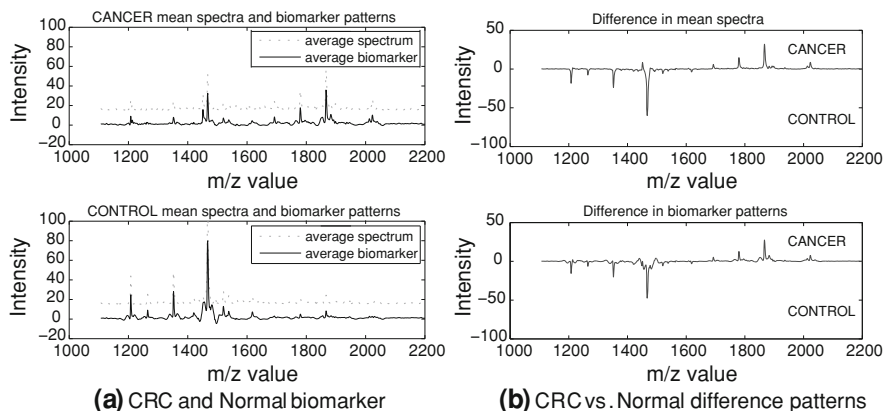


Fig. 2 Biomarker mean spectra and the differences of biomarkers. **a** CRC and Normal biomarker, **b** CRC versus Normal difference patterns

computed as shown in Fig. 2b. At the same m/z range 1200–2050 Da, our method detects 11 peaks with the largest difference are marked 1208, 1265, 135, 1467, 1520 for control spectra and 1451, 1693, 1779, 1867, 2023 for cancer spectra. The method in [6] omitted two difference peaks 1520 for control spectra and 1451 for cancer spectra. The difference of biomarkers are almost the same as the difference mean spectra.

We also find the biomarker patterns of HCC and cirrhosis and difference between class-discriminating parts of spectra (HCC-cirrhosis) using SWTBF-SVM. The difference of biomarker patterns with (KS, BH) testing method, for example, still keeps 10 largest peaks marked as 958, 2379 Da for HCC cancer spectra and 1352, 1451, 1467, 1683, 1779, 1867, 2605, 2605 and 2935 Da for cirrhosis cancer spectra as the difference of mean spectra.

3.3 Classification Results

To evaluate the performance of the proposed procedures, an average of classification results corresponding to six methods of statistical feature selection in [6] is calculated. The SWT in Step 1 is carried out at level decomposition $L = 5$ using bior6.8 wavelet, which has the largest filter length among bi-orthogonal wavelets. The statistical feature selection in Step 2 is performed at significant level $\alpha = 5\%$. The SWTBF-SVM applies the double 5-fold cross-validation SVM classifier in Step 3. The SWTBF-MBoost uses bootstrapping resampling and Modest AdaBoost classifier. The parameter ρ is set at 0.8 to select the same number of training samples and testing samples as 5-fold cross-validation SVM for the purpose of comparison. Modest Adaboost classifier is then applied with the number of

Table 1 A comparison of average classification results for detecting colorectal cancer from healthy patterns and detecting HCC from cirrhosis patterns

Methods	CRC dataset				HCC dataset			
	TRR (%)	Sens. (%)	Spec. (%)	Runtime (min)	TRR (%)	Sens. (%)	Spec. (%)	Runtime (min)
SWTBF-SVM	96.9	97.9	95.5	326	95.6	97.0	93.5	205
SWTBF-MBoost	96.8	97.1	96.4	22	95.1	97.1	91.9	6
CONVDWT	95.5	96.6	94.1	143	90.7	92.7	87.6	113
APPDWT	96.6	97.7	95.1	171	90.4	93.0	86.6	150

Table 2 A comparison of average classification results for detecting HCC from cirrhosis cancer with additional Gaussian noise

Method	$\sigma_n = 100$			$\sigma_n = 200$		
	TRR(%)	Sens.(%)	Spec.(%)	TRR(%)	Sens.(%)	Spec.(%)
SWTBF-SVM	95.1	96.6	92.8	94.1	96.2	90.9
SWTBF-MBoost	94.9	97.2	91.7	95.1	96.9	92.3
CONVDWT	89.9	92.5	85.9	89.7	91.7	86.6
APPDWT	90.1	92.3	86.6	90.1	91.9	87.3

boosting iterations set at 100 and the number of tree splits to the constructor chosen at 2. Our proposed methods are compared with CONVDWT, APPDWT methods in [6].

The average classification results for CRC and HCC dataset are shown in Table 1. The table shows that the SWTBF-SVM and SWTBF-MBoost have comparable results. For CRC data set, the SWTBF-SVM outperforms the CONVDWT and APPDWT 1.4 and 0.3% for average TRR, respectively. The SWTBF-MBoost also yields better results than CONVDWT and APPDWT, improving 1.3 and 0.2% for average TRR, respectively. For HCC dataset, the SWTBF-SVM exceeds the CONVDWT and APPDWT 4.9 and 5.2% for average TRR, 4.3 and 4.0% for average sensitivity, 5.9 and 6.9% for average specificity. The SWTBF-MBoost also gives better performance than the CONVDWT and APPDWT, improving 4.4 and 4.7% for average TRR, 4.4 and 4.1% for average sensitivity, 4.3 and 5.3% for average specificity.

In order to evaluate thoroughly the proposed methods' performances, several experiments are employed with additional Gaussian noise for HCC dataset. In Table 2, Gaussian noise is added at different levels of $\sigma_n = 100, 200$. At $\sigma_n = 100$, the SWTBF-SVM outperforms the CONVDWT and APPDWT 5.2 and 5.0% for average TRR, 4.1 and 4.3% for average sensitivity, 6.9 and 6.2% for average specificity. At $\sigma_n = 200$, the SWTBF-SVM is superior to the CONVDWT and APPDWT 4.4 and 3.9% for average TRR, 4.5 and 3.9% for average sensitivity, 4.3 and 4.0% for average specificity.

Modest AdaBoost classifier with bootstrapping resampling improves significantly computational efficiency in comparison with double cross-validation SVM classifier. In Table 1, the runtimes (run on PC with Pentium Dual Core 1.86GHz, 4GB DDR2) of the SWTBF-MBoost are faster than those of double cross-validation SVM approximate 6–15 times for CRC dataset and approximate 19–34 times for HCC dataset. The SWTBF-MBoost has better performance than double cross-validation SVM based methods in terms of computational efficiency since the former method is not involved in solving any quadratic programming problems as well as grid searching ones. The experimental results show that the classification performance of SWTBF-MBoost is comparable with that of the SWTBF-SVM.

4 Conclusion

In this paper, we develop a robust-to-noise feature extraction method which combined the SWT and a bivariate shrinkage estimator to transform MS data into robust stationary wavelet domain. Biomarker selection corresponds to new feature extraction method is also presented. To evaluate the performance of proposed methods, robust cancer classification algorithms, SWTBF-SVM and SWTBF-MBoost, are deployed on real MALDI-TOF datasets. The SWTBF-SVM has generalizability while the SWTBF-MBoost gives significantly faster experimental runtime. The experiments show that two proposed classification procedures provide better performance than other methods. The SWTBF-MBoost's performance is comparable with the SWTBF-SVM's one. Therefore, the SWTBF-MBoost is potential for MS analyzing applications which strictly require fast experimental runtime.

References

1. Morris JS, Coombes KR, Koomen J, Baggerly KA, Kobayashi R (2005) Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum. *Bioinformatics* 21(9):1764–1775
2. Petricoin EF 3rd, Ornstein DK, Paweletz CP, Ardekani A, Hackett PS, Hitt BA, Velasco A, Trucco C, Wiegand L, Wood K, Simone CB, Levine PJ, Linehan WM, Emmert-Buck MR, Steinberg SM, Kohn EC, Liotta LA (2002) Serum proteomic patterns for detection of prostate cancer. *J Natl Cancer Inst* 94(20):1576–1578
3. Ransohoff DF (2004) Rules of evidence for cancer molecular-marker discovery and validation. *Nat Rev Cancer* 4(4):309–314
4. Yu JS, Ongarello S, Fiedler R, Chen XW, Toffolo G, Cobelli C, Trajanoski Z (2005) Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. *Bioinformatics* 21(10):2200–2209

5. de Noo ME, Mertens BJA, Ozalp A, Bladergroen MR, van der Werff MPJ, van de Velde CJH, Deelder AM, Tollenaar RAEM (2006) Detection of colorectal cancer using maldi-tof serum protein profiling. *Eur J Cancer* 42(8):1068–1076
6. Alexandrov T, Decker J, Mertens B, Deelder AM, Tollenaar RAEM, Maass P, Thiele H (2009) Biomarker discovery in maldi-tof serum protein profiles using discrete wavelet transformation. *Bioinformatics* 25(5):643–649
7. Schleif FM, Lindemann M, Diaz M, Maa P, Decker J, Ellsner T, Kuhn M, Thiele H (2009) Support vector classification of proteomic profile spectra based on feature extraction with the bi-orthogonal discrete wavelet transform. *Comput Visual Sci* 12:189–199
8. Resson HW, Varghese RS, Drake SK, Hortin GL, Abdel-Hamid M, Loffredo CA, Goldman R (2007) Peak selection from maldi-tof mass spectra using ant colony optimization. *Bioinformatics* 23(5):619–626
9. Coifman RR, Donoho DL (1995) Translation-invariant de-noising. Technical report, Department of Statistics
10. Donoho DL (1995) De-noising by soft-thresholding. *Info Theory IEEE Trans* 41(3):613–627
11. Sendur L, Selesnick IW (2002) Bivariate shrinkage with local variance estimation. *IEEE Signal Process Lett* 9(12):438–441
12. Sendur L, Selesnick IW (2002) Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. *IEEE Trans Signal Process* 50(11):2744–2756
13. Donoho DL, Johnstone JM (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81(3):425–455
14. Mertens BJA, De Noo ME, Tollenaar RAEM, Deelder AM (2006) Mass spectrometry proteomic diagnosis: enacting the double cross-validators paradigm. *J Comput Biol* 13(9):1591–1605
15. Schapire RE (1999) A brief introduction to boosting. In: *Ijcai-99: Proceedings of the sixteenth international joint conference on artificial intelligence, Vols 1 and 2*, pp 1401–1406, 1452
16. Friedman J, Hastie T, Tibshirani R (2000) Additive logistic regression: a statistical view of boosting. *Ann Stat* 28(2):337–374
17. Vezhnevets A, Vezhnevets V (2005) Modest adaboost-teaching adaboost to generalize better. *Graphicon-2005*. Novosibirsk Akademgorodok, Russia

Part II
Future Security Technologies

An Improved User Authentication Scheme for Wireless Communications

Woongryul Jeon, Jeeyeon Kim, Junghyun Nam, Youngsook Lee and Dongho Won

Abstract Recently, wireless communications using mobile device are growing rapidly. The most advantage of wireless communication is that user can transfer various information to anywhere at any time using mobile device. However, to ensure security of communications, authentication is being magnified as an important issue in wireless communication. Recently, in 2011, Cui and Qin pointed out that Wu et al.'s scheme is failed to provide user anonymity and proposed an improved scheme. However, in this paper, we discuss that Cui et al.'s scheme is still vulnerable to malicious FA and does not provide perfect forward

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 (2011-0004751)

W. Jeon · J. Kim · D. Won (✉)

School of Information and Communication Engineering, Sungkyunkwan University,
300 Cheoncheon-dong, Jangan-gu, Suwon-si, Gyeonggi-do 440-746, Korea
e-mail: dhwon@security.re.kr

W. Jeon

e-mail: wrjeon@security.re.kr

J. Kim

e-mail: jeeyeonkim@paran.com

J. Nam

Department of Computer Science, Konkuk University, 322 Danwol-dong,
Chungju-si, Chungcheongbuk-do 380-701, Korea
e-mail: jhnam@kku.ac.kr

Y. Lee

Department of Cyber Investigation Police, Howon University, 727 Weolha-li,
Impi-Myeon, Gunsan-si, Jeonrabuk-do 573-718, Korea
e-mail: ysooklee@howon.ac.kr

secrecy. Finally, we will propose an improved scheme to overcome these vulnerabilities.

Keywords Wireless communication · Anonymity · User authentication · Security

1 Introduction

Recently, wireless communications are growing rapidly, and furthermore with growth of smartphone in mobile market, wireless network using mobile device is also growing fast. However, wireless network is susceptible to various attacks, because of its openness [1]. Thus, sometimes, wireless network security is considered as more complex than that of wired network. To ensure security of communications via wireless network, authentication is being magnified as an important issue. The authentication is fundamental for establishing secure communication channels over public insecure networks [2].

After Zhu and Ma proposed an authentication scheme to provide anonymity service for wireless communications [3], many subsequent studies have found and correct the security weaknesses [3–6]. In 2006, Lee, Hwang and Liao pointed out that Zhu et al.'s scheme is vulnerable to forgery attack and proposed an improved scheme [4]. Two years later, in 2008, Wu et al. discussed that both authentication scheme does not provide anonymity as claimed and then proposed a modified scheme [5]. Recently, in 2011, Cui and Qin pointed out that Wu et al.'s scheme also fails to provide anonymity service and then, proposed a modified scheme [6].

However, in this article, we show that Cui et al.'s scheme is still susceptible to malicious FA and does not provide perfect forward secrecy. To deter these vulnerabilities, we propose an improved authentication scheme with anonymity for wireless communications.

2 Review of Cui et al.'s Scheme

In this section, we review Cui et al.'s scheme for wireless communications. Following shows all notations used throughout in this paper.

- HA: Home agent of a mobile user.
- FA: Foreign agent of the network.
- MU: Mobile user.
- Att: Malicious attacker.
- ID_A : Identity of an entity A.
- PW_A : Password of an entity A.
- T_A : Timestamp generated by an entity A.
- $Cert_A$: Certificate of an entity A.

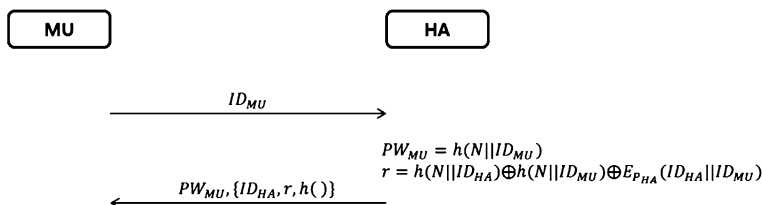


Fig. 1 Initial phase of Cui et al.'s scheme

- $(M)_K$: Encryption of a message M using a symmetric key K .
- $E_K(M)$: Encryption of a message M using an asymmetric key K .
- $h(\cdot)$: Cryptographic one-way hash function.
- \oplus : Bitwise exclusive-or(XOR) operation.

As previous researches, Cui et al.'s scheme consists of three phases: initial phase, first phase and second phase.

2.1 Initial Phase

Before starting communications in wireless network, mobile device has to register itself to HA for authentication. Following shows register process of mobile device which wants to use wireless network. The statement $A \rightarrow B: M$ denotes that A sends message M to B .

(1) $MU \rightarrow HA : ID_{MU}$

First, MU sends its own ID to HA for registration.

(2) $HA \rightarrow MU : PW_{MU}, \{ID_{HA}, r, h(\cdot)\}_{Smartcard}$

When HA receives a message from MU, HA computes PW_{MU} and r as follows;

$$PW_{MU} = h(N || ID_{MU})$$

$$r = h(N || ID_{HA}) \oplus h(N || ID_{MU}) \oplus E_{P_{HA}}(ID_{HA} || ID_{MU})$$

where N is a secret value kept by HA, and $E_{P_{HA}}$ denotes encryption using public key of HA. Finally, HA sends PW_{MU} and a smartcard which includes ID_{HA} , r and $h(\cdot)$ to MU. As a result, HA keeps a password table for user authentication (Fig. 1).

2.2 First Phase

In wireless network, MU can freely move to other FAs from HA. First phase shows that how FA authenticates MU and issues a temporary certificate.

- (1)
- $MU \rightarrow FA : n, c, ID_{HA}, T_{MU}$

When MU wants to access FA, MU computes n and C using r which is stored in his/her smartcard as follows:

$$n = r \oplus PW_{MU} = h(N||ID_{HA}) \oplus E_{P_{HA}}(ID_{HA}||ID_{MU})$$

$$C = (ID_{MU}||x_0||x)_L$$

where $L = h(T_{MU} \oplus PW_{MU})$ is a temporary symmetric encryption key between MU and HA, and both x_0 and x are random numbers generated by MU. Finally, MU sends n, C, ID_{HA} and T_{MU} to MU.

- (2)
- $FA \rightarrow HA : b, n, c, T_{MU}, E_{S_{FA}}(V), Cert_{FA}, T_{FA}$

Upon receiving the request message from MU, FA verifies T_{MU} . If T_{MU} is valid, then FA generates a random number, b and computes a digital signature $E_{S_{FA}}(V)$ with its private key, where $V = h(b, n, C, T_{MU}, Cert_{FA})$. Finally, FA sends $b, n, C, T_{MU}, E_{S_{FA}}(V), Cert_{FA}$, and T_{FA} to HA.

- (3)
- $HA \rightarrow FA : c, W, E_{S_{HA}}(b, c, W, Cert_{HA}), Cert_{HA}, T_{HA}$

First, HA verifies both $Cert_{FA}$ and T_{FA} . If they are valid, HA retrieves MU's identity from n , where $n = h(N||ID_{HA}) \oplus E_{P_{HA}}(ID_{HA}||ID_{MU})$. If MU is a legal user of HA, HA computes $L = h(T_{MU}||PW_{MU})$ and decrypts C to obtain x_0 and x . Using both x_0 and x , HA computes $W = E_{P_{FA}}(h(PW_{MU})||x_0||x)$, and generates a random number c . Finally, HA computes a digital signature $E_{S_{HA}}(b, c, W, Cert_{HA})$ and sends $c, W, E_{S_{HA}}(b, c, W, Cert_{HA}), Cert_{HA}$, and T_{HA} to FA.

- (4)
- $FA \rightarrow MU : (TCert_{MU}||h(x_0||x))_{sk}$

FA verifies HA's timestamp and certificate. If they are valid, then FA decrypts W and computes a session key $sk = h(h(PW_{MU})||x_0||x)$. Finally, FA issues a temporary certificate $TCert_{MU}$ and sends $(TCert_{MU}||h(x_0||x))_{sk}$ to MU.

- (5) MU computes
- sk
- and decrypts the message. Then MU verifies
- $h(x_0||x)$
- with his/her random numbers. If it is valid, then MU shares
- sk
- with FA (Fig. 2).

2.3 Second Phase

The second phase is invoked when MU wants to update its session key shared with FA. In order to enhance efficiency of the protocol, while MU stays with the same FA, the new session key sk_i can be derived from unexpired previous secret knowledge x_i and the fixed secret x as, $sk_i = h(h(PW_{MU})||x_i||x)$ where $i = 1, 2, 3, 4, \dots, n$.

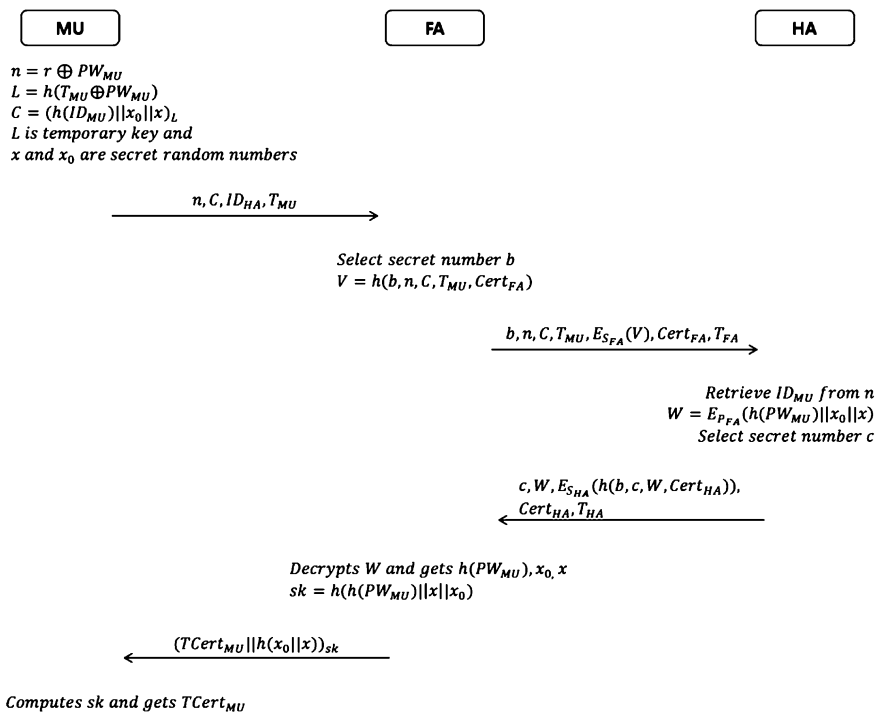


Fig. 2 First phase of Cui et al.'s scheme

3 Security Weakness of Cui et al.'s Scheme

This section demonstrates that Cui et al.'s authentication scheme is insecure against an active adversary.

In the wireless network, service boundary of FA is not strictly defined. Thus, MU can be located in two or more FA's service boundaries. This means that MU can communicate several FA's simultaneously. Now, we assume an adversary Att as a legal FA, which can communicate MU.

3.1 Session Key Exposure

Let's assume that MU starts the first phase with a FA in Att's service boundary. In this circumstance, MU is located in both FA and Att's service boundaries, and intended service provider is FA. Now, Att can obtain a secret session key, sk between MU and FA.

Following steps show session key exposure by Att.

(1) $MU \rightarrow FA : n, C, ID_{HA}, T_{MU}$

When MU sends the request message to FA, an adversary Att captures this message.

(2) $Att \rightarrow HA : b', n, c, T_{MU}, E_{S_{Att}}(V), Cert_{Att}, T_{Att}$

Att generates a random number, b' and computes a digital signature $E_{S_{Att}}(V)$ with its private key, where $V = h(b', n, C, T_{MU}, Cert_{Att})$. Finally, Att sends $b', n, C, T_{MU}, E_{S_{Att}}(V), Cert_{Att}$ and T_{Att} to HA.

(3) $HA \rightarrow FA : c, W, E_{S_{HA}}(b', c, W, Cert_{HA}), Cert_{HA}, T_{HA}$

First, HA verifies both $Cert_{Att}$ and T_{Att} . Because Att is also a legal FA, this verification is always valid. Then, HA retrieves MU's identity from n , where $n = h(N || ID_{HA}) \oplus E_{P_{HA}}(ID_{HA} || ID_{MU})$. If MU is a legal user of HA, HA computes $L = h(T_{MU} || PW_{MU})$ and decrypts C to obtain x_0 and x . Using both x_0 and x , HA computes $W = EP_{Att}(h(PW_{MU}) || x_0 || x)$, and generates a random number c . Finally, HA computes a digital signature $ESHA(b', c, W, cert_{HA})$ and sends $c, W, E_{S_{HA}}(b', c, W, Cert_{HA}), Cert_{HA}$, and T_{HA} to FA.

(4) Upon receiving the message from HA, Att decrypts W with its private key and compute $sk = h(h(PW_{MU}) || x_0 || x)$ which is the secret session key between MU and FA.

Actually, this attack is also available to [3–5], because MU does not notice ID_{FA} to HA neither. Therefore when HA receives messages from FAs simultaneously, HA cannot determine which FA is involved with MU in communication.

3.2 Perfect Forward Secrecy

The security of session key depends on the both random values, x_i and x . As mentioned in Sect. 2, to enhance efficiency of the protocol, MU just updates x_i to compute i-th session key.

When MU visits FA at i-th session, MU sends the message $(TCert_{MU} || x_i || other\ information)_{sk_{i-1}}$ where $sk_{i-1} = h(h(PW_{MU}) || x_i - 1 || x)$ is the current session key. Upon receiving the message, FA updates session key as $sk_i = h(h(PW_{MU}) || x_i - 1 || x)$.

However, Att can capture this message, too. When Att obtains this message from wireless communication, then Att decrypts this message using the exposed session key. Then Att obtains x_i and can compute the next session key. As a result, Cui et al.'s scheme does not provide perfect forward secrecy.

Table 1 Security Analysis

Schemes	Session key exposure	Perfect forward secrecy
Zhu et al. [3]	X	X
Lee et al. [4]	X	X
Wu et al. [5]	X	X
Cui et al. [6]	X	X
Our scheme	O	O

4 Improved Scheme

This section demonstrates an improved authentication scheme for wireless communications. From the above analysis, we can easily conclude that HA has to identify FA which MU intends to communicate with. Due to this, we modify C in authentication phase as follows:

$$C = (h(ID_{MU}) || ID_{FA} || x_0x)_L$$

Likewise, when HA decrypts C in the first phase, HA verifies FA’s identity with ID_{FA} . Att cannot replace ID_{FA} with ID_{Att} , because C is encrypted by L .

5 Security Analysis

This section demonstrates that how our modified scheme prevents attacks and provides perfect forward secrecy.

In order to impersonate FA, an adversary Att has to replace ID_{FA} with ID_{Att} in C . However, it is impossible because C is encrypted by L . To compute L , Att has to know PW_{MU} . Actually, Att can easily get T_{MU} in wireless communications. Nevertheless, Att cannot compute L , because Att cannot compute PW_{MU} , where $PW_{MU} = h(N || ID_{MU})$. To compute PW_{MU} , Att has to know both N and ID_{MU} , but it is impossible, because N is the strong secret key of HA. Therefore, our scheme is secure against to a malicious adversary.

Table 1 shows comparison between our scheme and previous schemes.

6 Conclusion

Recently, in 2011, Cui and Qin pointed out that Wu et al.’s scheme is failed to provide user anonymity and proposed an improved scheme.

However, in this paper, we discussed that Cui et al.’s scheme is still susceptible to malicious FA and does not provide perfect forward secrecy. To deter these vulnerabilities, we proposed a modified scheme and achieved perfect forward secrecy.

References

1. Nam J, Paik J, Kang H, Kim U, Won D (2009) An off-line dictionary attack on a simple three-party key exchange protocol. *IEEE Commun Lett* pp 205–207
2. Jeong H, Won D, Kim S (2010) Weakness and improvement of secure hash-based strong password authentication protocol. *J Inform Sci Eng* 26(5):1845–1858
3. Zhu J, Ma J (2009) A new authentication scheme with anonymity for wireless environments. *IEEE Trans Consum Electron* 50(1) 13(3): 230–234
4. Lee CC, Hwang MS, Liao IE (2006) Security enhancement on a new authentication scheme with anonymity for wireless environments. *IEEE Trans Ind Electron* 53(5):1683–1687
5. Wu CC, Lee WB, Tsaur WJ (2008) A secure authentication scheme with anonymity for wireless communications. *IEEE Commun Lett* 12(10):722–723
6. Cui X, Qin X (2011) An enhanced user authentication scheme for wireless communications. *IEICE Trans Info Syst* E94-D(1):155–157

An Improved Protection Profile for Multifunction Peripherals in Consideration of Network Separation

Changbin Lee, Kwangwoo Lee, Namje Park and Dongho Won

Abstract Multifunction peripherals, capable of networking and equipped with several hardcopy functions with various security functions, are taking place of printers and other printing devices in office workplaces. However, the security functions within a multifunction peripheral and its IT environments may have vulnerabilities. The information transmitted in multifunction peripherals includes very sensitive data since the device is networked to transmit data including confidential information. There have been international efforts to mitigate this anxiety of consumers through common criteria. In 2009, a series of standards for multifunction peripherals were developed. These protection profiles are classified in accordance to four different operational environments. However, though multifunction peripherals treat confidential information, network separation issue is not

This research was supported by the The Ministry of Knowledge Economy (MKE), Korea, under the “ITRC” support program supervised by the National IT Industry Promotion Agency (NIPA) (NIPA-2011-C1090-1001-0004).

C. Lee · K. Lee · D. Won (✉)
Information Security Group, Sungkyunkwan University, Suwon, Korea
e-mail: dhwon@security.re.kr

C. Lee
e-mail: cblee@security.re.kr

K. Lee
e-mail: kwlee@security.re.kr

N. Park
Department of Computer Education, Teachers College,
Jeju National University, Jeju-si, Korea
e-mail: namjepark@jejunu.ac.kr

regarded in classifying the operational environments. Thus, in this paper, we present an operational environment and propose a protection profile that is appropriate for the new environment.

Keywords Network separation · Common criteria · Security evaluation · Multifunction peripheral · Virtual personal network

1 Introduction

Multifunction peripherals (MFP for short hereinafter) are office machines which incorporate the functionality of multiple devices such as printer, copier, scanner, and fax machine in one. Recently, several security functions have been added to MFP to protect the confidential information and intellectual property related with various areas including extremely sensitive information such as industry technology information from leakage. To evaluate and assure that these security functions are well developed and functioning, Common criteria (CC for short hereinafter) evaluation assurance on MFP is in progress in Japan, USA, and South Korea. However, there was no criterion that is well-documented to evaluate security functions of MFP. The insufficient criterion made developers and evaluators difficult to evaluate the security functions of MFP. To solve this problem, IEEE P2600 Working Group has been developing protection profiles for MFP. The developed protection profiles are classified according to their usage in four different operational environments. Although these protection profiles tried to consider all of the possible operational environments for MFP, they still did not consider every possible operational environment of MFP. The missing environment is highly considerable. Therefore, it is required to add a protection profile to existing protection profiles to evaluate additional security functions of MFP. Thus, in this paper, we present an operational environment and propose a protection profile that is appropriate for the new environment [1, 2].

This paper is organized as follows. In [Sect. 2](#), we briefly describe existing standards for MFP and discuss the operational environments of each. Then, we present a new operational environment for MFP and propose supplementary classes along with the environment in [Sect. 3](#). Finally, we summarize and conclude our research in [Sect. 4](#).

2 Background and Related Work

In this section, we present a list of standards with brief background information, and analyze each of them with regard to their operational environments.

Table 1 IEEE P2600 family of standards

Standard	Year issued	Description
IEEE 2600	2008	IEEE 2600 is a core document of IEEE P2600 family of standards. This defines security requirements and identifies security exposures for MFP. This document instructs manufactures and developers on appropriate security capabilities and instructs users on appropriate usage of security capabilities
IEEE 2600.1	2009	IEEE 2600.1 is a protection profile for MFP in operational environment A
IEEE 2600.2	2009	IEEE 2600.2 is a protection profile for MFP in operational environment B
IEEE 2600.3	2010	IEEE 2600.3 is a protection profile for MFP in operational environment C
IEEE 2600.4	2010	IEEE 2600.4 is a protection profile for MFP in operational environment D

2.1 Existing Standards

IEEE P2600 Working Group is an approved standardization project that is sponsored by the IEEE Information Assurance Standards Committee of the IEEE Computer Society. The working group mainly endeavors to develop system security in MFP and its system. The working group has developed several standards as shown in Table 1 [3–5]. The operational environments A, B, C and D will each be explained in next subsection.

2.2 Existing Operational Environments

An operational environment denotes the total environment in which an MFP operates, including the consideration of the value of assets and controls for operational accountability, physical security, and personnel. In this subsection, we describe the classification of operational environments mentioned in the previous subsection [3].

Operational Environment A. Operational environment A processes restrictive information in which high security and assurance are required. In this environment, the facility is typically relatively large building or campus with a large population of networked devices. Many visitors will be present in this environment while there is some security. The hospital will be a good example of this environment since there are visitors everywhere and the processed information is partially very sensitive, for instance, patients’ personal and medical information.

Operational Environment B. Operational environment B processes restrictive information in which moderate security and assurance are required. In this environment, the facility is typically medium to large businesses, some governmental agencies, and organizations requiring managed telecommuting systems and remote offices. An average number of visitors will be present in this environment while

there is some security. A high-tech international company will be a good example of this environment since it treats important information such as product plans and company intellectual property with moderate security.

Operational Environment C. Operational environment C processes public-facing information in which security of documents is not guaranteed, but access control and accounting of usage are important. A public library will be a good example of this environment since it is open to publics and the processed information is usually not very significant.

Operational Environment D. Operational environment D processes private and personal information in presence of very high physical security. Small level of network security is needed for protection. A home user with stand-alone systems, possibly using wired or wireless home networks are good example of this environment.

2.3 Necessity of Network Separation

A device with wired or wireless network capabilities are exposed to attacks, malwares, and viruses since attackers may access the network as well as expected users [6]. To alleviate this problem, there have been efforts to separate networks into internal and external networks. There are physical and local ways to achieve network separation .

Physical Network Separation. Physical network separation is a method that literally separates internal network and external network physically. Figure 1 shows an example of physically separated network.

Device-based Separation. Device-based separation requires two devices; one device is for external network and another for internal network.

Switch-based Separation. In switch-based separation, hard disk, IP, and routing information are divided to separate network, then a switch in a peripheral component interconnect (PCI) card form is used to make the system operate as if there are two devices.

Logical Network Separation. Logical network separation usually utilizes virtualization technology to divide a network. The network is not divided physically, but operates as there are two different networks. Figure 2 shows an example of logically separated network.

Server-based Computing Separation. A terminal is used to install a virtual machine on a server. Then the application on the server processes the data.

Centralized-server Virtualization based Separation. A virtualized operating system operated by server is used through a program installed on client device that is run by a hypervisor on a centralized-server.

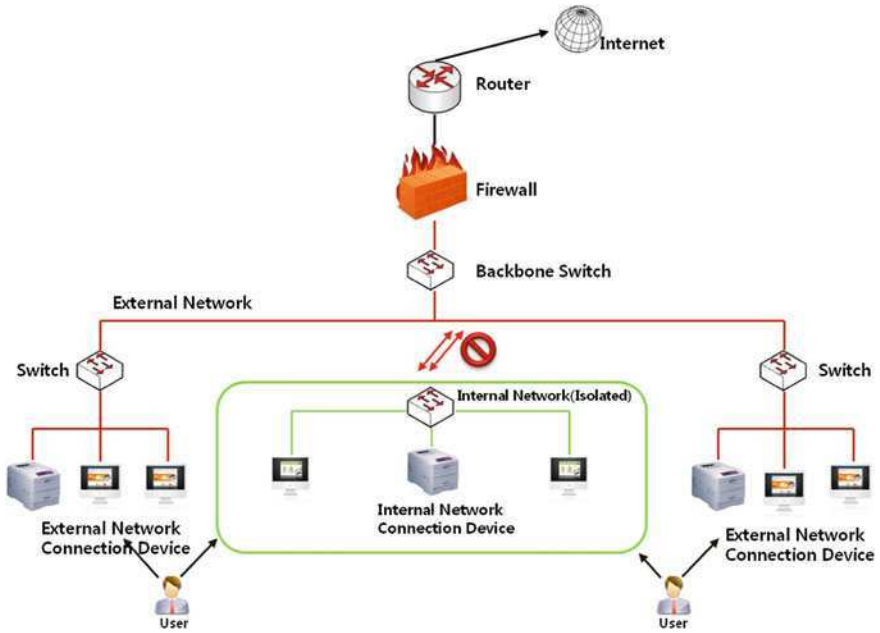


Fig. 1 A switch-based network separation

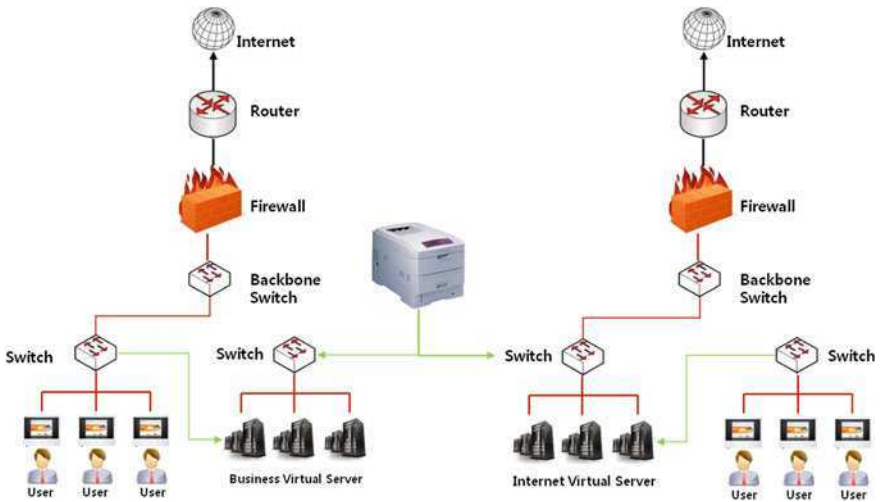


Fig. 2 A logically separated network

3 Improved Protection Profile

In this section, we will present a new operational environment that need be considered and propose a protection profile for that environment conforming CC version 3.1 revision 3 [7–9].

3.1 *New Operational Environment*

Recently, network separation in government agencies and military forces has been a significant issue. There have been reactions to this issue, including hardcopy devices. In this environment, network is separated into internal and external networks, and virtual personal network is used to communicate between two networks. However, as shown in Sect. 2.2, this environment is not taken into consideration in previous standards. Therefore, we propose a new operational environment where there is very sensitive information is processed and thus network is separated to protect the data. Recently, network separation in government agencies and military forces has been a significant issue. In this environment, network is separated into internal and external networks, and virtual personal network is used to communicate between two networks. As shown in Sect. 2.2, however, this environment is not taken into consideration in previous standards. Therefore, we propose a new operational environment where there is very sensitive information being processed and thus network is separated to protect the data.

3.2 *Protection Profile for New Environment*

In this subsection, we propose a protection profile for the environment stated in the previous subsection. Since this environment requires high assurance, we decided to inherit or keep the classes that were already specified in previous protection profiles, more specifically, IEEE 2600.1. We have selected several classes to evaluate and assure in this new environment. The selected classes are as follows in Table 2 [1, 7–9].

Information flow control functions: the components below are necessary to control information transmitted in separated network. The dependencies are reviewed, and identified to exist in the previous protection profile.

FDP_IFF.1.1 The TSF shall enforce the [MFP information flow control SFP] based on the following types of subject and information security attributes: [VPN related subjects and information]. MFP information flow control SFP should be set by manufacturers. The VPN related subjects and information covers functions regarding VPN.

Table 2 Additional security functional requirements

Class	Component	
Information flow control functions	FDP_IFF.1	Simple security attributes
	FDP_IFF.6	Illicit information flow monitoring
	FDP_IFC.1	Subset information flow control
Security management	FMT_MSA.3	Static attribute initialization

FDP_IFF.1.2 The TSF shall permit an information flow between a controlled subject and controlled information via a controlled operation if the following rules hold: [when the VPN related rules defined in FDP_IFF.1.3 are enforced successfully]. This component checks if the rules defined in FDP_IFF.1.3 are enforced upon permitting information flow.

FDP_IFF.1.3 The TSF shall enforce the [specified VPN related rules]. The VPN related rules should be defined by manufacturers; and this component ensures that the defined rules are enforced.

FDP_IFF.1.4 The TSF shall explicitly authorize an information flow based on the following rules: [no explicit authorization rules].

FDP_IFF.1.5 The TSF shall explicitly deny an information flow based on the following rules: [when the VPN related rules defined in FDP_IFF.1.3 are not enforced successfully]. This component ensures that the flow which does not abide by the rules is denied.

FDP_IFF.1.6 The TSF shall enforce the [MFP information flow control SFP] to monitor [VPN related illicit information flows] when it exceeds the [appropriate capacity]. This component protects MFP from illicit information flows through VPN.

FDP_IFC.1.1 The TSF shall enforce the [MFP information flow control SFP] on [VPN related subjects, information, and operations]. The VPN related subjects, information, and operations cover functions regarding VPN.

Security management: the components below are necessary to manage security functions need in separated network environments. The dependencies are reviewed, and identified to exist in the previous protection profile.

FMT_MSA.3.1 The TSF shall enforce the [MFP information flow control SFP] to provide [restrictive] default values for security attributes that are used to enforce the SFP.

FMT_MSA.3.2 The TSF shall allow the [none] to specify alternative initial values to override the default values when an object or information is created.

4 Conclusion

The development in information and communication technology and the prevalence of IT devices have made our lives more convenient and abundant. As an integrated hardcopy device with various functions, MFP is letting us to be more

efficient in many of office workplaces nowadays. However, evaluation and assurance of security functions of MFP needs to be achieved before we utilize this device. We have discussed insufficiency of previous criteria of MFP in regard to operational environment and proposed measures to make up to the deficiency. We hope that our research facilitates the evaluation process on behalf of MFP manufactures, purchasers, and evaluators.

References

1. Lee H, Won D, Kim S (2010) Protection profile for E-certificate issuance system. In: Proceedings of ICCC 2010, 11th international common criteria conference
2. Lee K, Lee Y, Won D, Kim S (2010) Protection profile for secure E-voting systems. In: Proceedings of ISPEC 2010, information security practice and experience conference 2010, Springer, LNCS 6047, Seoul, pp 386–397
3. Common Criteria (2009) Common Criteria for Information Technology Security Evaluation; Part 3: Security assurance components, Version 3.1 R3, CCMB-2009-07-003
4. IEEE: IEEE Standard for Information Technology (2008) Hardcopy device and system security, IEEE Std. 2600-2008
5. IEEE (2009) IEEE standard for a protection profile in operational environment A, IEEE Std. 2600.1-2009
6. Lee K, Lee C, Park N, Kim S, Won D (2011) An analysis of multi-function peripheral with a digital forensics perspective. In: Proceedings of CNSI 2011, international conference on computers, networks, systems, and industrial engineering, Jeju Island, May 23–25, 2011, pp 252–257
7. Smith TF, Waterman MS (1981) Identification of common molecular subsequences. *J Mol Biol* 147:195–197
8. Common Criteria (2009) Common criteria for information technology security evaluation; Part 1: Introduction and general model, Version 3.1 R3, CCMB-2009-07-001
9. Common Criteria (2009) Common criteria for information technology security evaluation; Part 2: Security functional components, Version 3.1 R3, CCMB-2009-07-002
10. IEEE (2009) U.S. government protection profile for hardcopy devices, IEEE 2600.2-2009
11. Park N, Song Y, Won D, Kim H (2008) Multilateral approaches to the mobile RFID security problem using web service. In: Zhang Y, Yu G, Bertino E, Xu G (eds) APWeb 2008. LNCS, vol 4976. Springer, Heidelberg, pp 331–341
12. 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
13. Park N, Kim H, Kim S, Won D (2005) Open location-based service using secure middleware infrastructure in web services. In: Gervasi O, Gavrilova ML, Kumar V, Laganá A, Lee HP, Mun Y, Taniar D, Tan CJK (eds) ICCSA 2005. LNCS, vol 3481. Springer, Heidelberg, pp 1146–1155
14. Park N, Kim S, Won D (2007) Privacy preserving enhanced service mechanism in mobile RFID network. In: ASC, advances in soft computing, vol 43. Springer, Heidelberg, pp 151–156
15. Park N (2010) Security scheme for managing a large quantity of individual information in RFID environment. In: CCIS, communications in computer and information science, vol 106. Springer, Heidelberg, pp 72–79
16. Park N, Kim S, Won D, Kim H (2006) Security analysis and implementation leveraging globally networked mobile RFIDs. In: PWC 2006. LNCS, vol 4217. Springer, Heidelberg, pp 494–505

17. Park N, Kim Y (2010) Harmful adult multimedia contents filtering method in mobile RFID service environment. In: LNAI, lecture notes in artificial intelligence, vol 6422. Springer, Heidelberg, pp 193–202
18. Park N, Song Y (2010) Secure RFID application data management using all-or-nothing transform encryption. In: WASA 2010. LNCS, vol 6221. Springer, Heidelberg, pp 245–252

Security Improvement to an Authentication Scheme for Session Initiation Protocol

Youngsook Lee, Jeeyeon Kim, Junghyun Nam
and Dongho Won

Abstract Recently, Yoon et al. proposed authentication scheme suited for session initiation environments. Our analysis shows that Yoon et al.'s scheme does not achieve its fundamental goal of password security. We demonstrate this by mounting an undetectable on-line password guessing attack on Yoon et al.'s scheme. We then figure out how to eliminate the security vulnerabilities of Yoon et al.'s scheme and improved over their scheme.

Keywords Authentication scheme · Session ignition · Password · Undetectable on-line password guessing attack · Session key

This work was supported by Howon University in 2011.

Y. Lee

Department of Cyber Investigation Police, Howon University,
727 Weolha-li, Impi-Myeon, Jeonrabuk-do, Gunsan-si, 573-718, Korea
e-mail: ysooklee@howon.ac.kr

J. Kim · D. Won (✉)

School of Information and Communication Engineering,
Sungkyunkwan University, 300 Cheoncheon-dong,
Jangan-gu, Gyeonggi-do, Suwon-si, 440-746, Korea
e-mail: dhwon@security.re.kr

J. Kim

e-mail: jeeyeonkim@paran.com

J. Nam

Department of Computer Science, Konkuk University, 322 Danwol-dong,
Chungcheongbuk-do, Chungju-si, 380-701, Korea
e-mail: jhnam@kku.ac.kr

1 Introduction

The Session Initiation Protocol (SIP) is an International Engineering Task Force (IETF)-defined signaling protocol, widely used for controlling multimedia communication sessions such as voice and video calls over Internet Protocol (IP) [1–7]. The protocol can be used for creating, modifying and terminating two-party (unicast) or multiparty (multicast) sessions consisting of one or several media streams. The modification can involve changing addresses or ports, inviting more participants, and adding or deleting media streams. Other feasible application examples include video conferencing, streaming multimedia distribution, instant messaging, presence information, file transfer and online games.

Recently in [8], Yoon et al. presented an efficient user authentication scheme suited for session initiation protocol. In their article, they claim that the proposed scheme authentication is fundamental for establishing secure communication channels over public insecure networks [4].

After Zhu and Ma proposed an authentication scheme to provide anonymity service for wireless communications [5], many subsequent studies have found and correct the security weaknesses [1, 5–7]. In 2006, Lee, Hwang and Liao pointed out that Zhu et al.'s scheme is vulnerable to forgery attack and proposed an improved scheme [6]. Two years later, in 2008, Wu, Lee and Tsaur discussed that both authentication scheme does not provide anonymity as claimed and then proposed a modified scheme [7]. Recently, in 2011, Cui and Qin pointed out that Wu et al.'s scheme also fails to provide anonymity service and then, proposed a modified scheme [1].

However, in this article, we uncover that Yoon et al.'s scheme does not guarantee its main security goal of password security. We show this by mounting an undetectable on-line password guessing attack on Yoon et al.'s scheme. What we do in this work is to report these security vulnerabilities of Yoon et al.'s scheme and to show how to eliminate them. The remainder of this paper is organized as follows. [Section 2](#) reviews Yoon et al.'s user authentication scheme. Then, [Sect. 3](#) presents our attack on Yoon et al.'s. and continuously [Sect. 4](#) descriptions the improved two schemes which offer a security patch for it. In [Sect. 5](#), we provide a security analysis of the proposed two schemes. Finally, we conclude this work in [Sect. 6](#).

2 Review of Yoon et al.'s Authentication Scheme

This section reviews an authentication scheme proposed by Yoon et al. [8]. The scheme participants include a user and an authentication server. For simplicity, we denote the user and the servers by U and S . Yoon et al.'s scheme consists of two phases: registration phase and authentication phase. The registration phase is performed only once per user when a new user registers itself with the

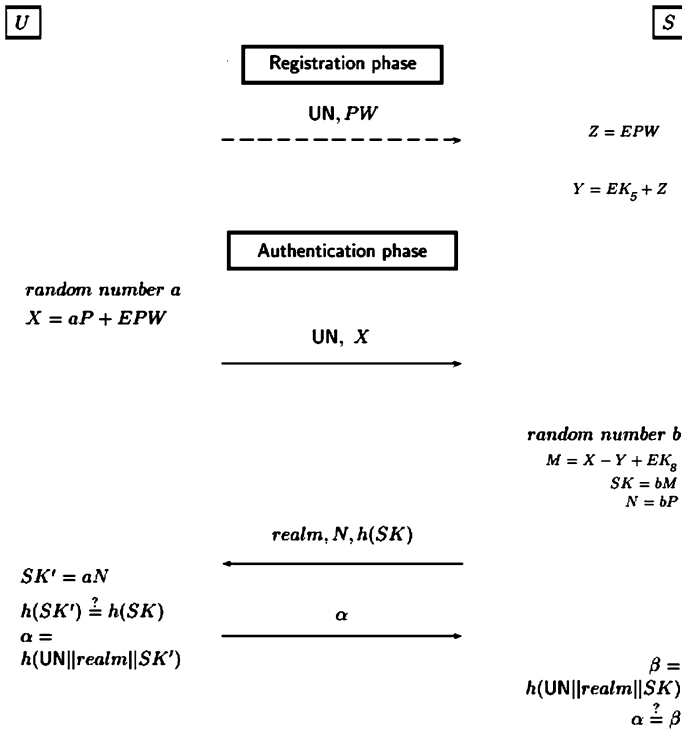


Fig. 1 Yoon’s registration and authentication phase

authentication server. The authentication phases are carried out whenever a user wants to gain access to the authentication server. Before the registration phase is performed for the first time, the user U and the authentication server S decide on the following system parameters: a one-way hash function h , an elliptic curve E , a finite field $GF(q)$, a generating element P of $E(GF(q))$. U and S choose an elliptic curve E over a finite field $GF(q)$ be an additive group of points on an elliptic curve E over a finite field $GF(q)$. Let P be the generating element of $E(GF(q))$. A high level depiction of the scheme is given in Fig. 1, where dashed lines indicate a secure channel, and a more detailed description follows:

2.1 Registration Phase

This is the phase where a new registration of a user takes place. The registration proceeds as follows:

Step 1. A user U , who wants to register with an authentication server S , chooses its password PW at will and submits a registration request, consisting of its username UN and PW , to the authentication server S via a secure channel.

Step 2. Upon receiving the request $\langle \text{UN}, PW \rangle$, S computes a secret value $Z = EPW$ which is an elliptic curve point in $E(GF_{(q)})$ from the password PW and $Y = EKS + Z$ by using its secret key EKS .

Step 3. The authentication server S stores the user's user name UN and Y in the user account database.

2.2 Authentication Phase

When U wants to log into the system, he enters his username UN and password PW into the client system.

Step 1. The user U generates a random number a and computes $X = aP + EPW$. Then U sends a login request message $\langle \text{UN}, X \rangle$ to the server S .

Step 2. When the login request arrives $\langle \text{UN}, X \rangle$, the server S first chooses the random number b and computes $M = X - Y + EKs$, a secret session key $SK = bM$, and $N = bP$. Then S sends the response message $\langle \text{realm}, N, h(SK) \rangle$ to the user U .

Step 3. Having received $\text{realm}, N, h(SK)$ from the server S , the user U computes $h(SK') = h(aN)$. U verifies the correctness of $h(SK)$ by checking that $h(SK')$ equals $h(SK)$. If correct, U accepts as the authentic server, computes $\alpha = h(\text{UN} \parallel \text{realm} \parallel SK')$ and sends $\langle \alpha \rangle$; otherwise, aborts its login attempt.

Step 4. After receiving α , S computes $\beta = h(\text{UN} \parallel \text{realm} \parallel SK)$. The server S checks that α equals β . If they are equal, S believes that he is talking to legal user. If they are not equal, S believes that he is talking to illegal user and aborts the scheme.

3 Cryptanalysis of Yoon et al.'s Scheme

Unfortunately, Yoon et al.'s scheme described above is completely insecure in the presence of an active adversary. To show this, we present an undetectable on-line password guessing attack that exploits password security weaknesses in the scheme [9].

3.1 Undetectable On-Line Password Guessing Attack

An attacker also may try to verify a guessed password in an on-line transaction; he verifies his guess using responses of a server. If his guess fails, he starts a new transaction with the server using another guessed password. However, in successful attack, a failed guess cannot be detected and logged by the server, as the server is not able to distinguish an honest request from a malicious one.

In Yoon et al.'s scheme, now the following description represents our undetectable on-line password guessing attack mounted by the attacker U_a against U 's password: The attacker U_a , who wants to find out PW , now guesses possible passwords and checks them for correctness.

1. The attacker U_a , who wants to find out the user U 's passwords, chooses the random number a' and computes $X' = a'P + EPW'$ using guessed password PW' . Then, U_a posing as U_i , sends $\langle U, X' \rangle$ to the server S .
2. Since, from S 's point view, UN, X' are indistinguishable from UN, X of an honest execution, S believes that the message $\langle UN, X' \rangle$ is from U . Hence, S operates as specified in protocol using the received messages from U_a . The authentication server S computes N , and SK and sends the message $\langle realm, N, h(SK) \rangle$ to the attacker U_a posing as U .
3. Now, an attacker U_a upon receiving $h(SK)$ and N from S , computes $h(SK'_{a'}) = a' N$. U_a then verifies the correctness of PW' by checking the equality $h(SK'_{a'}) = h(SK)$. Notice that if PW' and PW are equal, then $h(SK'_{a'}) = h(SK)$ ought to be satisfied.
4. U_a repeats a new transaction with the server using another guessed password until a correct password is found.

4 The Proposed Two Schemes

In this section we propose two authentication schemes which enhance on previous scheme, Yoon et al.'s scheme [8]. Like Yoon et al. scheme, the improved authentication scheme consists of two phases: the registration phase, the authentication phase. In describing the scheme, we will omit the registration phase because the registration phase of the proposed two schemes is equal to Yoon et al.'s registration phase. Before the registration phase is performed for the first time, the user U and the authentication server S decide on the following system parameters: a one-way hash function h , an elliptic curve E , a finite field $GF_{(q)}$, a generating element P of $E(GF_{(q)})$. U and S choose an elliptic curve E over a finite field $GF_{(q)}$ be an additive group of points on an elliptic curve E over a finite field $GF_{(q)}$. Let P be the generating element of $E(GF_{(q)})$. A high level depiction of the scheme is given in Figs. 2 and 3, where dashed lines indicate a secure channel, and a more detailed description follows:

4.1 The Proposed Scheme 1

Authentication Phase. When the user U wants to log into the system, he enters his username UN and password PW into the client system.

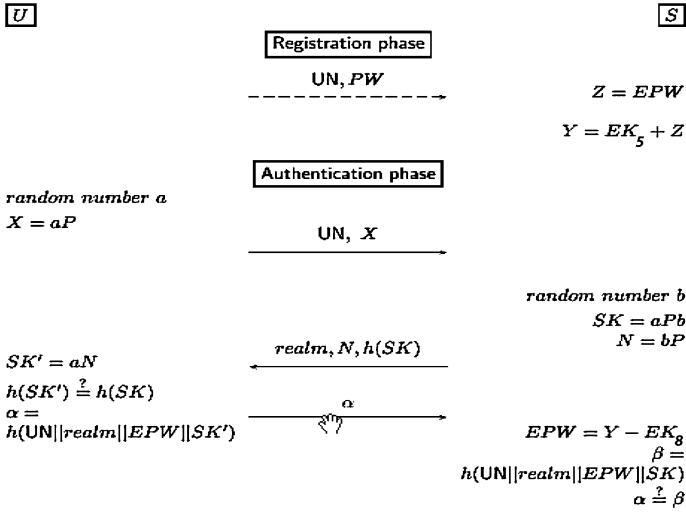


Fig. 2 The proposed scheme 1

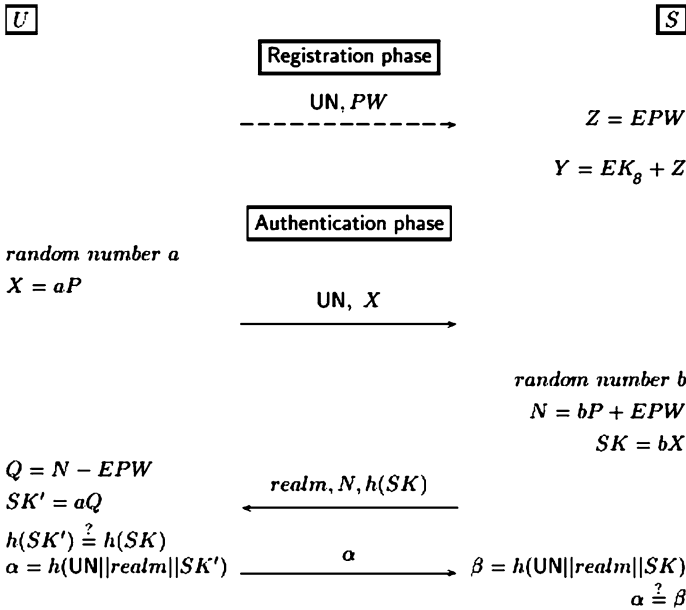


Fig. 3 The proposed scheme 2

Step 1. The user U generates a random number a and computes $X = aP$. Then U sends a login request message $\langle \text{UN}, X \rangle$ to the server S .

Step 2. When the login request arrives $\langle \text{UN}, X \rangle$, the server S first chooses the random number b and computes a secret session key $SK = bX$ and $N = bP$.

Step 3. Then S sends the response message $\langle \text{realm}, N, h(SK) \rangle$ to the user U .

Having received $\text{realm}, N, h(SK)$ from the server S , user U computes $SK' = aN$. The user U verifies the correctness of $h(SK)$ by checking that $h(SK')$ equals $h(SK)$. If correct, U accepts as the authentic server, computes $\alpha = h(\text{UN} \parallel \text{realm} \parallel \text{EPW} \parallel SK')$ and sends $\langle \alpha \rangle$; otherwise, aborts its login attempt.

Step 4. After receiving α , S computes $EPW = Y - Eks$ and $\beta = h(\text{UN} \parallel \text{realm} \parallel \text{EPW} \parallel SK)$. The server S checks that α equals β . If they are equal, S believes that he is talking to legal user. If they are not equal, S believes that he is talking to illegal user and aborts the scheme.

4.2 The Proposed Scheme 2

Authentication Phase. When the user U wants to log into the system, he enters his username UN and password PW into the client system.

Step 1. The user U generates a random number a and computes $X = aP$. Then U sends a login request message $\langle \text{UN}, X \rangle$ to the server S .

Step 2. When the login request arrives $\langle \text{UN}, X \rangle$, the server S first chooses the random number b and computes $N = bP + EPW$ and a secret session key $SK = bX$. Then S sends the response message $\langle \text{realm}, N, h(SK) \rangle$ to the user U .

Step 3. Having received $\text{real}, N, h(SK)$ from the server S , user U computes $Q = N - EPW$ and $h(SK') = h(aQ)$. U verifies the correctness of $h(SK)$ by checking that $h(SK')$ equals $h(SK)$. If correct, U accepts as the authentic server, computes $\alpha = h(\text{UN} \parallel \text{realm} \parallel SK')$ and sends $\langle \alpha \rangle$; otherwise, aborts its login attempt.

Step 4. After receiving α , S computes $\beta = h(\text{UN} \parallel \text{realm} \parallel SK)$. The server S checks that α equals β . If they are equal, S believes that he is talking to legal user. If they are not equal, S believes that he is talking to illegal user and aborts the scheme.

5 Security Analysis

We now figure out what is wrong with the scheme and how to fix it. The fixed scheme is given mainly to provide a better insight into the failure of Yoon et al.'s scheme. In this section, we only provide a heuristic security analysis of the proposed scheme, considering a variety of attacks and security properties.

Offline Password Guessing Attack. In this attack, an attacker may try to guess password and then to check the correctness of the guessed password off-line. If his guess fails, the attacker tries again with another password, until he finds the proper

Table 1 Comparison of security properties between our protocols and previously published two protocols

Security property	Tasi et al. [11].	Yoon et al. [8].	Our scheme 1	Our scheme 2
Off-line dictionary attack	Y	Y	Y	Y
Undetectable on-line password guessing attack	N	N	X	Y
Replay attack	Y	Y	Y	Y
Man in the middle attack	Y	Y	Y	Y
Denning-Sacco attack	Y	Y	Y	Y
Stolen verifier attack	Y	Y	Y	Y
Mutual authentication	Y	Y	Y	Y
Known-key security	Y	Y	Y	Y
Session key security	Y	Y	Y	Y
Perfect forward secrecy	N/A	Y	Y	Y

one. In the proposed protocol, the only information related to passwords is $\alpha = h(\text{UN}||\text{realm}||\text{EPW}||\text{SK}')$ but because SK' is secret value, this value does not help the attacker to verify directly the correctness of the guessed passwords. Thus, off-line password guessing attacks would be unsuccessful against the proposed protocol.

Undetectable On-Line Password Guessing Attack. At the highest level of security threat to password authenticated key exchange protocols are undetectable on-line password guessing attacks [9] where an attacker tries to check the correctness of a guessed password in an on-line transaction with the server, i.e., in a fake execution of the protocol; if his guess fails, he starts a new transaction with the server using another guessed password. Indeed, the possibility of an undetectable on-line password guessing attack in the three-or-more-party setting represents a qualitative difference from the two-party setting where such attack is not a concern. However, this attack is meaningful only when the server is unable to distinguish an honest request from a malicious one, since a failed guess should not be detected and logged by the server. In our scheme, the server is the first who issues a challenge and the client is the first who replies with an answer to some challenge. It is mainly due to this ordering that the scheme is secure against undetectable on-line password guessing attacks.

Implicit Key Authentication. The fundamental security goal for a key exchange protocol to achieve is implicit key authentication. Loosely stated, a key exchange protocol is said to achieve implicit key authentication if each party trying to establish a session key is assured that no other party aside from the intended parties can learn any information about the session key. Here, we restrict our attention to passive attackers; active attackers will be considered in the full version of this paper. Given aP and bP , the secret value $K = abP$ cannot be computed, since no polynomial algorithm has been found to solve the computational Elliptic Curve Diffe–Hellman problem. Thus, if the random numbers a and

b are unknown, then the session key sk cannot be computed since H is a one-way hash function. Hence, the secrecy of the session key is guaranteed based on the computational Elliptic Curve Diffie–Hellman assumption in the random oracle model [10].

In the Table 1, we compare the proposed schemes with previously published authentication schemes using the ten security properties. It is easy to see that our proposed authentication schemes can achieve all of the seven security properties.

6 Conclusion

This work has considered the security of Yoon et al.’s authentication scheme [8] for session initiation scheme. We demonstrate this by an undetectable on-line password guessing attack that completely compromises the password security of the scheme. Besides reporting the security problem, we proposed two secure authentication schemes over Yoon et al.’s scheme.

References

1. Veltri L, Salsano S, Papalilo D (2002) SIP security issues: the SIP authentication procedure and its processing load. *IEEE Netw* 16(6):38–44
2. Wikipedia. <http://en.wikipedia.org>
3. Franks J et al (1999) HTTP authentication: basic and digest access authentication. IETF RFC2617, June 1999
4. Handley M et al (1999) SIP: session initiation protocol. IETF RFC2543, March 1999
5. Thomas M (2001) SIP security requirements. IETF Internet Draft (draftthomas-sip-sec-reg-00.txt), Nov 2001 (work in progress)
6. Rosenberg J et al (2002) SIP: session initiation protocol. IETF RFC3261, June 2002
7. Arkko J et al (2002) Security mechanism agreement for SIP sessions. IETF Internet Draft (draft-ietf-sipsecagree-04.txt), June 2002
8. Yoon E-J, Yoo K-Y (2009) A new authentication scheme for session initiation protocol. International conference on complex, intelligent and software intensive system, pp 550–554
9. Ding Y, Horster P (1995) Undetectable on-line password guessing attacks. *ACM SIGOPS Oper Syst Rev* 29(4):77–86
10. Bellare M, Rogaway P (1993) Random oracles are practical: a paradigm for designing efficient protocols. In: *Proceedings of ACM CCS 1993*, pp 62–73
11. Tsai JL (2009) Efficient nonce-based authentication scheme for session initiation protocol. *Int J Netw Secur* 8(3):312–316
12. Yang CC, Wang RC, Liu WT (2005) Secure authentication scheme for session initiation protocol. *Comput Secur* 24:381–386
13. Diffie W, Hellman M (1976) New directions in cryptology. *IEEE Trans Inf Theory* 22(6):644–654
14. Durlanik A, Sogukpinar I (2005) SIP authentication scheme using ECDH. *World Enformatika Soc Trans Eng Comput Technol* 8:350–353
15. Koblitz N (1987) Elliptic curve cryptosystems. *Math Comput* 48:203–209
16. NIST (1999) Recommended elliptic curves for federal government use, July 1999

A Study on the Development of Security Evaluation Methodology for Wireless Equipment

Namje Park, Changwhan Lee, Kwangwoo Lee and Dongho Won

Abstract Recently, there has been an increased interest in wireless security equipment because of the proliferation of distributed wireless networks and wireless equipment. The security functions of wireless security equipment being used by organizations and companies provide remote users with access to a system after performing user identification and authentication, and allow encrypted data to be exchanged. However, since a consistent methodology for evaluating the vulnerability of wireless equipment has not yet been developed, it is difficult for evaluators and developers to properly evaluate the security of wireless equipment. To solve these problems, we propose an environment for vulnerability testing and outline trends in the development of wireless security equipment and security functions.

This research was supported by the Ministry of Knowledge Economy (MKE), Korea, under the “ITRC” support program supervised by the National IT Industry Promotion Agency (NIPA) (NIPA-2011-C1090-1001-0004). Also, this work was supported by grant project (No. 2011-034) of KISA.

N. Park
Department of Computer Education Teachers College,
Jeju National University, Jeju, Korea
e-mail: namjepark@jejunu.ac.kr

C. Lee · K. Lee · D. Won (✉)
Information Security Group, Sungkyunkwan University, Suwon, Korea
e-mail: dhwon@security.re.kr

C. Lee
e-mail: chlee@security.re.kr

K. Lee
e-mail: kwlee@security.re.kr

Keywords Wireless security equipment · Common criteria · Evaluation methodology

1 Introduction

There has recently been a convergence of the smart grid, smart work, and the smart phone in the IT industry. This convergence is carried out based on both wired/wireless networks. Wireless networks especially provide convenience to the organization and company by providing user authentication and data exchange using wireless equipment. However, wireless networks are subject to threats such as eavesdropping and message modification. Therefore, wireless network security is an important research area because it is directly related to confidentiality and privacy. For this reason, many security functionalities for wireless networks have been developed. In the case of IT security products, ISO/IEC 15408 (Common Criteria, referred to herein as CC) can provide a basis for security. CC guarantees that an IT security product is secure against specific threats and attacks. CC provides classes that provide information assurance per the standard. In part III of CC, the ATE and AVA classes define methods for carrying out security analysis that use functionality testing and vulnerability testing. To evaluate the security of wireless network equipment, the tests should be carried out according to a pre-defined test plan using specific items to be tested. However, security testing methods and a vulnerability test environment for wireless equipment have not yet been developed, causing the developers and evaluators to encounter difficulties. Moreover, various types of wireless security equipment continue to be released. For these reasons, security function tests and a vulnerability analysis methodology will be needed when carrying out equipment design, implementation, and tests.

To solve these problems, we will discuss the status of wireless security equipment and CC evaluation to construct a vulnerability test environment and to suggest a vulnerability test. The test environment and test items presented will be useful to the developer or evaluator who is doing development or evaluating wireless security equipment.

This paper is organized as follows. In [Sect. 2](#), we briefly discuss wireless security equipment, and security functions. [Section 3](#) is about the threats the equipment is faced with and the vulnerabilities of wireless security equipment, and we here propose a vulnerability test environment and a vulnerability test methodology. We summarize and conclude our research in [Sect. 4](#).

2 Related Work

In this section, we present the status of wireless security equipment and security.

Table 1 CC Evaluated wireless security equipment [9, 10]

Equipment	Manufacture	Evaluation assurance level
WS5100 Wireless Switch and RFS7000 RF Switch	Motorola	EAL 4
Cisco Systems Wireless	Cisco	EAL 3
Cisco Wireless Local Area Network (WLAN) Access System with Integrated Wireless Intrusion Detection System (WIDS)	Cisco	EAL 2
Fortress Secure Gateway (AF2100, AF7500, FC-X)	Fortress	EAL 3
Radimaster	SecureDataSystems	EAL 3
AirFront	AirCube	EAL 4
AGS-NPS	AirCube	EAL 2
AnyClick AUS	Unetsystem	EAL 4
PPX-AnyLink	Entrolink	EAL 3

2.1 Overview of Wireless Security Equipment

Wireless security equipment provides access control for remote wireless users who need to access a database storing important data or secret information. It also provides data encryption during data exchange between users and the system over a wireless connection. Therefore, wireless security equipment has been installed in public institutions, financial institutions, or companies where it is needed to control access to secure information and to protect personal and private information.

Wireless security equipment consists of three parts. First, there is a user authentication system for authenticating users of the wireless network. Second, there is equipment providing data encryption between users and the system over a wireless network, which realizes a wireless data transmission system. Finally, there is a solution using equipment that provides user authentication and data encryption for the wireless network. Table 1 describes the CC evaluation status of wireless security equipment according to Korea Evaluation and Certification Scheme (KECS) [1].

2.2 Security Functions of Wireless Security Equipment

In this section, we analyze the security functions of wireless security equipment. Figure 1 shows a wireless security system consisting of an administrator, a wireless user, an authentication server, and a log server [2–5].

Data security on the wireless section. Fundamentally, the wireless security equipment provides data exchange between users and equipment. However, data exchanged over a wireless network is subject to threats, such as eavesdropping, interception, and tampering. Therefore, it is necessary to provide security for data on a wireless network using wireless security equipment. For providing end-to-end

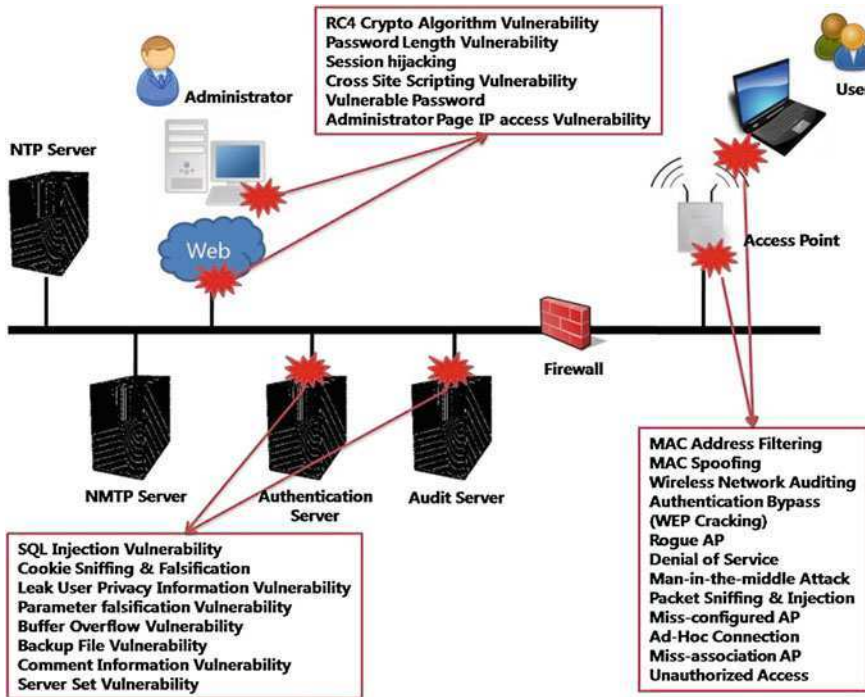


Fig. 1 Vulnerabilities and threats in Home Healthcare Service [2, 3]

data security on a wireless section, equipment uses the security mechanisms of EAP-TLS, EAP-FAST or WPA2-PSK and a 128 bit cipher, AES-CCM (CCMP).

System access control. System access control is necessary for wireless security equipment because the equipment is usually being operated at a company or public institution. Access to secret information or important data must be limited to authenticated or authorized users, and so wireless security equipment typically provides system access control. Examining the security functions thereof, all users are typically authenticated by an ID/Password method or a user MAC address identification method. Also, the user management and user authorized dynamic set use a centralized system.

User identification and authentication. Wireless security equipment performs user identification and authentication for wireless users. The wireless security equipment as stated in Table 1 provides user authentication and access control to protect against unauthenticated access. The equipment uses the authentication mechanisms of WiFi certified WPA2 which includes IEEE 802.1X port control. The equipment also stores the user ID and MAC address on a per session basis, and carries out private or public certification via an authority's authentication certificate.

Management user log. There has been a need for wireless security equipment to maintain a continuous user log. A system log has entries for key generation, distribution, deletions, information about related wireless endpoints, self-testing results, and encryption settings and changes to the system environment.

In Sect. 3, we will analyze the security functions and construction of a vulnerability test environment before providing a methodology for evaluating wireless security equipment.

3 Proposed Methodology for Security Evaluation

3.1 Vulnerabilities for Wireless Security Equipment

In order to analyze the vulnerabilities of wireless security equipment, we referred to vulnerability analysis web sites, related journals, and portal sites that offer vulnerability information. We compiled from the sources a list of threats to be used when analyzing vulnerabilities. Table 2 explains the possible vulnerabilities. Figure 1 depicts various vulnerabilities of wireless security equipment.

3.2 Testing Method for Security Functions of Wireless Security Equipment

The security functions that are implemented by the wireless security equipment should be operating properly. Moreover, to guarantee that it has properly eliminated the possibility of security holes, an evaluator performs vulnerability tests for testing/analyzing different security functions and uses an evaluation methodology specific to the wireless security equipment to be tested. The general tests and analysis method are composed of the five steps of preparing to test, making test plan, building a test environment, performing a functional specification test, and analyzing test results. Each of these test steps must take into account the specific vulnerabilities. TOE security environment [6].

Preparing to test. Preparing to test is a prerequisite to performing a test. The first step is to clearly establish the goals of the test. The goal of evaluating wireless security equipment is to identify that the security functions of wireless security equipment are operating perfectly. After clearly establishing the goals of testing, we specify how to write up the analysis, which can be done using the three techniques of an informal technique, a semi-formal technique, and a formal technique. Examples of the various techniques are natural language description as an informal technique, using diagrams or tables as a semi-formal technique, and mathematical description as a formal technique. Vulnerability evaluation preparation includes drawing up a vulnerability list.

Table 2 Vulnerabilities of wireless security equipments [11, 12]

Vulnerability	Explanation
MAC Address Filtering	Acquire MAC address or falsification through wireless network sniffing
MAC Spoofing	Copy internal AP MAC address and SSID for external AP faked same as internal AP
Wireless Network Auditing	Acquire information through wireless network auditing
Authentication Bypass (WEP Cracking)	Attack wireless packet security function WEP, and acquire WEP key through attack
Rogue AP	Set rogue AP at internal for malicious purpose, and leak internal information or attack system
Denial of Service	Send massive packet to AP or system for denial of service
Unauthorized Access	Unauthorized user access to system
Man-in-the-middle Attack	Malicious attacker fake valid server between users and valid server, so he acquire communication information between users and server
Packet Sniffing & Injection	Eavesdropping between valid users communication
Miss-configured AP	Unauthenticated user can connect system, because of non-security set of AP
Ad-Hoc Connection	External unauthorized user illegal connection with internal network user using Ad-Hoc method
Miss-association AP	To leak information of company, attacker lead internal user connected other network using that wireless equipment automatic find SSID and connect property
RC4 Crypto Algorithm Vulnerability	Leak information on wireless network, using cipher algorithms vulnerability
Password Length Vulnerability	Using password length's vulnerability, attacker do social engineering hacking to get key value
Session hijacking	Attacker intercepts active session after valid user session was operated. So attacker can observe and control all operation
Cross Site Scripting Vulnerability	If user operate page, script which was transferred attacker's code would operate
SQL Injection Vulnerability	Attacker force injection SQL command to target database, and operate data leak, falsification, or administrator authentication bypass
Cookie Sniffing & Falsification	Attackers steal or control cookies in users' web browser
Leak User Privacy Information Vulnerability	User privacy information was leaked from database
Parameter falsification Vulnerability	Attacker modify normal system parameter to occur abnormal operation
Buffer Overflow Vulnerability	Attack input over buffer size of data, to purpose of operating malicious command
Backup File Vulnerability	Vulnerabilities of backup files like .bak which was generated during developed server
Comment Information Vulnerability	A comment generated during developed system possibly leak system information
Vulnerable Password	Vulnerabilities of below secure password length or possibly guess combination password
Server Set Vulnerability	Vulnerabilities of default server set value which was not changed
Administrator Page IP access Vulnerability	Attacker can unauthorized access to system because IP access control dose not operate on administrator page

Table 3 Test Document Form [6]

Test List	Explanation
Test goal	A goal that we do test or vulnerability evaluation test
Test environment	A environment that we do test or vulnerability evaluation test
Test subordinate relationship	Specify a test preceded this test or vulnerability evaluation test
Test processes	Detail process of this test or vulnerability evaluation test
Expected result	Expected result of this test or vulnerability evaluation test
Actual result	Actual result of this test or vulnerability evaluation test

Table 4 Illustration of vulnerability test document

Test number	VA.3	Test vulnerability	Denial of service
Test purpose		This test measures wireless security equipment toward Denial of Service attack	
Test environment		Authentication Server System: MS Win 2000 Server, CPU: Intel III Xeon Memory: 1,024 Mb, HDD: 120G SCSI Disk Device Authentication Client Windows System: Microsoft Windows XP CPU: Inter Core i7, Memory: DDR2 2G, HDD: 60G Network Adapter: 802.11n Wireless network card	
Test subordinate relationship		Nothing	
Test processes		<ol style="list-style-type: none"> 1. Operate authentication server and authentication client 2. Operate ‘airdump’ network traffic recorder 3. Perform denial of service attack using tool ‘Stick’ 4. Measure unusual network traffic state using ‘airodump’ 5. Perform denial of service attack using tool ‘Synk4’ 6. Measure unusual network traffic state using ‘airodump’ 	
Expected result		The service provided by wireless security equipment and wireless network will be normal state or little increase traffics despite of denial of service attack tools	
Actual result		Result of attack using denial of service attack tools, the wireless security equipment was normal state when we performed ‘Stick’ tool. However, when we performed ‘Synk4’ tool, equipment state’s is abnormal and service interrupt occurred	

Drawing up a test plan. The test plan is put in place after preparations have been put in place. To establish a test plan, the equipment to be used in the test must be identified. The components of wireless security equipment generally used in testing are a wireless security server, wireless security client, authentication server, log server, and the users.

Build test environment. The build test environment is established after a test plan has been established. First, we build a test wireless security server and client. Second, we specify a vulnerability test list and the detailed security functions that are to be used in the test environment. The application programs and driver information must be clearly recorded.

Functional specifications test. In the functional specification test, the wireless security equipment is actually tested. The functional specification test covers security functions and the vulnerability test list of wireless security equipment. The security functions of wireless security equipment are categorized into data encryption for the wireless section, user identification and access control, and management of the user audit log. The vulnerabilities are categorized into attack tests which are established for each vulnerability. Table 3 describes the form of a test document [7].

Test result analysis. When all the test steps have been completed, the test results are analyzed. The conformity of test results is analyzed in order to provide assurance that all security functions and all items on the vulnerability list were tested. Finally, if the expected results are different from the actual results, those items should be modified or supplemented. Table 4 is an illustration of a vulnerability test document [8].

4 Conclusion

In the coming years, the use of wireless security equipment will widen to cover general users at companies and institutions. The information handled by the security equipment at these locations will accordingly become more and more important. At the same time, the number of detected vulnerabilities that wireless communication and security equipment will face will also increase. Therefore, we provide a wireless security equipment methodology so that an evaluator can efficiently evaluate wireless security equipment. In the future, a detailed list of the items that wireless security equipment should be tested for and a detailed test process will be required.

References

1. Common Criteria for Information Technology Security Evaluation, version 3.1, CCMB-2006-09
2. National Cyber Security Center (2008) Wireless authentication system protection profile
3. US Government Information Assurance Directorate (2007) Wireless local area network client protection profile for basic robustness environment
4. Common Criteria Portal Certified Products Security Target. <http://www.commoncriteria.portal.org>
5. National Cyber Security Center Certified Products Certified Report. <http://service1.nis.go.kr/>
6. Lee K, Won D, Kim S (2011) A secure and efficient E-Will system based on PKI. *Inf Int Interdiscip J Int Inf Inst* 14(7):2187–2206
7. Lee Y, Kim S, Won D (2010) Enhancement of two-factor authenticated key exchange protocols in public wireless LANs. *Elsevier Comput Electr Eng* 36(1):213–223

8. Lee K, Lee Y, Won D, Kim S (2010) Protection profile for secure E-voting systems. In: Proceedings of ISPEC 2010, Information security practice and experience conference 2010, LNCS 6047. Springer, Seoul, Korea, pp 386–397, 12–13 March
9. Common Criteria Portal Certified Products. <http://www.commoncriteriaportal.org/products/>
10. National Cyber Security Center Certified Products. <http://service1.nis.go.kr/>
11. CVE, Internet site for vulnerability analysis. <http://cve.mitre.org/>
12. Andrew A, Konstantin V, Andrei A (2004) WI-FOO: the secrets of wireless hacking, Pearson, Upper Saddle River
13. Park N, Song Y, Won D, Kim H (2008) Multilateral approaches to the mobile RFID security problem using web service. In: Zhang Y, Yu G, Bertino E, Xu G (eds) APWeb 2008. LNCS, vol 4976. Springer, Heidelberg, pp 331–341
14. 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
15. Park N, Kim S, Won D (2007) Privacy preserving enhanced service mechanism in mobile RFID network. In: ASC, Advances in soft computing, vol 43. Springer, Heidelberg, pp 151–156
16. Park N (2010) Security scheme for managing a large quantity of individual information in RFID environment. In: CCIS, Communications in computer and information science, vol 106. Springer, Heidelberg, pp 72–79
17. Park N, Kim S, Won D, Kim H (2006) Security analysis and implementation leveraging globally networked mobile RFIDs. In: PWC 2006. LNCS, vol 4217. Springer, Heidelberg, pp 494–505

Computer Application in Elementary Education Bases on Fractal Geometry Theory Using LOGO Programming

Jaeho An and Namje Park

Abstract This paper suggested a way of using LOGO programming, the educational programming language elementary school students can easily learn, and fractal geometric theory for computer education in elementary school in accordance with the theme of creativity, the educational goal of elementary school curriculum. Future curriculum of computer education will include the areas of algorithm and programming. As using an educational programming language is an integral part of algorithm and programming education, research should be carried out urgently regarding the use of programming languages. When LOGO programming is taught with fractal geometric theory, students can easily understand mathematical notions e.g., regularity, repetition, similarity and resemblance. Therefore, it is expected that students will be able to learn LOGO programming more effectively when it is taught in connection with what should be taught in math curriculum such as figures, measurement, regularity and problem solving.

Keywords LOGO · Fractal geometry · Elementary computer education

This paper is extended from a conference paper presented at the journal of Korean Institute of Information Technology (Vol.9, No.8). The author is deeply grateful to the anonymous reviewers for their valuable suggestions and comments on the first version of this paper.

J. An · N. Park (✉)
Department of Computer Education, Teachers College,
Jeju National University, 61 Iljudong-ro,
Jeju-do, Jeju-si 690-781, Korea
e-mail: namjepark@jejunu.ac.kr

J. An
e-mail: profirean@jejunu.ac.kr

1 Introduction

Recently, computer education in Korea has been criticized for focusing on applying computer software. Accordingly, there is an increased sense that it is necessary to enhance student creativity, logical thinking, and problem solving when teaching programming languages. Also, since teaching programming languages has focused on the talented students and not on the average students, the methods of educating the average student have received insufficient attention. As such, the methods of educating the average student should be developed further.

Computer programming has many distinct features. First, teaching programming not only provides an understanding of programming languages but also has an impact on skills in other areas, such as enhancing problem solving, logical thinking, and creativity. In light of this, when students study a programming language, the students are expected to improve not only in the computer field, but in other fields as well. Therefore, in this paper, we are going to suggest an easy method of providing computer education to average students wherein elementary school computer education is combined with fractal geometry theory and the LOGO programming language. Furthermore, we are going to pave the way for studying how to apply this method using various materials.

2 Using LOGO Programming in Education

LOGO is an educational computer programming language which uses functional programming. The educational aspect of LOGO lies in the distinctness of its features. First, LOGO commands are cyclic, procedural, easy and simple. Secondly, LOGO commands and procedures are interactive and a user can directly view on the screen the results of processing commands. Third, it elevates student interest.

2.1 *Logical Thinking*

Elementary school students have the tendency to think and act instinctively when they are faced with certain problems; however, if they study algorithms when studying programming, the ability to think systematically can be enhanced. One has to think algorithmically especially when composing procedures with LOGO, making it extremely helpful for developing logical thinking when it is applied to elementary school computer education. This is because to resolve various errors that occur in the progress of composing procedures, the students have to think about why the error occurred. They then have to modify the procedure and try it again. Overall, this process enhances their ability to think logically.

2.2 Enhancing Creativity and Problem Solving

In programming, there are several ways to compose a procedure even when the goal is the same. Elementary school students can therefore complete a study project using a variety of approaches. This is beneficial because composing a procedure in their own way will enhance their creativity. In addition, the process of composing a procedure may force them to solve problems that arise. Trying to solve their errors in their own creative way will also enhance their creativity and ability to solve problems.

3 School Computer Education Using LOGO and Fractals

3.1 Linking Computer Education to Elementary School Mathematics Courses

The ultimate goal of elementary school computer education is using computer education to enhance creativity, problem solving ability, and logical thinking. The same goal applies to mathematics education. A lot of the learning experiences present in computer education can help students achieve the goals of their mathematics education, and vice versa. Therefore, linking elementary school mathematics education to elementary school computer education can produce powerful synergy.

The five sections of elementary mathematics education are numbers and calculations, figures, measurements, probability and statistics, and finally regulation and problem solving ability. Elementary school computer education applied to LOGO and fractals can be connected to several of these sections as shown in Table 1.

First, the basic commands used to compose fractal procedures in LOGO can be related to figures and measurements. Using simple movement commands and changing a turtle's angle allows the angles of simple figures to be understood. Moreover, students can learn how to measure sections by measuring the angles on the screen. There is also a third benefit of using polygons to improve the student's ability to understand figures.

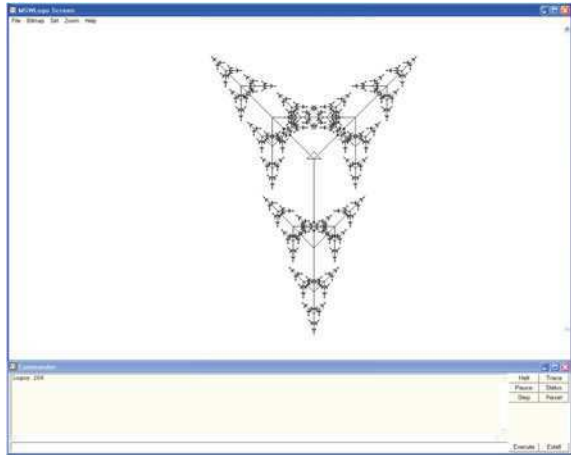
Second, recursive calls can be related to the learning of ratios which is related to problem solving. Changing the size of the factor in recursive calls sets the ratio by which to reduce a fractal. The ratio can be confirmed with the eyes because the change in size of the fractal depending on the ratio will be printed on the screen. This learning process allows ratios which are a part of regulation and problem solving ability to be learned.

Third, learning fractals can be connected to learning regulations which is a part of problem solving. It is possible to learn about regulations when using procedures to realize fractals because fractals have the features of regulation, repetitiveness, and similarity. Also, part of regulations and problem solving is learning how to make regular patterns and so learning this section is very helpful and useful.

Table 1 Connection among curriculum of elementary math, LOGO and education of fractal

Grade	Scope and contents	LOGO programming
4	<ul style="list-style-type: none"> • Shape <ul style="list-style-type: none"> - Different angles and triangles 	<ul style="list-style-type: none"> • Basic commands <ul style="list-style-type: none"> - fd :size lt 90 bk :size/2 - fd :size lt 135 bk :size*2/3
4	<ul style="list-style-type: none"> • Measurement <ul style="list-style-type: none"> - Angle 	
4	<ul style="list-style-type: none"> • Pattern and problem solving <ul style="list-style-type: none"> - Making pattern - Pattern and response 	<ul style="list-style-type: none"> • Fractal geometry (Figs. 1 and 5) reference
5	<ul style="list-style-type: none"> • Shape <ul style="list-style-type: none"> - Congruence - Symmetry 	
5	<ul style="list-style-type: none"> • Pattern and problem solving <ul style="list-style-type: none"> - Ratio 	<ul style="list-style-type: none"> • Recursive call <ul style="list-style-type: none"> - Logot :size/2 - Logoy :size*2/3

Fig. 1 One Y fractal geometry screen developed by LOGO programming



3.2 Development of Education Method for Y Shaped Fractal Geometrical Theory

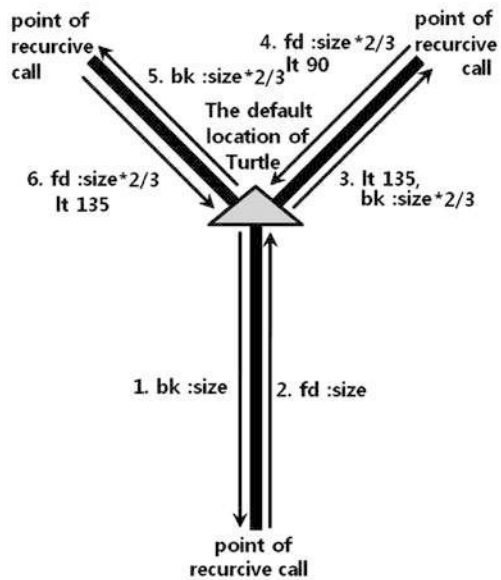
The fractal form suggested in Fig. 1 is designed based on the English alphabet shape. This procedure is designed in a way that each part is drawn by turtle from the cross point in alphabet Y. If the recursive call is conducted at each corner of Y, a fractal form with a regularity and self similarity is made. If the fractal form is different from the basic form, it would be difficult for an elementary school student to understand. Therefore, the total shape is made in a way that it is similar to the basic form.

Analysis of newly developed Y shaped fractal and algorithm is as follows. The ‘Logo-Y’ procedure is composed of basic commands which an elementary school child may easily understand. The Y fractal patterns designed in this study are as follows if they are expressed in Logo language procedure (Fig. 2).

Fig. 2 Y fractal procedure screen developed by LOGO programming

```
to logoy :size
  if :size<2 [stop]
  bk :size
  logoy :size/2
  fd :size
  lt 135
  bk :size*2/3
  logoy :size/2
  fd :size*2/3
  lt 90
  bk :size*2/3
  logoy :size/2
  fd :size*2/3
  lt 135
end
```

Fig. 3 Fundamental figure of Y fractal in LOGO programming



As the newly developed Y fractal procedure is designed to make fractal form using the alphabet Y as its basic form, the understanding of the basic form is required. To make the shape of Y exposed well, the angle between left and right branches is made 90°. Accordingly, the Y fractal is a little different from T fractal in terms of angle and length. T fractal uses the right angle while Y fractal uses 135 and 90° to form 360°. It means that the Y fractal requires more careful thinking than the T fractal. The basic form, which is the foundation of Y fractal, is as shown in Fig. 3.

The 'Logo-Y' procedure uses the basic command and three recursive calls. Accordingly, the basic command and recursive call can be repetitiously learned using this fractal. The flow chart for this procedure can be expressed as shown in Fig. 4.

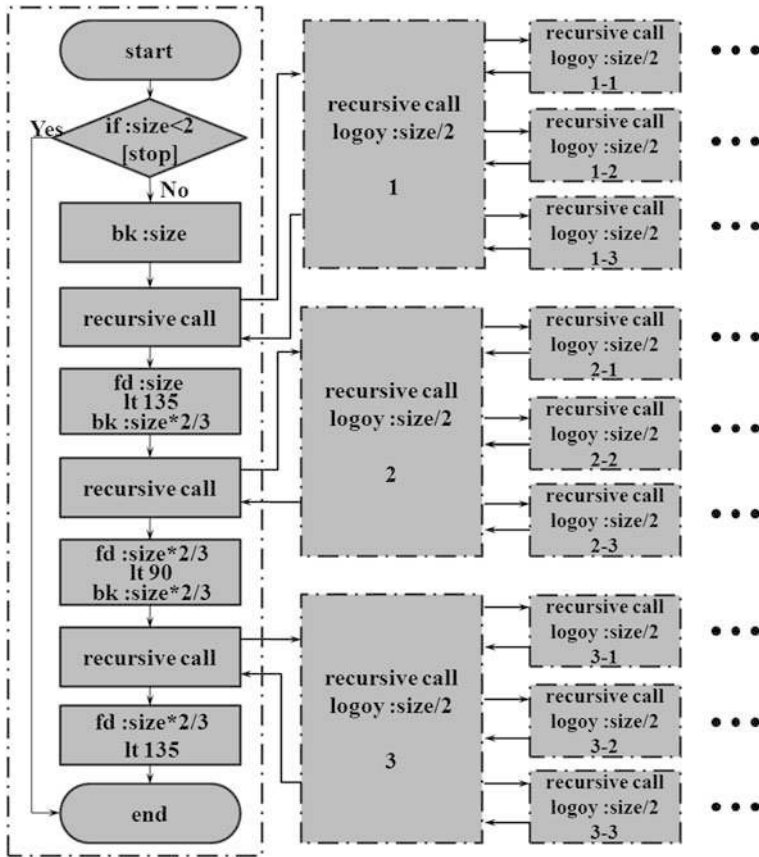


Fig. 4 Flow chart of Y fractal procedure

If the procedure is called along with size factor, check if the input size factor is less than 2. If so, the procedure is finished. If not so, the next command is to be conducted. Unlike the T fractal procedure, its branches do not make the angle of 90° but 135°, making turtle move more backward. Then, the recursive call is conducted. The recursive call makes this size reduced to two-third of the original size and the command of 'bk:size' is conducted until the procedure is in compliance with the conditions of stop in 'if' sentence. If the size becomes less than 2 after several recursive calls, the procedure stops and the command after the first recursive call is conducted for the size just before becoming less than 2. If the command is conducted like this, the bottom part of alphabet Y fractal is first completed and then the right and left parts are made before the fractal is completed. The procedure for making the fractal is as shown in Fig. 5.

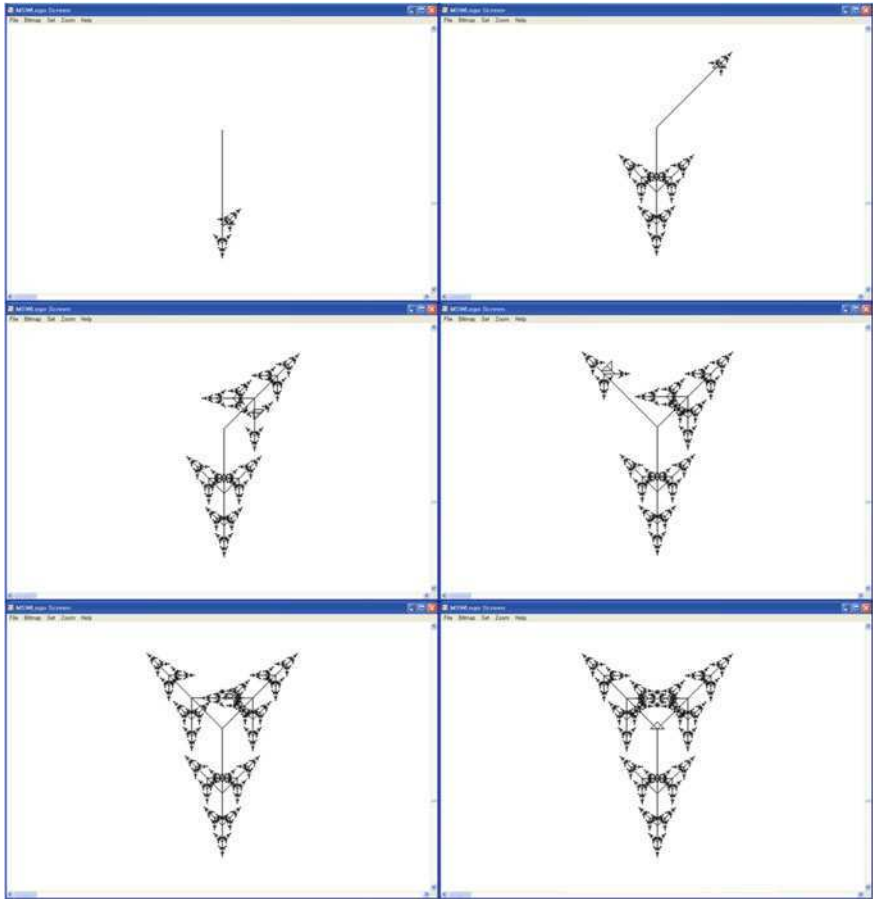


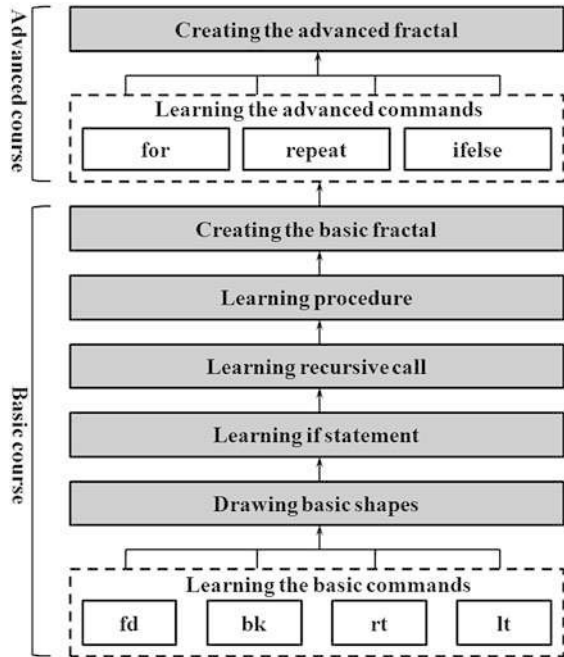
Fig. 5 Process of Y fractal procedure in LOGO programming

3.3 *Methods of Applying LOGO to Elementary School Computer Education*

We drew up lesson plans for a 4 h computer education course that used LOGO and fractal theory. The goal was focused not on studying the basic functions of programming but on getting the maximal effect from applying the minimum of functions. The lesson plan is shown in Fig. 6.

In the first lesson, we studied basic LOGO commands and the concept of fractals. In the second lesson, we made and studied Basic-Y fractals using LOGO. In the third lesson, we made Y fractals which were more difficult than Basic-Y

Fig. 6 Flow of study



fractals. In the last lesson, we made fractal procedures using figures which were covered in the first to third lessons and that the students had learned and in their own way. The overall flow of study was composed with basic and intensified process for graded lesson.

First, the level of interest of the students has to be taken into consideration. It is possible that students will lose interest when the teacher teaches basic commands, and accordingly, the teacher has to minimize the presentation of functions. Also, the teacher can increase student interest by providing various applications which are suitable for the level of the students. Secondly, the teacher has to emphasize procedures and recursive calls because they are crucial to making fractals. Third, when the teacher teaches the Y fractal, intervention has to be minimized to offer plenty of opportunities because it applies what was learned about Basic-Y fractals. Lastly, advanced lessons should be given to students who get good results, and remedial education should be given to the other students.

4 Conclusions and Tasks

What is important in programming is not results but the process whereby they are achieved. Regardless of the result, learners will do many activities in the process of programming. During this process, problem solving and the ability to think

logically will be strengthened. These advantages of learning programming also apply to elementary school students. Therefore studying programming is very important in elementary school.

We have introduced the beneficial effects of applying LOGO and fractals to elementary school computer education. One of the merits of the programming language is that the effects gained from studying it can be made stronger by connecting it to the material of other courses. So if we a variety of content that has been connected with other subjects is taught by relating it to a programming language, not only can we teach the programming language more effectively but it will also be beneficial for students as they will learn other subjects. As a result, the methods of utilizing various materials in programming language and computer education have to be studied.

References

1. Harvey B (1997) Computer science logo style. MIT Press, Cambridge
2. Park N (2011) Implementation of terminal middleware platform for mobile RFID computing. *Int J Ad Hoc Ubiquitous Comput*
3. Park N, Song Y, Won D, Kim H (2008) Multilateral approaches to the mobile RFID security problem using web service. In: Zhang Y, Yu G, Bertino E, Xu G (eds) *APWeb 2008, LNCS 4976*. Springer, Heidelberg, pp 331–341
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 3842*. Springer, Heidelberg, pp 741–748
5. Leron U (1983) Some problems in children's Logo learning. In: *Proceedings of the 7th international conference for the psychology of mathematics education, Israel*
6. Noss R (1984) Children learning logo programming: interim report no. 2 of the chiltern logo project. *Advisory Unit for Computer Based Education, Hatfield*
7. Park N, Kim H, Kim S, Won D (2005) Open location-based service using secure middleware infrastructure in web services. In: Gervasi O, Gavrilova ML, Kumar V, Laganá A, Lee HP, Mun Y, Taniar D, Tan CJK (eds) *ICCSA 2005, LNCS 3481*. Springer, Heidelberg, pp 1146–1155
8. Park N, Kim S, Won D (2007) Privacy preserving enhanced service mechanism in mobile RFID network. In: *ASC, Advances in soft computing, vol 43*. Springer, Heidelberg, pp 151–156
9. Park N (2010) Security scheme for managing a large quantity of individual information in RFID environment. In: *CCIS, Communications in computer and information science, vol 106*. Springer, Heidelberg, pp 72–79
10. Park N (2008) *Reliable system framework leveraging globally mobile RFID in ubiquitous era*. Ph.D. thesis, Sungkyunkwan University, South Korea
11. Park N, Song Y (2010) Secure RFID application data management using all-or-nothing transform encryption. In: *WASA 2010, LNCS, vol 6221*. Springer, Heidelberg, pp 245–252

Construction of a Privacy Preserving Mobile Social Networking Service

Jaewook Jung, Hakhyun Kim, Jaesung You,
Changbin Lee, Seungjoo Kim and Dongho Won

Abstract The social-network application comes on, as the smart phone has come into wide use. The Social Network Service is making huge effect to the new relationship among the people. With this new wave, the development of the social-network application which is based on the smart phone is activated. Some of the social network application including the service, allow users to search other users who is close to the current location, for example, “WhosHere” of iPhone. These applications provides user checker which is based on the information of the other users such as age, gender, interest. However, this method demands the congruity of the information, i.e., the information of the searcher and the surveyee must

This research was supported by the Ministry of Knowledge Economy (MKE), Korea, under the “ITRC” support program supervised by the National IT Industry Promotion Agency (NIPA) (NIPA-2011-C1090-1001-0004).

J. Jung · H. Kim · J. You · C. Lee · S. Kim · D. Won (✉)
Information Security Group, School of Information
and Communication Engineering, Sungkyunkwan University,
300 Cheoncheon-dong, Jangan-gu, Gyeonggi-do, Suwon 440-746,
Republic of Korea
e-mail: dhwon@security.re.kr

J. Jung
e-mail: jwjungdhwon@security.re.kr

H. Kim
e-mail: hhkimdhwon@security.re.kr

J. You
e-mail: jsyoudhwon@security.re.kr

C. Lee
e-mail: cbleedhwon@security.re.kr

S. Kim
e-mail: skim71@korea.ac.kr

perfectly match. Improving this aspect, in this paper, we propose a method to enhance privacy of social networking service, while preserving its original objective as a social hub. The proposed method is based on fuzzy vault scheme; the main contribution of our scheme is that the matching ratio that sets the degree of information correlation is variable and can be set by user.

Keywords Fuzzy vault · Matching ratio · Privacy · Social networking service (SNS) · Smart-phone

1 Introduction

The social-network application comes on, as the smart phone has come into wide use. The Social Network Service (SNS) is making huge effect to the new relationship among the people. With this new wave, the development of the social-network application which is based on the smart phone is activated. Some of the social network application including the service allow users to search other users who is close to the current location, for example, “WhosHere” of iPhone. These applications provides user checker which is based on the information of the other users such as age, gender, interest. However, this method demands the congruity of the information, i.e., the information of the searcher and the surveyee must perfectly match. Improving this aspect, in this paper, we propose a method to enhance privacy of social networking service, while preserving its original objective as a social hub. The proposed method is based on fuzzy vault scheme. The organization of this paper is as follows. In [Sect. 2](#), we briefly describe fundamental knowledge on social networking service, fuzzy vault scheme. In [Sect. 3](#), we present our proposed method, and explain the architecture of our system. We analyze our proposed scheme in [Sect. 4](#), and summarize and conclude our research in [Sect. 5](#).

2 Related Work

2.1 Social Networking Service

A social networking service usually refers to an online service, in which people can establish and build social networks or relations amongst each other. Most social networking services are web based and provide means for users to interact over the internet, such as e-mail and instant messaging. Although online community services are sometimes considered as a social networking service in a broader sense, social networking service usually means an individual-centered

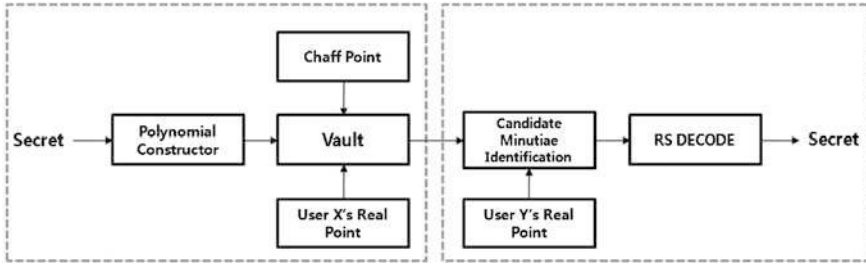


Fig. 1 Fuzzy vault scheme

service whereas online community services are group-centered. Social networking sites allow users to share ideas, activities, events, and interests within their individual networks.

2.2 Fuzzy Vault Scheme

Fuzzy vault scheme proposed by Ari Juels and Madhu Sudan is a privacy-preserving protocol [5]. The protocol is divided into two parts: vault creating and vault unlocking (Fig. 1).

The vault creating part is as follows:

1. The scheme makes a polynomial using user X’s secret, the information that user X intends to protect.
2. User X creates real point and chaff point.
3. User X puts real point into a function poly() to make a vault. Chaff points are substituted with arbitrary values which have not been mapped into the polynomial.

The vault unlocking part is as follows:

1. The scheme takes user Y’s real point (user Y may be the same person as user X) and compares them with user X’s real point to collect equivalent point. As many points are identical, more point will be collected from the vault.
2. The scheme makes a polynomial using point collected from the vault through comparison process. In this step, not all of real point may be the same, and a few chaff points are collected instead. This minor error is put through “RS (Reed–Solomon) DECODE” process, resulting in recovering the original polynomial. If the error is significant, the recovered polynomial will not be accurate.

Table 1 Notations

Notation	Description
Fuzzy encryption	Processing of creating a vault
Fuzzy decryption	Processing of unlocking a vault
Preference	User can select these n-preferences
Real point	Denotes the polynomial coordinate (x, y) of selected user's preferences
Chaff point	Set of random point for the purpose of security (for protect real point)
Vault	Set of Hash value (hashing the polynomial coefficients) and Points(real & chaff)
Hash value	Made of coefficients which apply to SHA-1 algorithm
User information	Nickname, Preferences, Distance, Matching Rate

3 Proposed Scheme

In this article, we propose a new architecture, which compensates for short comings of previous social networking applications.

Fuzzy vault scheme used in fingerprint recognition needs the error correcting process for reducing the error. However, our scheme does not use existing error correcting process. Instead, we take the notion of matching rate in our scheme. In addition, proposed fuzzy vault scheme in this paper is used as protocol to find the friend that have the same input information. Above all, there is considerable meaning that fuzzy vault concept is used in the social networking service (Table 1).

In this section, we describe the privacy enhanced social networking service based on fuzzy vault scheme. The following notations are used throughout this paper.

3.1 Entire Scheme Overview

Figure 2 depicts our system structure.

In Fig. 2, The procedures are briefly described as follows:

Step 1: In this phase, a process that User X sets his nickname, preferences, distance and matching rate. After setting User X's information, vault is made using his information (that is a series of encryption process).

Step 2: At this step, user X submits these information to server to be registered (initial connection phase).

Step 3: User X sends a vault to other users via server, and then requests to other users for decryption.

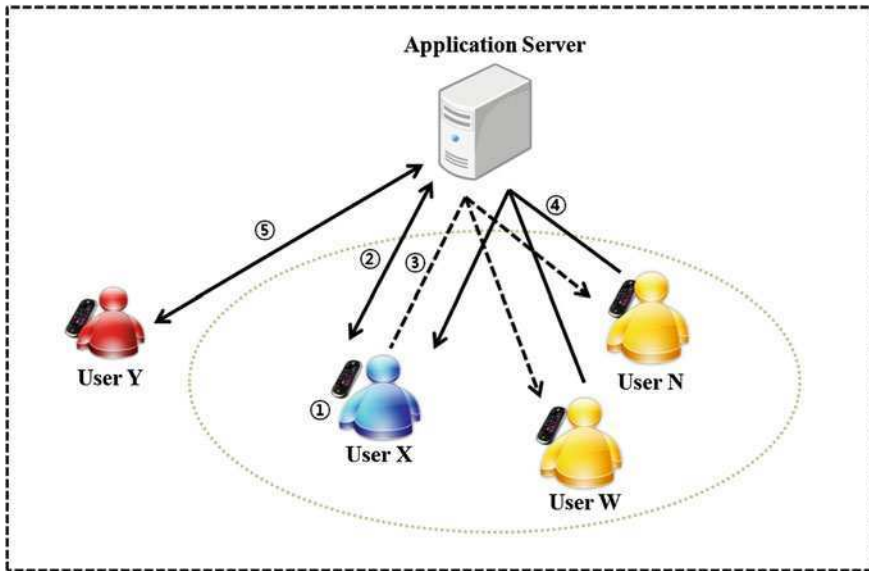


Fig. 2 System structure

Step 4: Vault is unlocked by other users, and then transmits result to user X via server. If the result is success, user X can add other users to his friends list (join protocol phase).

Step 5: When the application is finished, user Y disconnects from the system and server (leave phase).

For more details, we describe in Sects. 3.2 and 3.3.

3.2 Basic Structure of Our Scheme

Fuzzy Encryption, Fuzzy Decryption

Our construction supplements the problems that previous applications have, and enhances the security at the same time.

Figure 3 represents a flowchart that shows fuzzy encryption process and decrypting process.

First, User X selects k -preferences of total n -preferences according to his taste. Polynomial coefficients are randomly picked up and then polynomial is created.

Second, calculate real & chaff points through the created polynomial. Third, concatenate their coefficients and put into SHA-1 to make a hash value. Through these processing, a vault is created finally. A vault consists of this hash value, matching rate, real points, and chaff points.

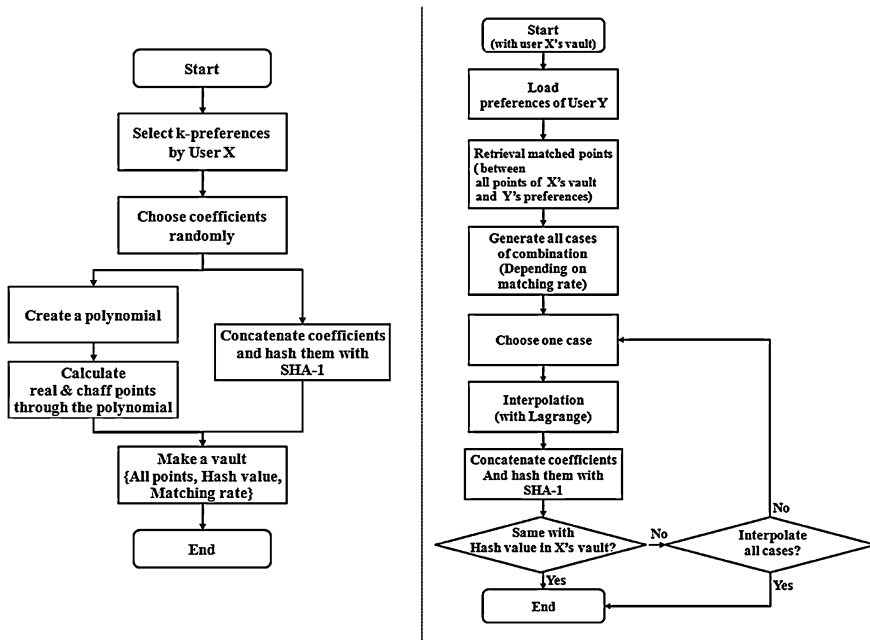


Fig. 3 Flowchart of fuzzy encryption process and decrypting process

In decrypting process, upon receiving a vault from user X, user Y unlocks the vault. First, load preferences of user Y and retrieval matched points between all points of user X’s vault and user Y’s preferences. Second, generate all cases of combination based on set of matching rate. Third, choose the one of some case and create polynomial through Lagrange interpolation based on chosen case. Then concatenate their polynomial coefficients and put into SHA-1 to make a hash value. This hash value is then compared with a hash value in user X’s vault. If they are identical, the decryption process is successfully completed, and otherwise, decryption process keep continuing until equal to these two hash value.

3.3 Protocol

In this section, we describe three-protocol based on fuzzy vault: initial connection protocol, join protocol, leave protocol. In addition, for understanding our proposed protocol, we will set k-parameter value; n-parameter value and matching rate value: $k = 5$, $n = 40$ and matching rate = 60%.

Initial Connection Phase

Initial connection protocol is a primary process for communicate each other properly. For example, user X selects 5 preferences from a list which contains 40

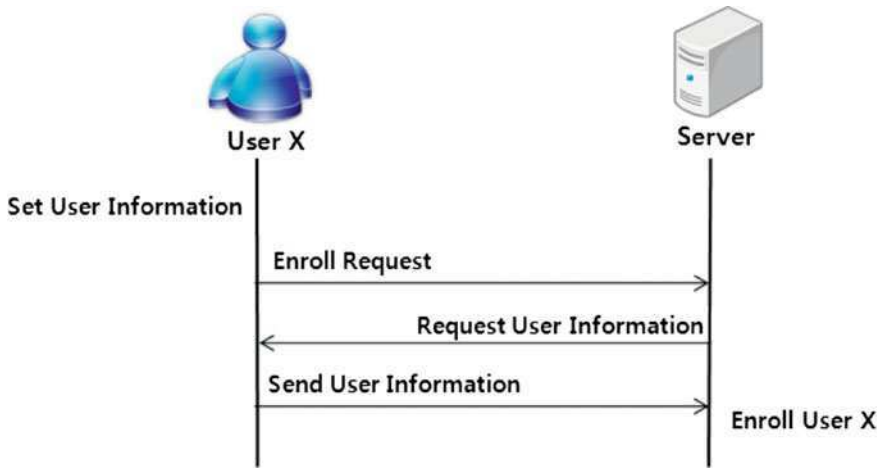


Fig. 4 Initial connection protocol

preferences. Selected preferences are then mapped with 5 points with preset coordinates. These 5 points are real points.

After user X selects preferences, fuzzy encryption process is run. The figure below shows initial connection phase; a process that user X is registered to the server. Figure 4 shows the initial connection protocol. Initial connection protocol proceeds as follows. First, User X sets his nickname, preferences, distance and matching rate and creates a vault, and then submits that information to server to be registered. When server requests user X for his information, user X sends user information to server. The server registers user X using received information and creates his ID.

Join Protocol Phase

We have seen the initial connection protocol. However, we have to consider the case for addition of a new user.

If user W selects 5 preferences, user W would create his vault depending on his information. In case user X want to find friends, user X's vault send to user W. We assume that User W already connect to server. User W's preferences are picked out by using user X's real points and matching rate in vault. At this time, because of 60% matching rate, user W's 5 preferences are calculated number of cases through ${}_5C_3$ operation. In a similar method, polynomial is created by user W's preferences, and then extracts polynomial coefficients. Then concatenate their polynomial coefficients and put into SHA-1 to make a hash value, and then compare with hash value in user X's vault and user W's hash value. Finally, if these two values equal to each other, fuzzy decryption process is successfully completed.

Figure 5 shows a protocol in which user W is a new user and user X is adding user W to his friends list. In this protocol, User Y is a user who is out of boundary and excluded from communication. We assume that user X and user Y are already

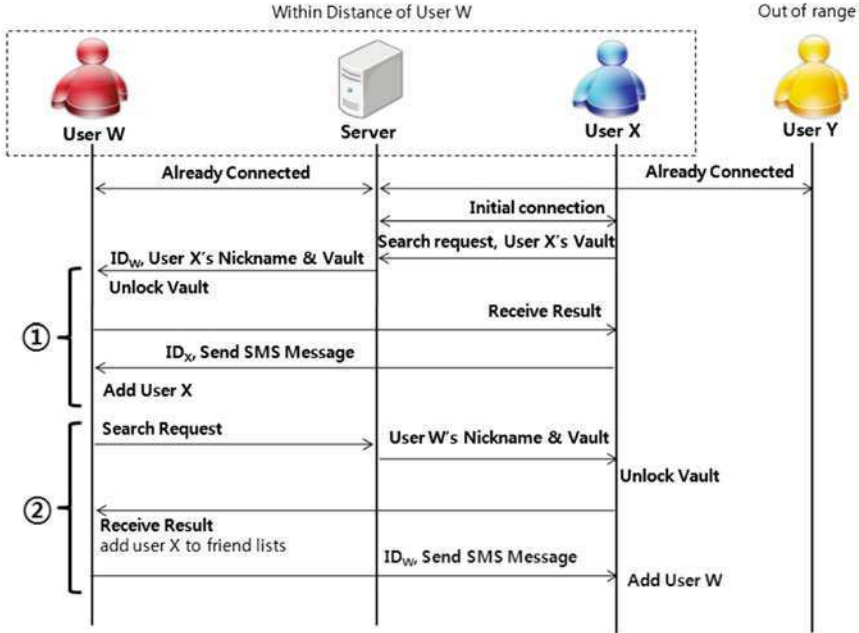


Fig. 5 Protocol of friend addition process

connected to server. Also, we assume that server and user Y are already in initial connection state. There are two methods for this procedure.

Case 1

One of the methods is user X adding user W to his friends list. When user X is in initial connection status, user X sends join request & user X's vault to server, and the server sends user X's nickname and vault to user W, whose inbound of specific distance. User W decrypts received vault and sends the result to user X. If the result is positive, user X and user W are now friends.

Case 2

Another method is User W adding User X to his friends list. When User W requests server for refresh, the server sends User W's nickname and vault to User X. Then user X decrypts received vault, and sends the result of decryption to user W. Then user W adds user X to his friend list. Now user W and user X are set for message transfer.

Leave Phase

Figure 6 shows the protocol how implements the system when user disconnects the system. It is simple comparing the user inserting process. There are two methods for this procedure.

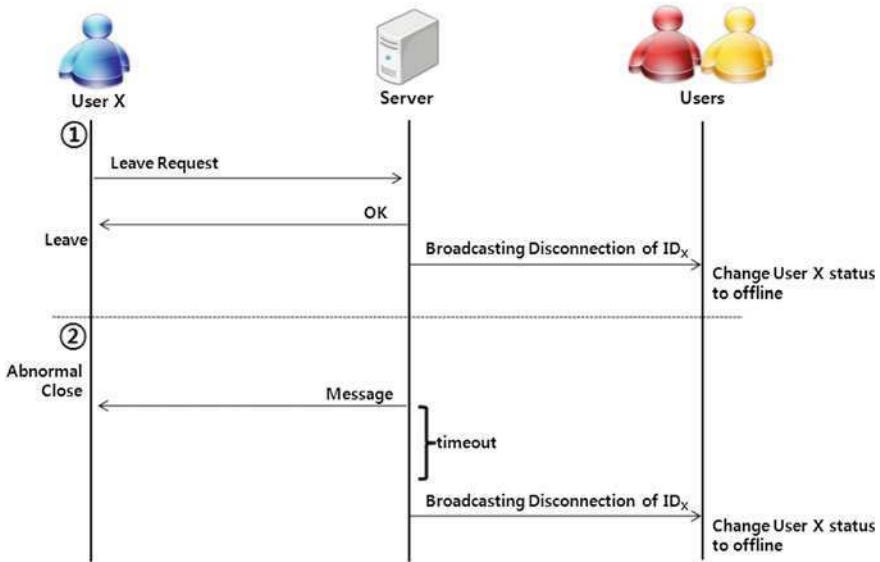


Fig. 6 Disconnecting protocol

Case 1

User X requests a server “leave request” and Server disconnects User X’s connecting state. And then, translates information to another user that User X’s state is disconnection. Finally, another user is modified information that condition of User X is “off-line” state.

Case 2

In this case, if user X abnormally terminates and at the same time, other users send a friend request or a text message. Obviously, user X will not respond, since user X is no longer available. The server waits for user X’s response for a while (a specific time period), then lets other users know that user X’s session has been terminated. Upon receiving the message from the server regarding termination, other users set user X’s status as “off-line”.

4 Analysis

4.1 Security Analysis

The privacy preservation capability of our method is rather intuitive. Since our method provides users with capability to set various matching ratio, one cannot be sure about which personal interests is common between his/her and the friend, unless one sets the matching ratio as 100%. Thus, if the application with our

Table 2 Comparison analysis in communication aspects

	Using fuzzy vault	Not using fuzzy vault
Points	$37 * 4 \text{ bytes (integer value)} * 2$	Not use
Index	Not use	$10 * 4 \text{ bytes (integer value)}$
Hash value	20 bytes	20 bytes
Matching rate parameter	4 bytes	4 bytes
Total	312 bytes	60 bytes

method prohibits setting the matching ratio as 100%, one cannot know the exact information about his/her friend.

4.2 Communication Analysis

Server Phase

In our proposed method, the role of server is solely a communication channel. Since we developed a client oriented distributed processing application, no computation is processed in the server. That is, the computational cost for the server is 0.

Client Phase

We analyzed the size of data both using fuzzy vault and not using fuzzy vault. Basically, a fuzzy vault consists of 37 indexes and their polynomials, hash value, and matching rate parameters. Thus, calculation of data size of using fuzzy vault scheme is as follow:

- $37 * 4 \text{ bytes (integer value)} * 2 + 20 \text{ bytes (hash value)} + 4 \text{ bytes (matching rate parameter)} = 312 \text{ bytes}$.

In case of not using fuzzy vault scheme, it consists of 10 index and mating rate parameter. Thus, calculation of data size of not using fuzzy scheme is as follow:

- $10 * 4 \text{ bytes (integer value)} + 20 \text{ bytes (matching rate parameter)} = 60 \text{ bytes}$.

In addition, the packet size (312 bytes) is significantly smaller than allowable size stated in 802.11standard (2,272 bytes), meaning that the data size is not a problem (Table 2).

5 Conclusion and Future work

This paper yields the disadvantage of the existing social-network applications and suggests the option to solve the problem by applying concept of fuzzy vault. The social network application which is suggested on this paper makes it possible to

find other users who have similar tastes, even though the preferences of each user do not perfectly match. As a consequence, the network between the users tends to be broadened. In the future, we will work on making the social networking application based on our proposed method.

References

1. Baltensperger R (2000) Improving the accuracy of the matrix differentiation method for arbitrary collocation points. *Appl Numer Math* 33(1–4):143–149
2. Berrut JP, Trefethen LN (2004) Barycentric Lagrange interpolation. *Soc Ind Appl Math* 46(3):501–517
3. Facebook’s Privacy Policy. <http://www.facebook.com/policy.php>
4. Gross R, Acquisti A (2007) Information revelation and privacy in online social networks. In: *Proceedings of the 3rd ACM workshop on privacy in the electronic society*, pp 71–80
5. Juels A, Sudan M (2006) A fuzzy vault scheme. *Des Codes Cryptogr* 28(2):237–257
6. Balachander K, Wills CE (2008) Characterizing privacy in online social networks. In: *Proceedings of the first workshop on online social networks*, pp 37–42
7. Miluzzo E, Lane ND, Eisenmau SB, Cambell AT (2007) CenceMe—injecting sensing presence into social networking applications. In: *Proceedings of the 2nd European conference on smart sensing and context*, pp 1–28
8. Moon D, Choi W, Moon K (2009) Fuzzy fingerprint vault using multiple polynomials. *J Korea Inst Inf Secur Cryptol* 19(1):125–133
9. Strahilevitz LJ (2004) A social networks theory of privacy. *American law and economics association annual meetings*, no. 230

Part III
IT-Agriculture Convergence

Standardization Trend of Agriculture-IT Convergence Technology in Korea

Se-Han Kim, Chang Sun Shin, Cheol Sig Pho, Byung-Chul Kim
and Jae-Yong Lee

Abstract Recently, USN for the Agriculture-IT convergence is used for the automation for Greenhouse and Plant Factory. These are mainly composed of the sensor and actuator, control gateway, operating system, and energy resources. The efforts for the standardization to improve of productivity of the crop, decrease in labor and reduce of the investment costs are progressed. In this paper, we introduce the standardization trend of USN-centered Agriculture-IT Convergence in Korea.

Keywords Greenhouse · Plant factory · Vertical farm · USN · Agriculture-IT Convergence · WSN

S.-H. Kim · C. S. Pho
RFID/USN Research Division, ETRI, Daejeon, Korea
e-mail: shkim72@etri.re.kr

C. S. Pho
e-mail: cspyo@etri.re.kr

C. S. Shin
Department of Information and Communication Engineering,
Sunchon National University, Suncheon, Korea
e-mail: csshin@sunchon.ac.kr

B.-C. Kim (✉) · J.-Y. Lee
Department of Information and Communication Engineering,
Chungnam National University, Daejeon, Korea
e-mail: byckim@cnu.ac.kr

J.-Y. Lee
e-mail: jy1@cnu.ac.kr

1 Introduction

The Agriculture-IT convergence aims at the high-tech agriculture infrastructure technology development for the drawn Green Industry to become number one of nation. For this purpose, the gradual foundation of development adapting the new agricultural paradigm through the substantial efficiency consideration of the agricultural production activity, parts industry, and mutual complementary technology of the agriculture through the energy and the service problem solving and IT is prepared with the object [1–5].

The agriculture in which the IT technology is applied can be classified as outdoor culture, the greenhouse and plant factory.

In the bare ground, Agriculture-IT technology utilizes the ubiquitous sensor network including the monitoring of the crop and planting condition and does the simple control watering, and etc. The relative technique undergoes the difficulty to the product engineering deficit of the system unit and investment comparison productivity deficit.

The greenhouse Agriculture-IT technology does the crop improvement of productivity through the various environmental control including the inside of the greenhouse and external environment, rearing information monitoring of the crop, watering of the greenhouse, nutrient solution, CO₂, side wall window, roof window and light source, and etc. It is in the commercialization introduction stage through the research of technique by crop and the normalizing of the integrated control technology putting the energy and manufacture efficiency an emphasis is progressed.

The plant factory can do in being the integrated material of the Agricultural IT in which the IT technology including the various sensor and control elements energy source, service and construction, and etc. is applied. The plant factory works the complete control factory type agriculture reached to the various forms (the single story and multilayer) with the object and can be said to be the bloom of the future agriculture and the main issue is the application of the IT technology for the securing economical efficiency for the crop production through the energy and reduction of manpower.

In this paper, we introduce the standardization trend of Agriculture-IT centered about USN in KOREA. In the article “[Spam Host Detection using Ant Colony Optimization](#)”, we describe the Agriculture-IT driving strategy of standardization in USN. The article “[Location Estimation of Satellite Radio Interferer Using Cross Ambiguity Function Map for Protection of Satellite Resources](#)” illustrate the status of standardization for Environmental Control and Monitoring System for agriculture automation in Greenhouse. The article “[Korean Voice Recognition System Development](#)” tell the status of standardization for Plant Factory. Finally, the conclusion and future works are presented.

2 Agriculture-IT Driving Strategy of Standardization in USN

The standardization by application is the initial step with the technology which USN is applicable in the various field including the national defense, manufacture, construction, traffic, medical treatment, environment, education, agriculture, and etc. Particularly, USN is the representative IT technology for applying to the agriculture, can do in being the composite of the various technology including the service model, application, middleware, network, sensor node, and etc.

In Korea, USN related standardization propels the requirements gathering of the industry through the various technical standardization forums including USN forum, IP-USN forum, u-City forum, and etc. and TTA RFID/USN standardization group (PG 311), IPv6 standardization group (PG 210) information and communication group standard. In addition, the international standardization is pursued based on the adopted standard in the International Organization for Standardization including ISO, ISO the/IEC JTC 1, ITU-T, IETF, IEEE, and etc.

In the agriculture, USN focuses that the productivity of the crop is enhanced and the labor is minimized through the monitoring and optimized controlling through the sensor and controller.

The standard for the Agriculture-IT convergence was real initiated from the field of USN since 2010 in Korea. And the following existing standards are utilized in the agriculture field.

“Sensor Node Identification Code and Data Structure” in TTA Standard provides the definition, structure, generation rule and procedure of sensor node identification code. In here, it explains the definition of S-Code (sensor node identification code) and deals with the structure of S-Code such as Issuing Agency Code (IAC), Company Code (CC), Prefix, Usage Code (UC), and Serial Code (SC) and explains and provides detailed examples about those parts of S-Code. In addition, it is to identify, manage and distribute the sensor and sensing information more easily by using the unique identifier of wire and wireless sensor nodes in the ubiquitous environment. “Hierarchical Identification Scheme for Sensor nodes: hCode” in TTA Standard provides hierarchical identification scheme for sensor nodes considering sensor network environment and provide interoperability of sensor network services [6].

“The Standard Interface for Heterogeneous Sensor Networks” in USN Forum Standard includes the communication protocol, message formats between sensor network and host. In addition, it defines some standard sensor data types. By using this communication protocol, message formats and sensor data types, this standard provides sensor network abstraction. “Plug and Play based USN Sensor Access Interface Standard” in USN Forum show the physical interface and Transmission protocol between sensor platform and sensor module, the Standard API for Sensor Access, the HAL library for Sensor Device Driver, and the Standard Reference model [7] (Fig. 1).

In addition these standard, it uses international communication standard including the IEEE802.15.4/4e/4 g, ZigBee, 6LoWPAN in IEEE and IETF.

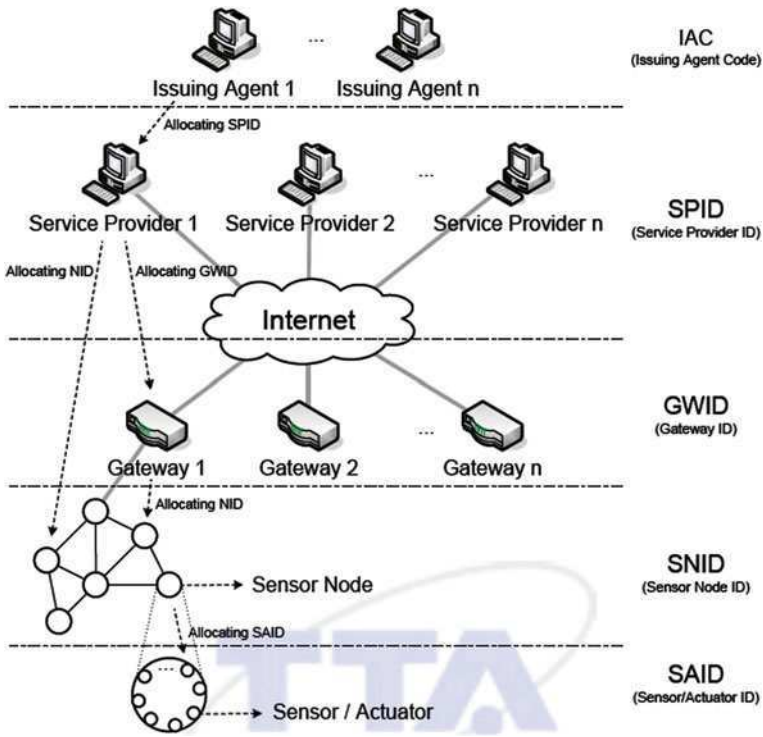


Fig. 1 Hierarchical identification at sensor network

3 The Status of Standardization for Environmental Control and Monitoring System in Greenhouse

3.1 Standard Scope in Greenhouse

The standard of the greenhouse control system defines the components in applying IT technologies to a greenhouse and specifies the requirements and the architecture for the technological issues. The system collects information for the growth management of crops and can control the facilities promoting the optimal growth environments in greenhouse. This system includes the growth environment management service, the growth environment control service, and etc.

Since 2010, the started fields of standardization are the interface standards among the sensor node, actuator node, control gateway, operating system, and the management system. Particularly, there is it lowers the cost of the green house equipment through the interface standard of the related inter system, it facilitates the various information exchange including the production, distribution, control, etc. (Fig. 2).

The standard of the greenhouse control system is planning to be completed until August 2011 six sections as follows;

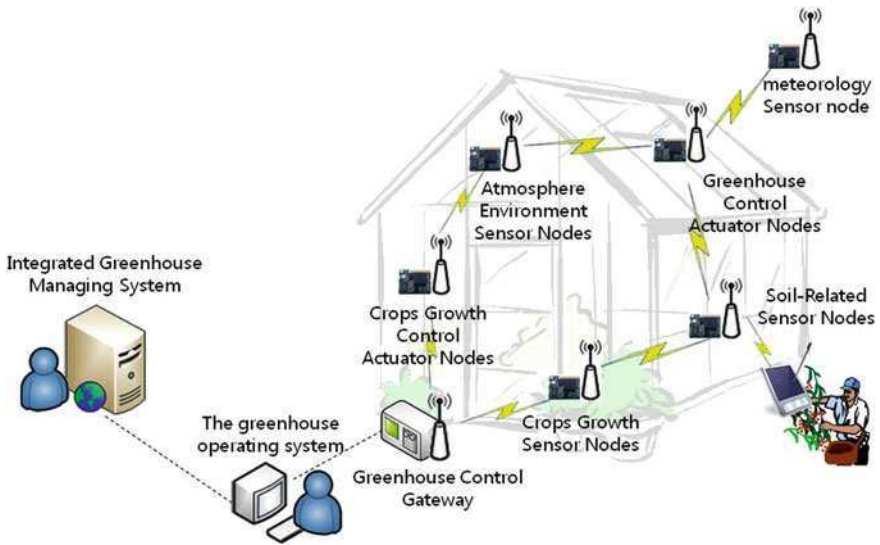


Fig. 2 System component in greenhouse

- Requirements profile for environmental control and monitoring system in greenhouse
- Greenhouse control system-Part 1: Interface for between sensor nodes and greenhouse control gateway
- Greenhouse control system-Part 2: Interface for between actuator nodes and greenhouse control gateway
- Greenhouse control system-Part 3: Interface for between greenhouse control gateway and greenhouse operating system
- Greenhouse control system-Part 4: Interface for between greenhouse operating system and integrated greenhouse management system
- Sensor data specification for greenhouse control

In addition, the standardization including the Service Structure and Service Specification, Service Management Specification, System Structure, Interface between Greenhouse Operating System and Energy Component, Energy saving and Management Data Specification, Energy Interface Specification for Energy rotation in Greenhouse, etc. is planning to be progressed from the third quarter of 2011 (Fig. 3).

3.2 Environmental Control and Monitoring Standard System in Greenhouse

The greenhouse environment monitoring is the service which shows the information collected from the sensor node to the user in order to check the inside of the

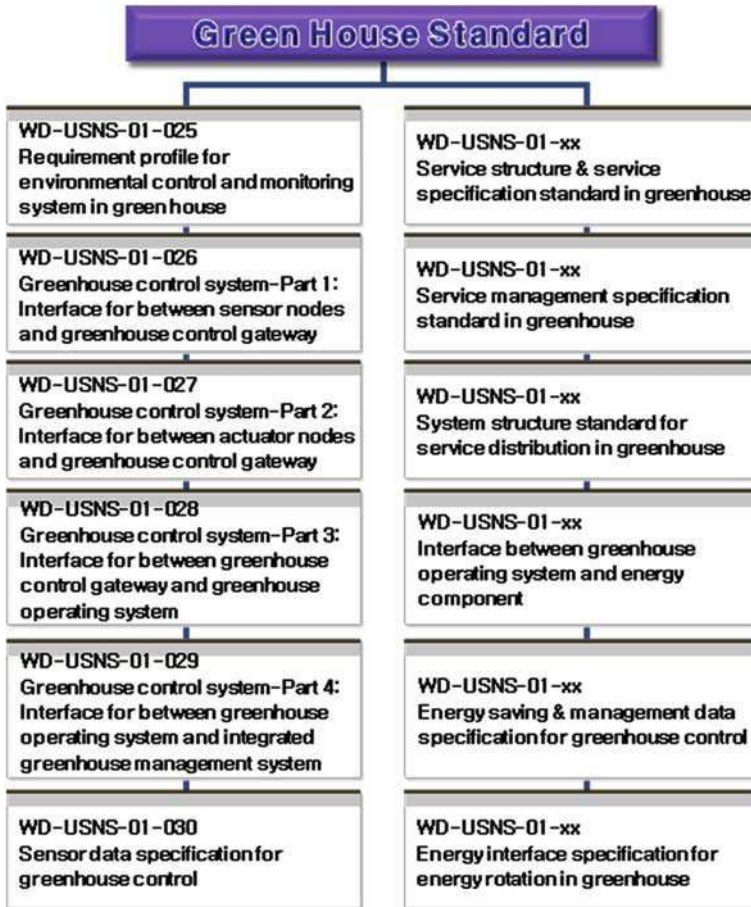


Fig. 3 Standard scope in greenhouse

greenhouse and external status. The monitoring objects including these standards are (1) information outside the greenhouse like climatic environment, light, temperature, humidity, direction and velocity of the wind, rain information, (2) information inside the greenhouse like atmospheric condition the light, temperature, humidity, carbon dioxide density, and (3) rhizosphere environment like the soil, hydroponics, and culture medium. The System collects the related sensing information of sensor nodes according to the decided periodic time or reports the sensing information of the aperiodic sensor nodes by the specific condition. For system operation, each system has the function of directly inputting which the user collects the audio, image and text. Also it display the real-time environmental information indication, graph about the user-specification period, period average and standard deviation indication, and the environmental information indication of out of control and warning for the dangerous situation of growth environment.

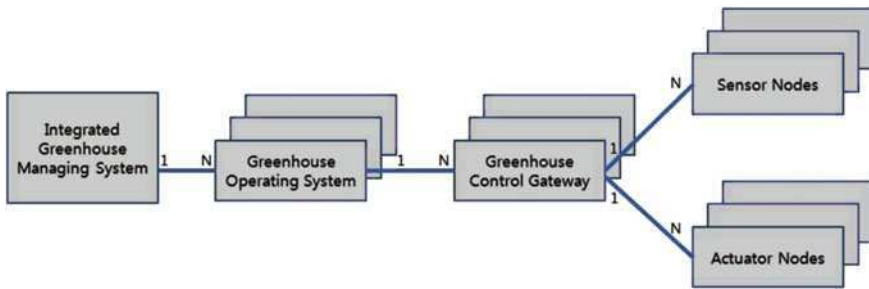


Fig. 4 Structure chart of the environmental control and monitoring standard system in greenhouse

The greenhouse environment automatic control is the service controlling automatically the control node based on the sensing information which is collected through the sensor node. The automatic control objects including these standards are (1) the air ventilation control through the side/roof window and ventilating PAN, (2) the temperature control of with hot water boiler, electric heater, air-conditioner and heat-pump, (3) the humidity control including dehumidifier, humidifier, (4) the watering control with drip-watering, nutrient supply instrument and sparkling machine for controlling moisture, pH, EC and humidity, (5) the amount of light control with the shading film and artificial/secondary light source regulator, and (6) the Co2 control inside of the greenhouse. The automatic control is supported the automatic control function through the optimal algorithm and the manual control function by user. For system operation, greenhouse control system has the function to the automatic saving function of the control record, the processing of the automatically collected information and input function, the display the control-sate presently, etc.

The greenhouse operation management is the service management method required of the greenhouse operation. And it operates with the external data server for the collection including the environment and the crop growth data, and etc. The operation management objects including these standards are the greenhouse profile management, the software installation at the sensor/actuator nodes, the rearing database management for each crop, the setting up the controlled environment in greenhouse control system, the operation and feedback with operation system, agriculture diary production and archiving facility, and etc.

As to Fig. 4, the interface of the greenhouse system elements moreover the standardization continues with figure showing the necessary present status of interface above various elements.

4 The Status of Standardization for Plant Factory

The standard of the Plant Factory defines the elements in applying IT and BT technologies and specifies the requirements and the architecture for the technological issues. The system collects information for the growth management of

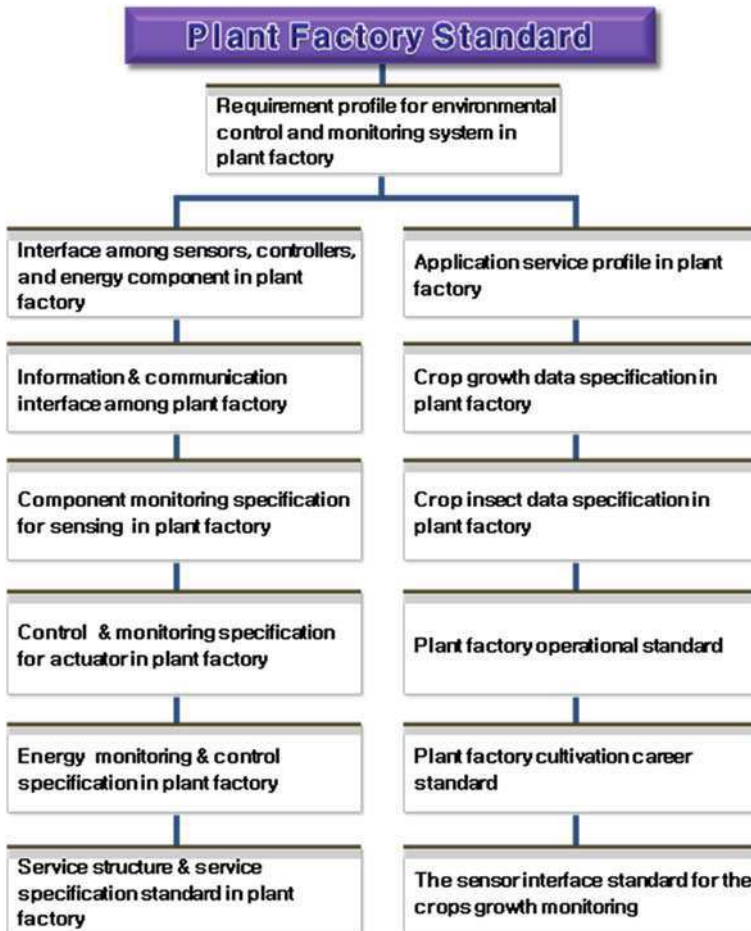


Fig. 5 Standard scope in plant factory

crops and can control the facilities promoting the optimal growth environments in vertical farm. Specially, the control, administration, and economic type integrated building environment control platform of component based is efficiently operating with utilizing the heat-pump and solar cell, vertical and horizontal equilibrium growing apparatus, environmental control with smart phone, precision nutrient solution, automation robot for the porting, harvest, and packing, and the software of rearing and insect prevention.

In 2011, ETRI, and Agriculture-IT related enterprise, USN Forum, Rural Development Administration (RDA) and the Ministry of Knowledge Economy (MKE) in Korea progress the standardization together with testbed in RDA. Figure 5 is the standard scope in 2011.

In addition, the Plant Factory standards include the various requirement profiles about the organization of the standard and range, establishment of the term, application service requirement (the inside of plant factory and external monitoring, automatic control, energy management, and operation management), component-to-component requirement, component-to-component interface definition, and etc. The service profile standard is the procedure and definition for installation of the plant factory and administrator, facility man and general user. It is also comprehensive of the setting up, data polling, aperiodic the sensor data transmission, emergency notice, monitoring, controlling, management process and etc. required for the optimization of the components of the Plant Factory. To expand the information of Plant farm, they have the facility operating, the energy management, the exchange of the sensor information and control information required for the broad area interface (Inter-working) between the Plant Factory falling to the local crop growth information.

5 Future Works

In this paper, we review the USN standard for the Agriculture-IT Convergence used in the agricultural fields including the Greenhouse and Plant Factory. These standards use together the existing USN standard like Sensor Node Identification Code, Data Structure and Communication Technology. The standard including the profiles, interfaces, data specification, management specification, energy technology, service, and etc. according to the characteristic of the Greenhouse and Plant Factory was initiated at Korea in 2010. In case the middleware and interface is provided through these standards, the farmer and companies build the new facility, the time and cost is minimized. The difficulty to operation and maintenance will be able to be solved too. This is very important for the country, company, local government, and farmer. In the future, the new standard should be developed about not only the standardization about the agricultural devices but also the new service and valuable business model. And continued research and development are comprised.

Acknowledgments This work was supported by the Industrial Strategic technology development program, 10037299, Development of Next Generation Growth Environment System and 10040125, Development of the Integrated Environment Control S/W Platform for Constructing an Urbanized Vertical Farm funded by the Ministry of Knowledge Economy (MKE, Korea).

References

1. Pawlowski A, Guzman JL, Rodríguez F, Berenguel M, Sánchez J, Dormido S (2009) Simulation of greenhouse climate monitoring and control with wireless sensor network and event-based control. *Sensors* 9:232–252
2. Janos S, Istvan M (2009) Distance monitoring and control for greenhouse systems via internet, Kopaonik Srbija Zbornik radova konferencije Yuinfo, pp 1–3

3. Matijevis I, Simon J (2010) Control of the greenhouse's microclimatic condition using wireless sensor network. *IPSI J TIR* 6(2):35–38
4. Hwang J, Shin C, Yoe H (2010) Study on an agricultural environment monitoring server system using wireless sensor networks. *Sensors* 10:11189–11211
5. Moon A, Li S, Kim K (2011) Components based integrated management platform for flexible service deployment in plant factory. Part I, *HCI International*, pp 524–528
6. Telecommunications Technology Association. <http://www.tta.or.kr>
7. USN Forum, <http://www.usnforum.or.kr>

Design and Implementation of Greenhouse Control System Based IEEE802.15.4e and 6LoWPAN

Se-Han Kim, Kyo-Hoon Son, Byung-Chul Kim
and Jae-Yong Lee

Abstract Recently, USN are becoming an important solution to agriculture automation. This paper describes the implementation and configuration of USN using IEEE802.15.4e and 6LoWPAN for the Greenhouse Control System which is comprised of the Sensor/Actuator Node, Greenhouse Control Gateway, Greenhouse Operating System and the Integrated Greenhouse Managing System. We apply IEEE802.15.4e and 6LoWPAN to our system in order to overcome the expandability and reliability of the Greenhouse having many control elements. Unlike unstable communication of ZigBee, this system supports the quality of service for the serious performance degradation by the frequent retransmission generated by the increasing of traffic and has the timeliness delivery of the sensing information.

Keywords USN · Agriculture-IT convergence · Greenhouse · Plant factory

S.-H. Kim · K.-H. Son
RFID/USN Research Division, ETRI, Daejeon, Korea
e-mail: shkim72@etri.re.kr

K.-H. Son
e-mail: sonkh@etri.re.kr

B.-C. Kim (✉) · J.-Y. Lee
Department of Information and Communication Engineering,
Chungnam National University, Daejeon, Korea
e-mail: byckim@cnu.ac.kr

J.-Y. Lee
e-mail: jy1@cnu.ac.kr

1 Introduction

Ubiquitous Sensor Network, USN technology connects the sensor-devices with all things, and develops the value added service. It has us share the information between the person and thing or thing and thing anytime and anywhere. This USN technology applies to the various field including the national defense, manufacture, construction, traffic, medical treatment, environment, education, physical distribution, etc. and recently utilizes for the agriculture, and is studied for the convenience and productivity. Generally, the systems utilizing USN technology served by processing great quantity of data that the wireless sensor network collects. Recently, in the field of USN, it stretches in the sensor-centered service. It expands to the field controlling the various controllers based on the information through the sensor. Particularly, in the case of the field of controlled agriculture, the efficient control including not only the inside of the Greenhouse and external sensing but also the window control, illumination (light), ventilation machine, and hot blast heater, etc. is important. Recently, the requirement of this market is reflected and the TG4e standard [1] enhancing IEEE802.15.4-2006 [2] is progressed in IEEE802.15.4 WG15. By using wired protocols like PLC, RS-232C, RS485, etc., the existing agricultural system generally controls the various control elements but the convenience for equipment expandability and installation is insufficient.

In this paper, for the reliability and expandability in Greenhouse automation show the implemented system by using the IEEE802.15.4e designed according to the factory automation characteristic. The Greenhouse Control System using 6LoWPAN in which the end-device setup and control is possible in the foreign network tries to be introduced. In the article "[Spam Host Detection using Ant Colony Optimization](#)", the other research and the IEEE802.15.4e and 6LoWPAN standard for the agricultural automation in which this paper becomes the core of the technology to be proposed. The article "[Location Estimation of Satellite Radio Interferer Using Cross Ambiguity Function Map for Protection of Satellite Resources](#)" illustrates the complete system, each detailed technique and implementation issue. Finally, the conclusion of this paper is presented.

2 Related Research for the Agriculture Automation

2.1 Automation System for Greenhouse

For a couple of years, the research about the various agriculture automation technologies had been being progressed. Generally, as to the agricultural technique, for confirming the result and having the various sensing and control elements according to the characteristic of the weather and crops hang much time. Recently, the various technical attempts utilizing the Ubiquitous Sensor Network

are progressed [3–5]. USN is a collection of sensor and actuators nodes linked by a wireless medium to perform distributed sensing and acting tasks. The sensor nodes collect data and communicate over a network environment to a computer system [6, 7]. The Environment of Greenhouse mainly influence on the crop growth. For the purpose of the optimum rearing environment, we utilize possible various technologies [8–12].

2.2 IEEE802.15.4e for Wireless Control

Presently, the PHY/MAC standard is determined based on the IEEE802.15 standard for the Wireless Personal Area Network (WPAN) and the ZigBee technology to apply to the related industries with the representative standard of USN. In addition, in order to graft the IP technology onto the sensor network the standardization is progressed around 6LoWPAN WG, ROLL WG, and CORE BoF of IETF. In addition, in the organization including HART, ISA, ISO, IEC, the independent standardization continues according to the communication layer or service object in which the object is. The new standard for the object of the higher layer service is based on the IEEE802.15.4, the standard on the transmission technology.

As to IEEE802.15.4, the standardization was progressed for the humidity and the temperature sensors having the small packet size, basic monitoring service of the low power for collecting metering data and the remote controller for the toy or electronics. Therefore, the IEEE802.15.4 MAC technology has the limitation that it cannot be satisfied the requested quality of the serious performance degradation by the frequent retransmission generated by the increase in the network traffic and has the timeliness delivery of the sensing information. In addition, the received signal quality degradation by the radio frequency interference of the homology or heterogeneous has the problem that it cannot exhibit the function as personal area network and it has the limitation to the activation of the related market.

Recently, the movement for replacing the wire monitoring apparatus for the production quality control with the cheap wireless foundation network in the field having the weak radio environment as the factory automation gets up in the industrial circles actively. Thus, in order to replace the networking between cable equipment wirelessly, HART managing the field electric installation standards for telecommunications established the Wireless HART standard in 2007. The sensor nodes following this standard applied to the factory automation by the members of the Wireless HART.

In addition, the factory automation standard association, ISA finished the wireless system standard, ISA-100.11a standard task for the industrial automation on September 2009 [13]. Thus, the standardization that it supplements the function of the existing IEEE 802.15.4-2006 MAC standard in IEEE802.15 in 2007 and it tries to secure the time of the reliability of the radio environment and delivery of the sensing information is progressed in TG4e [1]. The operation of PAN at the

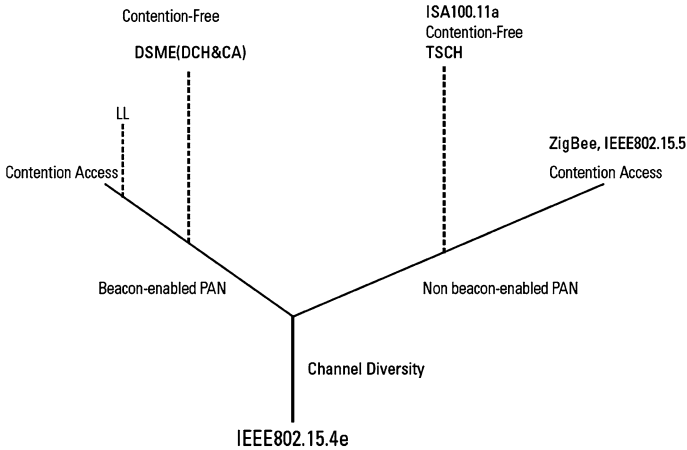


Fig. 1 MAC operation classification in IEEE802.15.4e

IEEE802.15.4e- is classified into the beacon enabled PAN and the non-beacon enable PAN. The beacon enabled PAN is used based the periodically broadcasted Beacon. The non-beacon enabled PAN mode requests the beacon by the non-periodic for the exchange of the communication frame and uses PAN. This has PAN the same action mode in order to maintain the compatibility with the IEEE802.15.4. The MAC function action mode of the IEEE802.15.4e is used with two PAN operations and Fig. 1 schematizes this [15].

The major characteristics of IEEE802.15.4e apply the time-sharing based channel diversity technique. The channel access of time-sharing reduces the retransmission by the packet collision caused from the characteristic of the random access channel like CSMA and minimizes the valid communication power. This MAC technology can improve the quality of the transferred information by guaranteeing the definite delay time for the alarm and the time is required delivery of the monitoring and controlling information. In addition, the channel diversity technique in which RF instrument of the homology or heterogeneous coexists like the industrial site and the radio frequency interference overcomes the receiving signal quality degradation by the severe environment and channel fading caused by the wireless channel can maximize the reliability of the RF link.

2.3 6LoWPAN

6LoWPAN is the architecture and protocol standard for using the Internet Protocol Version 6(IPv6) based IEEE802.15.4-Low Rate Wireless Personal Area Network among the working group of IETF. 6LoWPAN is the technology for delivering relatively slow IEEE802.15.4 packet (250 kbps/2.4 GHz, 40 kbps/915 MHz, 20 kbps/868 MHz) through the large-scale IPv6 packet with end-to-end.

In addition, the content about which will how perform the IPv6 automatic address establishment function using the MAC address (16bit or 64bit extended type address) which IEEE 802.15.4 technology uses is included [16, 17]. Because of using the protocol which 6LoWPAN gateway uses in the existing IP network from the IP layer to the application layer as it is, it can have the structure of the little simple gateway. Therefore, in the gateway, the management procedure can be simple and the processing time can reduce.

3 Environmental Control and Monitoring System in Greenhouse

3.1 The Organization of the Greenhouse Control System

In the Greenhouse, the Greenhouse control system administers the various information for the crop growth monitoring and controlling through the sensor and actuator nodes. It is the main controlling equipment, which can utilize the practical knowledge and creates the appropriate growth environment comprised in this system by the growth environment administration, application service function of the Greenhouse Control Gateway, the Greenhouse operating system, and the Integrated Greenhouse Managing System.

The communication between the Integrated Greenhouse Management System and the Greenhouse operating system, the Greenhouse operating system and the Greenhouse control gateway uses the IP-based communication like the Wi-Fi, CDMA, and Ethernet. The communication between the Greenhouse control gateway and sensor nodes/actuator nodes can use the wire-based ones including the Ethernet, RS485, CAN, etc. and the wireless-based ones like ZigBee, Wi-Fi and Bluetooth according to the environment selectively. The sensor nodes play the roles of delivering the sensing value (temperature, humidity, CO₂, illumination, solar radiation in the Greenhouse, rain, wind velocity outside the Greenhouse, and moisture, EC, PH and temperature of soil) to the Greenhouse Control Gateway. The Actuator nodes play the role that the actuator like window control, illumination (LED), ventilation machine, and hot blast heater is driven according to the message delivered from the Greenhouse Control Gateway and it controls the Greenhouse environment (Fig. 2).

The Greenhouse Control Gateway plays the role that it collects the information from the sensor nodes and delivers the order received from the Greenhouse operating system to the actuator nodes. According to the necessity, the role of converting the protocol is played and the simple logic can be included for the optimization control of crop growth. It is similar to the role of a gateway in USN and it integrates with the Greenhouse operating system according to the function or network architecture and it can manage.

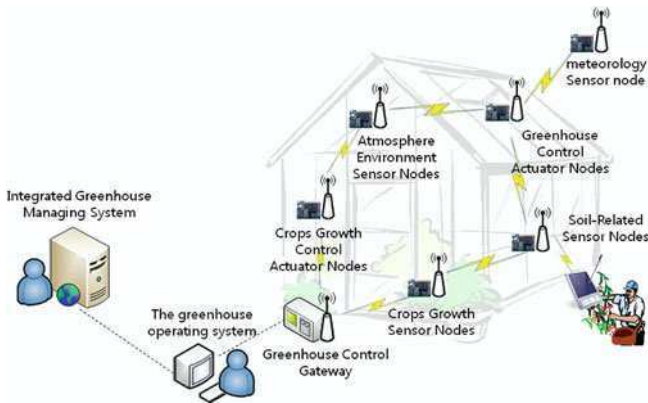


Fig. 2 System component in greenhouse

The Greenhouse operating system can monitor the environment and the crop growth information from the Greenhouse inside and external sensors. In addition, it is possible to control the optimization operation through the Greenhouse environment control algorithm. This system keeps collected data in the database and can record the agriculture diary and vegetation expertise based on stored data. In addition, the service and control software required for the Greenhouse operation downloads from the Integrated Greenhouse Managing System. The duty cycle of the sensor nodes and actuator nodes is managed. The Integrated Greenhouse Managing System, linked with the Greenhouse operating system and data server positioned in the other site, provides the feedback with the crop growth information to the Greenhouse system. In addition, installing the necessary software in the Greenhouse operating system according to the sensor nodes and actuator nodes is played.

3.2 Greenhouse Control Gateway

The Greenhouse Control Gateway controls the sensor and actuator required for the crop growth of the inside of the Greenhouse and does the gateway roles for the coupling with the external system like Greenhouse operating System. The Fig. 3 is the configuration diagram of the Greenhouse Control Gateway.

The Greenhouse Control Gateway is comprised of the mainboard of the ARM9 based the mainboard, the base module for the IEEE802.15.4e communication, the resistive type-touch screen, key-pad, Bluetooth for the I/O unit of users, and the Ethernet and Wi-Fi modules for the internet connections. After making the sensor module and RF communication transmit and receive sensor data, the base node communicates with the mainboard with the serial communication. The function of

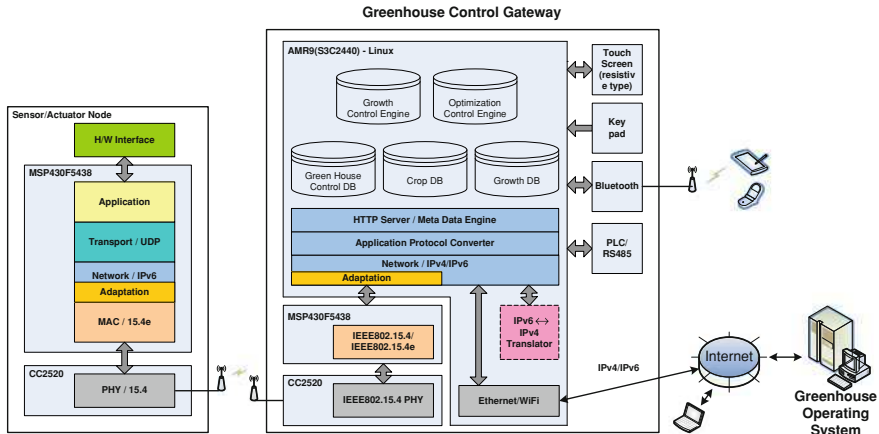


Fig. 3 Greenhouse control gateway

base module minimized to improve the packet processing ability for IEEE802.15.4e.

In the mainboard, 6LoWPAN function of directly controlling the Sensor Node at the Greenhouse operating system is mounted. Three databases and two control engines is mounted in order to control the optimum state of growth. As to the Greenhouse Control DB, the temperature, humidity, illumination, solar radiation sensor of the inside the Greenhouse and the wind direction and velocity, rain, temperature, humidity sensor outside the Greenhouse, etc. collect the information of environment in Greenhouse. The Growth DB collects the moisture, EC, PH and temperature of soil, CO₂, the temperature and area of leaves, the height of crop, etc. for the real-time rearing information gathering of the crops. The Crop DB is the fundamental rearing and insect information DB defined in advance. The Growth control Engine controls the temperature, CO₂, water, etc. controlled according to the growing level by the adaptive control algorithm for controlling the rearing of the crops. By using above three database information, it optimally-control and the inside of Greenhouse is controlled through the optimization control engine. The Optimization control engine controls the Greenhouse through the fuzzy algorithm with the mutual feedback of the actuator including the inside of the Greenhouse and external sensor information and the ventilation machine, LED, sodium lamp, fluorescent lighting, sprinkling of water, CO₂ machine, hot blast heater, and window control, etc. The Fig. 4 is the real picture of the Greenhouse Control Gateway.

3.3 Sensor/Actuator Nodes

After the sensor node implemented in this system collects the sensing value including the temperature, humidity, CO₂, EC, Ph., etc. and process the data through the microprocessor (MSP430) in order to collect the greenhouse inside



Fig. 4 Green control gateway H/W

growth environment information of the crops, by using RF Transceiver (CC2520 RF Chip), it transmits to the Greenhouse Control Gateway. In addition, after the actuator node receives the command from the Greenhouse Control Gateway and process the order, it does the control operation including the illumination, inner temperature, blower, side and roof wall control, etc. The radio frequency signal of the node was transmitted to the Transmit Power of 4 dBms from 2,400 to 2,483.5 MHz frequency bandwidth to 250 kbps.

3.4 The Implementation of IEEE802.15.4e MAC and 6LoWPAN

The existing Greenhouse Control System utilized PLC, RS-232C, and RS485 for the stability of the control but has the defect that the expandability can decrease. For using for automation control, the technologies making use of the existing ZigBee have used for the monitoring in spite of many advantages because the stability was insufficient. In our system, by using the IEEE802.15.4e, that is the wireless technology which was stable and in which the expandability is high, the technology in which the collision avoidance of existing radio frequency was possible and the reliability and real-time transmission are high. The IEEE802.15.4e MAC processing routine implemented based on Distributed Synchronous Multi-channel Extension (DSME) MAC standard of the IEEE802.15.4 e (2011) Draft standard. The function of WPAN can express with the message as the same sequence as the existing IEEE802.15.4 and develop each message sequences with the Primitive that is exchange information between the upper and the bottom layer.

The inter layer interface is as shown in Fig. 5. The between layers, interface is cooked based on FIFO in order to minimize the intervention of SW by MCU.

6LoWPAN implemented in the kernel of embedded Linux. Therefore, the unnecessary processing layer is omitted and the case of being not its own dwelling

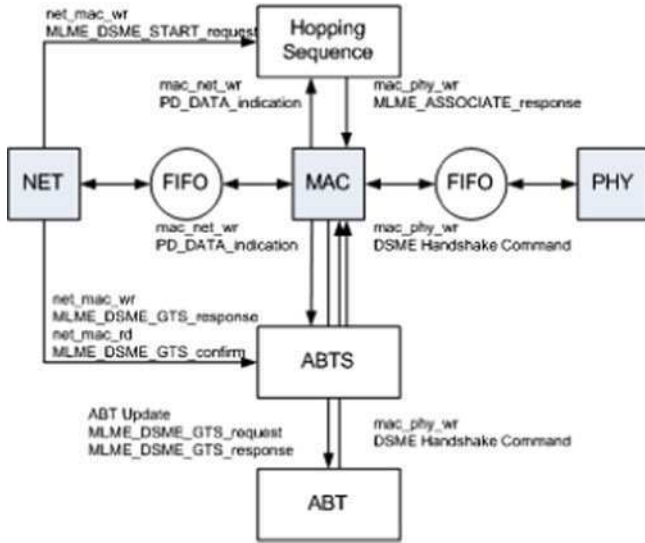


Fig. 5 Structure for IEEE802.15.4e MAC

is thrown in the kernel level through destination with directly, the routing routine of the IP layer. It rises to the higher rank in case of its own data. The complicated process step formed by coming up to the application is skipped and the processing speed is improved.

4 Conclusion

In this paper, we show the Greenhouse Control System to enhance the expandability and reliability of the Greenhouse having many control elements like the ventilation machine, LED, sodium lamp, fluorescent lighting, sprinkling of water, CO₂ machine, hot blast heater, and window control, etc. The new standard IEEE802.15.4e was applied to the Greenhouse Control Gateway and Sensor/Actuator nodes in order to overcome the limit of ZigBee for the monitoring. Also because of using the protocol which 6LoWPAN gateway uses in the existing IP network from the IP layer to the application layer as it is, it can delivery end-to-end transmission, sensor/actuator node to the Greenhouse Operating System for detailed operation.

Acknowledgments This work was supported by the Industrial Strategic technology development program, 10037299, Development of Next Generation Growth Environment System funded by the Ministry of Knowledge Economy (MKE, Korea).

References

1. IEEE P802.15.4e, IEEE Standard for Information technology—Telecommunications and information exchange between systems—Local metropolitan area networks—Specific requirements, Part 15.4: wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (WPANs), (2011), IEEE 802.15 WPANTM task group 4e, <http://www.ieee802.org/15/pub/TG4e.html>
2. IEEE Std 802.15.4-2006, IEEE Standard for Information technology—Telecommunications and information exchange between systems—Local metropolitan area networks—Specific requirements, Part 15.4: wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (WPANs) (2006)
3. Sigrimis N, Antsaklis P, Groumpos P (2001) Special issue on control advances in agriculture and the environment. *IEEE Control Syst Mag* 21:8–85
4. Sigrimis N, King R (1999) Special issue on advances in greenhouse environment control. *Comput Electron Agric* 26:217–374
5. Gonda L, Cugnasca CE (2006) A proposal of greenhouse control using wireless sensor networks. In proceedings of 4th world congress conference on computers in agriculture and natural resources, Orlando
6. Zhu YW, Zhong XX, Shi JF (2006) The design of wireless sensor network system based on ZigBee technology for greenhouses. *J Phys* 48:1195–1199
7. Feng X, Yu-Chu T, Yanjun L, Youxian S (2007) Wireless sensor/actuator network design for mobile control applications. *Sensors* 7:2157–2173
8. Rodríguez F (2002) Modeling and hierarchical control of greenhouse crop production (in Spanish). PhD thesis, University of Almería Spain. <http://aer.ual.es/TesisPaco/TesisCompleta.pdf>
9. Rodríguez F, Guzmán JL, Berenguel M, Arahal MR (2008) Adaptive hierarchical control of greenhouse crop production *Int J Adap Cont Signal Process* 22:180–197
10. Pawlowski A, Guzman JL, Rodríguez F, Berenguel M, Sánchez J, Dormido S (2009) simulation of greenhouse climate monitoring and control with wireless sensor network and event-based control. *Sensors* 9:232–252
11. Janos S, Istvan M (2009) Distance monitoring and control for greenhouse systems via internet Kopaonik. Srbija, Zbornik radova konferencije Yuinfo, pp 1–3
12. Matijevics I, Simon J (2010) Control of the greenhouse's microclimatic condition using wireless sensor network. *IPSI J TIR* 6(2):35–38
13. ISA-100.11a-2009 (2009) Wireless systems for industrial automation: process control and related applications, ISA
14. IEEE 802.15 WPANTM Task Group 4e. <http://www.ieee802.org/15/pub/TG4e.html>
15. Jeong W, Shin C (2010) The trend of MAC standardization at wireless sensor network, *TTA J* 129
16. Kim K, Montenegro G, Daniel Park S, Chakeres I, Yoo S Dynamic MANET On-demand for 6Lo-WPAN (DYMO-low) Routing. Draft-montenegro-6lowpan-dymo- low-routing-00, IETF(2007. 6)
17. Kim E, Kim Y (2007) The trend of standardization of 6LoWPAN based IP-USN. No.22 and electronic communication trend analysis

Accuracy Estimation of Hybrid Mode Localization Method Based on RSSI of Zigbee

HoSeong Cho, ChulYoung Park, DaeHeon Park and JangWoo Park

Abstract The Zigbee support RSSI (Received Signal Strength Indicator) for the localization measurement. But, RSSI occur error by signal attenuation. We study Hybrid method that the accuracy measured localization of Zigbee. The Hybrid methods are using RSSI and AOA method. In this article, we measured angle with distance of RSSI value based on hybrid method. Also, we get standard deviation using experiment data, angle and distance error measured by the simulation. We study on the influence of distance and angle on location error, we conformed our method by comparing our result to DV-hop and DV-distance results

Keywords Localization · RSSI · Position error · Simulation

1 Introduction

The localization technique is one of the core technology for control and handling given mission be influenced by changes in the environment to communicate with computer and person or object [1].

H. Cho · C. Park · D. Park · J. Park (✉)
Department Of Information and Communication Engineering,
Sunchon National University, 413 Jungangno, Suncheon, Jeonnam 540-742, Korea
e-mail: jwpark@sunchon.ac.kr

H. Cho
e-mail: thsgk1215@sunchon.ac.kr

C. Park
e-mail: naksu21@sunchon.ac.kr

D. Park
e-mail: dhpark@sunchon.ac.kr

The kinds of localization technique are TOA (Time of Arrival), TDOA (Time Difference of Arrival), AOA (Angle of Arrival), RSSI (Received Signal Strength Indicator) [2, 3]. TOA is method to calculate by distance to measure the absolute time for signal arrives between receiver and beacons [4]. It needed synchronization between receiver and beacons. TDOA is one of localization method using multiple beacon time difference of received signals. Unlike method of TOA, it needed synchronization between each beacons own. AOA is one of localization method by direction angle of received signal using array antenna [5]. RSSI is method to measure position between receiver and beacon using differences the strength of signal be influenced by distance [6].

The study of localization technique go along actively by various means, it find position to calculate coordinate and then representation by vector matrix using triangulation method and trigonometry [7], beside that besides that, using the triangulation method and VOR (VHF Omnidirectional Ranging) base station [8], and to calculate average of each hop distance of landmark and sensor node using triangulation and AOA method on Ad-hoc network [9], etc.

So, our goal is to reduce position errors in technique of Localization. Position error is affected the distance and angle error. For this reason, we were an experiment to measure the distance and angle using RSSI. We were calculated average and standard deviation of each distance and angle through the experiment. Based on this, we investigated about the impact of cumulative the distance error and angle error to position error through simulation.

This article consists of 5 sections. Section 2 is explains the distance and angle measurement test using RSSI. And Sect. 3 is explains method of angle and position measurement, Sect. 4 is explain results of simulation, Sect. 5 describes the conclusions.

2 Measurement of the Distance and Angle Using RSSI

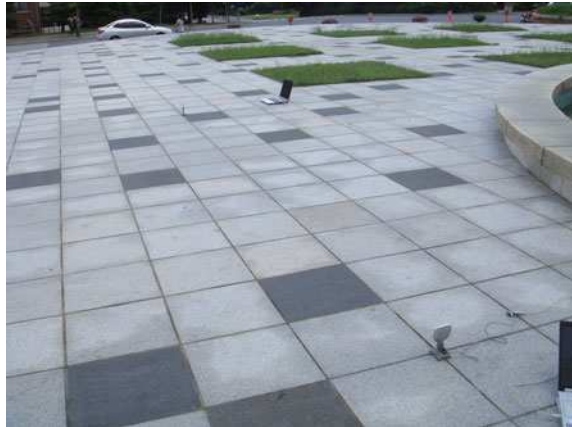
The experiments are used Zigbee module of Maxstream Inc and directional antenna of Com Interline Inc. The frequency range of antenna is 2.4–2.5 GHz, horizontality signal range is $65^\circ \pm 5^\circ$, verticality signal range is $60^\circ \pm 5^\circ$, and signal intensity is 6dBi. The Zigbee module is using Xbee2 pro module and we measured to RSSI using the protocol defined in module.

We chose an experimental environment to the spacious park for less likely to be affected to multipath fading.

Figure 1 shows measurement of distance experiment. The distance of experiment is measured from 2 m until 12 m, and 2 m intervals. Directional antenna is connecting to RP-SMA type with Zigbee. The power of transmitter and receiver is used laptops, and the receiver connected laptop is display to received signal intensity and save to file. The experiment is measured to 100 times for each distance. Figure 2 shows the distribution of the measured data each distance.

The Eq. 1 is theoretical power equation.

Fig. 1 Measurement experiment environment



$$20 \log d = P_{TR} - P_{RV} + C \tag{1}$$

Where $d[m]$ is the distance, $P_{TR}[dBm]$ and $P_{RV}[dBm]$ are the transmitted and received power levels, C is constant.

Figure 3 shows fitting graph to the theoretical power equation and distance data of the experiment.

The Eq. 2 shows similar forms although slight differences compare the Eq. 1.

$$17.21 \log d = P_{TR} - P_{RV} - 22.973 \tag{2}$$

Figure 4 show RSSI value for the 12 m measurement. In the distance of 12 m average was $-65.2[dBm]$, standard deviation was $0.574[dBm]$. We obtained average of standard deviation of all distance such as Fig. 4. It is $0.474[dBm]$. Represents the standard deviation of the value of the distance is 16 cm.

Also, we get to error of angle by experiment. The angle experiment was measured at distance of 5 m and divided into 5° of whole 70° . We found average by measuring each angle to 100 times. This is show in the Fig. 5.

Figure 6 is show RSSI value to measured angle 0° (beacon in the state of the receiver to the front facing). Average was $-49.156[dBm]$, standard deviation was $0.329[dBm]$. We obtained average of standard deviation of all angle such as Fig. 6. It is $0.73[dBm]$. Represents the standard deviation of the value of the angle is 2.2° .

3 Angle Measurement and Localization Method

All wireless sensor nodes assumed to be able to directly communicate neighboring nodes within their transmission range. The landmarks of nodes know their position coordinates and based on the direction. In addition, all sensor nodes can measured directions of incoming signal from neighbor nodes through own axis.

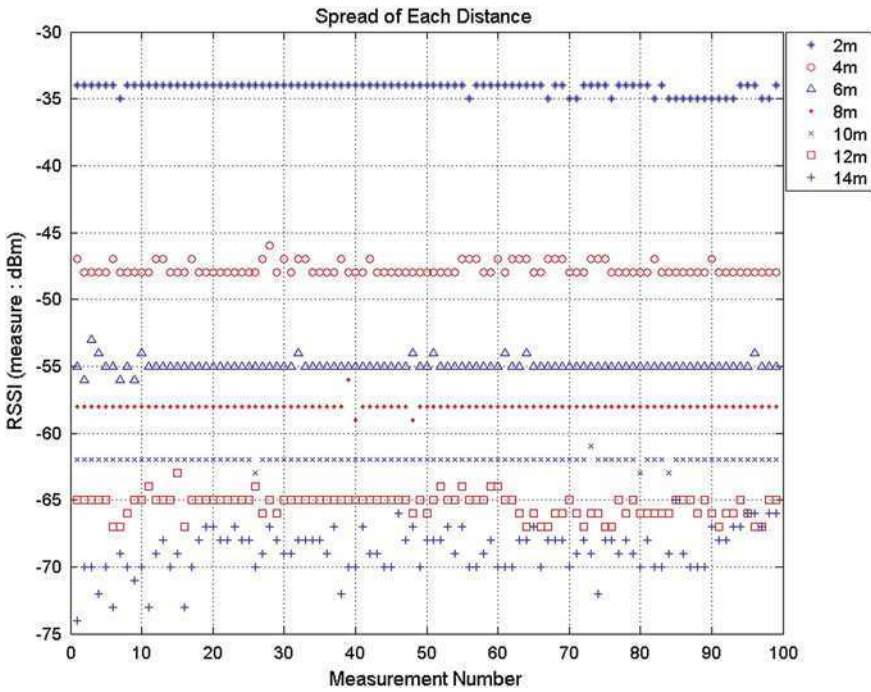


Fig. 2 The RSSI distribution of each distance

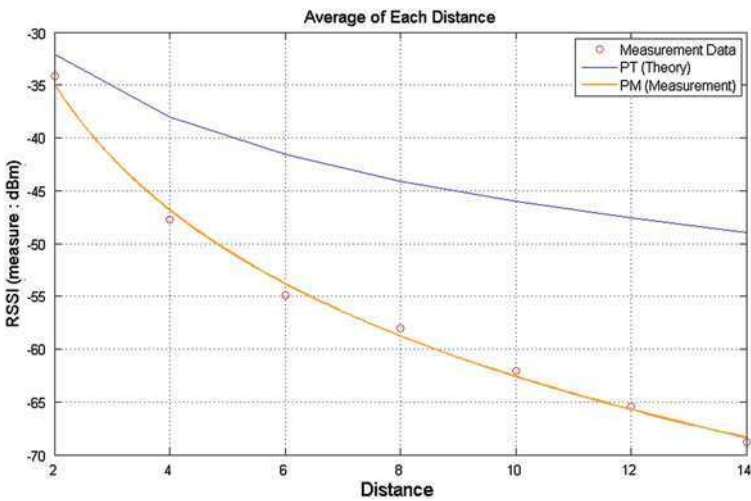


Fig. 3 Each distance power fitting graph using MATLAB (PT: The theoretical power equation, PM: The experiment data power equation)

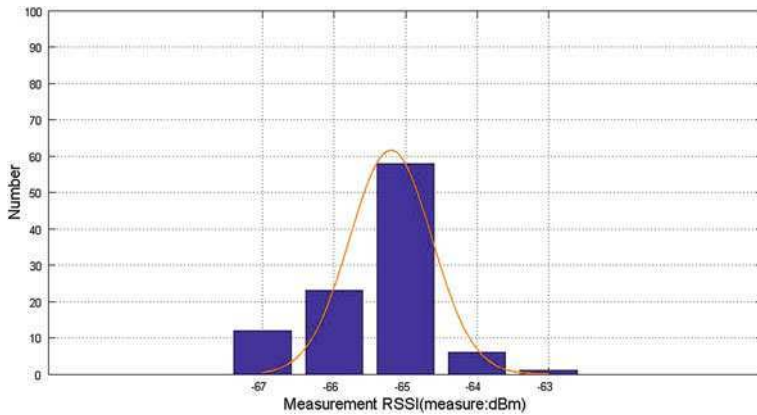


Fig. 4 Standard deviation for the 12 m measurement

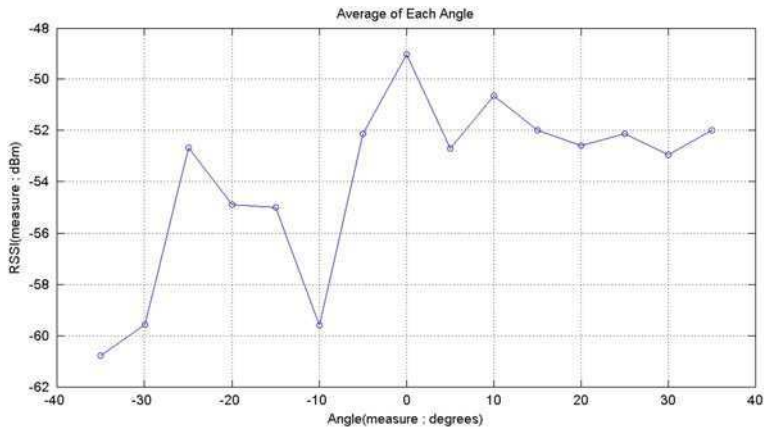


Fig. 5 The RSSI value according to each angle

Figure 7 shows wireless sensor nodes A, B, C how to calculate mutual angles. Blue dotted arrow indicates based on the direction. The azimuth is defined from axis of each node to base on the direction in Fig. 7. The azimuth of node A denote \hat{A} , node A to node C in each direction angle was expressed as a \widehat{AC} . θ show angle from based on the direction to neighbor nodes. In other words, θ_{AC} shows to face angle from based on the direction of node A to node C.

$$\theta_{AC} = \begin{cases} \widehat{AC} - \hat{A} & \text{for } \widehat{AC} \geq \hat{A} \\ 2\pi + (\widehat{AC} - \hat{A}) & \text{for } \widehat{AC} < \hat{A} \end{cases} \quad (3)$$

obtain as Eq. 3,

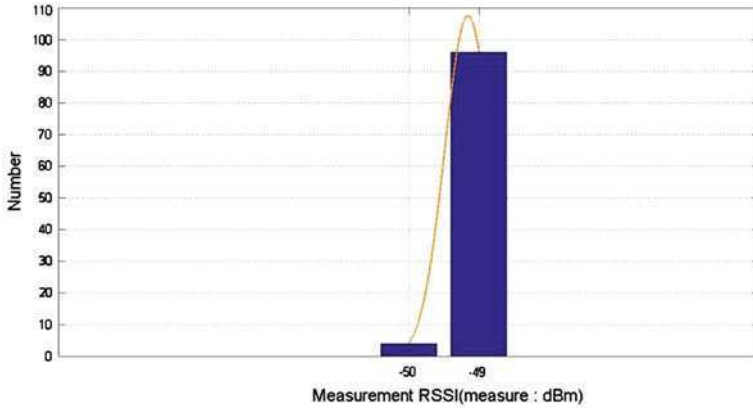
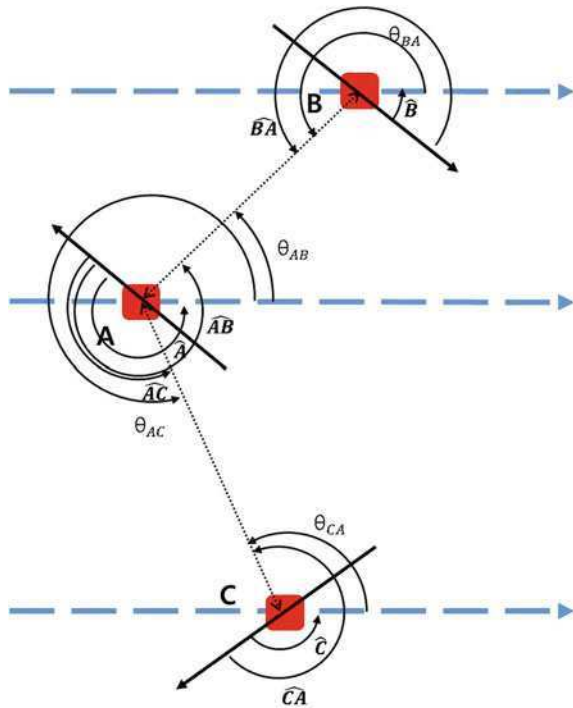


Fig. 6 Standard deviation of angle to 0°

Fig. 7 Getting the angle between the neighboring sensor nodes



In Fig. 7, node A and node C for example, find azimuth of node A,

$$\hat{A} = \begin{cases} \widehat{AC} - \theta_{AC} & \text{for } \widehat{AC} \geq \theta_{AC} \\ 2\pi + (\widehat{AC} - \theta_{AC}) & \text{for } \widehat{AC} < \theta_{AC} \end{cases} \quad (4)$$

be represented by Eq. 4.

All sensor nodes could be found own azimuth when performed repeatedly in this way starting at landmark to all neighbor node. The sensor nodes could be measured own position got at neighboring sensor nodes within the transmission distance and obtain azimuth and incident angle.

In this article, localization method consists of three steps. First step, all sensor node find distance between neighbors nodes to determine within own transmission range. Next step, neighbor nodes of landmark find their azimuth and incidence angle of landmark. The course repeat to know for own standard direction and incidence angle of neighbor nodes to all nodes. Third step, they calculate position of neighbor nodes using measured distance and angle to neighbor nodes, and inform the position of the neighbor node. For example, it can calculate the location of landmarks at least one nodes to own position.

Also, we considered coordinates update method of four kinds. First, this is minimum hop method. This method is to calculate the coordinates of a node from coordinate of the landmark having least hops to a node. The second is minimum distance method. In other words, this method is to determine the position based on the coordinate of nearest landmark. The third way is to find the average of calculating coordinate by all landmarks to know. Finally, the fourth way is to find the average of coordinates by consider to calculate all coordinates for various paths.

We are simulated considering the methods in [Sect. 4](#).

4 Simulation Results

This article will show the simulation results and compare with the results of DV-hop and DV-distance method [9]. The proposed method will be susceptible to angle error because the node's azimuth is calculated from a landmark far apart. This becomes severe when a landmark is far apart from. This article, measured distance is assumed to be Gaussian. And it is assumed the measured angle is also Gaussian.

$$r_{meas} = r_{exact}(1 + \sigma_r N(0, 1)) \quad (5)$$

$$\theta_{meas} = \theta_{exact} + \sigma_\theta N(0, 1) \quad (6)$$

where $r_{meas}(\theta_{meas})$ is the measured distance(angle), $r_{exact}(\theta_{exact})$ is the true value of the distance(angle), $\sigma_r(\sigma_\theta)$ is a specific constant [10], and $N(0, 1)$ is a normally distributed random variable. Therefore, the noise error in measured values is modeled as additive and can be varied by changing the specific constant $\sigma_r(\sigma_\theta)$, where in MATLAB program, $\sigma_r(\sigma_\theta)$ in Eqs. 5 and 6 roles standard deviation of normal distribution, so we will simply call it a standard σ deviation.

$$d = \frac{N(\pi R^2)}{A} \quad (7)$$

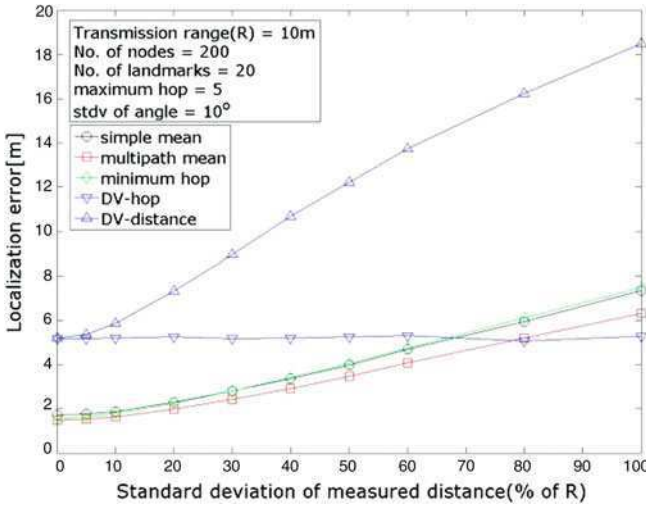


Fig. 8 Position error according to sensor node measured distance error

where N is the total number of sensor nodes deployed in sensor field, A is the area of sensor field, R means the transmission range of nodes which is the same among all nodes. Localization error obtained in the simulation is defined by

$$L_{error} = \frac{\sum_{i=1}^N |r_{calc} - r_{real}|}{N} \tag{8}$$

where N is the total number of nodes, r_{calc} is the calculated coordinate of a node, and r_{real} is the real coordinate of a node.

In this simulation, sensor field is considered to be square with length 100 m. The transmission range is 20 m. As shown in Eq. 7, node density can be changed according to varying the number of nodes or transmission range when the area of sensor field is fixed.

In Figs. 8 and 9, parameters ‘simple mean’ mean sensor nodes know their coordinates averaged values, ‘Multipath mean’ mean reaching landmark to average value of all paths for nodes to determine their position. ‘minimum hop’ mean the minimum number of hops for decide own coordinates from landmarks. The simulation was run from wider range to distance of experiment. Also, the simulation result compared DV-distance and DV-hop.

Figure 8 shows the localization error according to the measured distance errors. In this case, distance error is shown percent of transmission range. Our method and DV-distance shows the increased localization error. But our method result in good localization error until standard deviation of distance reaches 75% of transmission range. Also, DV-hop shows constant localization error.

As explained in Sect. 3, our method is sensitive to the error in angle measurement. Especially, the accumulation and propagation of angle error is seemed to be profound. Figure 9 shows the localization errors resulted from out method.

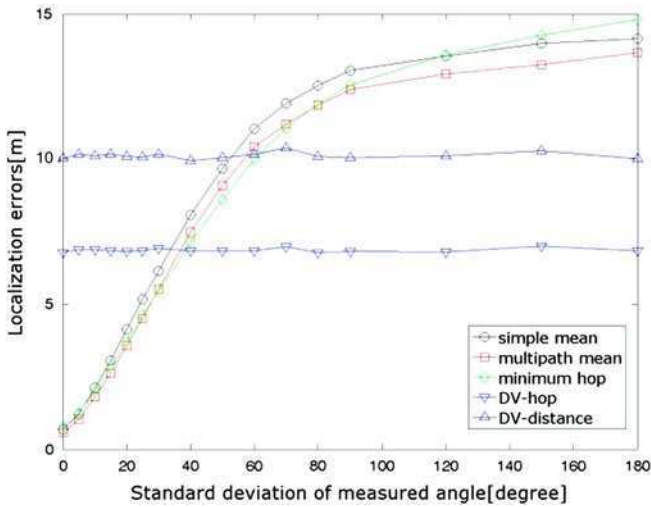


Fig. 9 Position error according to sensor node measured angle error

According to increasing the standard deviation of measured angle error, localization error is increased. But our method shows better results than DV-methods until reaching about 40° of standard deviation. When the angle error increases sufficiently, the localization error reaches saturation. It means measured angles are fully random and cannot contribute the accuracy of localization any more. DV-methods do not consider the angle measurement so that the result shows constant localization error regardless of angle error.

5 Conclusion

We started from the measurement using distance and angle a position of neighboring nodes of the landmarks. Then we experiment RSSI value according to distance and angles through distance and angles measuring experiment using RSSI, and shows angular measuring method. If position error angular error when standard deviation of less than approximately 40° , we are known relatively good performance compared to DV-distance and DV-hop. We are known the good performance compared to DV-distance and DV-hop if angle error have the standard deviation of less 40° . DV-hop algorithm is many no change to position error according to distance, however we could be found many the error to happen to DV-distance. In contrast, like better existing algorithm than we proposed localization method if less than 70% of distance error. We were known to important of angle measurement by proposed method from results of simulations.

Acknowledgments This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MEST) (No. 2011-0005294)

References

1. Cho HS, Park C-Y, Park D-H, Park J-W (2011) AoA localization system based on Zigbee experimentation and realization. *J Korea Navig Inst* 15(1):83–90
2. Niculescu D, Nath B (2003) Ad hoc positioning system (APS) using aoa. 22nd annual joint conference of the IEEE computer and communications societies, vol 3, pp 1734–1743
3. Wang X, Bischoff O, Laur R, Paul S (2009) Localization in wireless Ad-hoc sensor networks using multilateration with rssi for logistic applications. *Inst Electromagn Theory Microelectron (ITEM)* 1:461–464
4. Cheung KW, So HC, Ma W-K, Chan YT (2004) Least squares algorithms for time-of-arrival-based mobile location. *IEEE Trans Signal Process* 5(4):1121–1128
5. Peng R (2006) Angle of arrival localization for wireless for sensor networks. *Sensor and ad hoc communications and networks. SECON'06.2006 3rd annual IEEE communications society*, vol 1
6. Sugano M (2006) Indoor localization system using rssi measurement of wireless sensor network based on zigbee standard. In: *Wireless sensor network*
7. Betke M, Gurvits L (1997) Mobile robot localization using landmark. *Robotics Autom IEEE Trans* 13:251–263
8. Niculescu D, Nath B (2004) VOR base stations for indoor 802.11 positioning. 10th annual international conference on mobile computing and networking, pp 58–69
9. Niculescu D, Nath B (2001) Ad hoc positioning system (APS). *Glob Telecommun Conf* 5:2926–2931
10. Biswas P, Aghajan H, Ye Y (2005) Integration of angle of arrival information for multimodal sensor network localization using semidefinite programming. In: *Proceedings of 39th Asilomar conference on signals, systems and computers*, pp 1–9

A Study on the Failure-Diagnostic Context-Awareness Middleware for Wireless Sensor Networks

In-Gon Park and Chang-Sun Shin

Abstract We propose a middleware that diagnose any incorrect operation of a sensor or equipment occurring in the WSN application system adopting an indoor sensor network, and determines as to whether the incorrect operation is related to the reliability of the service offered by the system. The middleware proposed in this thesis makes up Data Management Module, Circumstance Information Related Module, Circumstance Analysis Module, Service Module, and Information Storage Module. The data retrieved from the interoperation between modules are analyzed through incorrect operation diagnosis algorithm, and thus it is determined whether a sensor or equipment operates incorrectly.

Keywords Failure-diagnostic · WSN · Context-awareness

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 (2011-0014742).

I.-G. Park · C.-S. Shin (✉)
Department of Information and Communication Engineering,
Sunchon National University, Suncheon, Korea
e-mail: csshin@sunchon.ac.kr

I.-G. Park
e-mail: pig9004@sunchon.ac.kr

1 Introduction

Ubiquitous computing converges IT technologies with other industries like construction, education, healthcare, agriculture and so on. Especially the Wireless Sensor Network (WSN), a spatially distributed monitoring network consisting of wireless devices using sensors, is considered as a state-of-the-art technology in ubiquitous computing. Rapid advances in wireless networks, embedded systems, and sensor technologies have introduced various WSN-related industries that are becoming more important in everyday life [1]. IP-based WSN network enables each sensor node to make IP communications, and IP-USN technology being able to directly interconnect with BcN draws attention from people as a core technology realizing an advanced Ubiquitous society [2]. Such changes in WSN paradigm have been developing based on the establishment of sensor infrastructure and its service expansion. Against this backdrop, studies on energy efficiency, pervasive, and the reliability [3–5] are in progress, and researchers in various areas are studying the WSN middleware supporting customized features and operation characteristics [6–9].

However, researches on how to deal with error is insufficient compared with that on R&D of fundamental technology of sensor and sensor network. Accordingly, the main purpose of this paper is to research sensor failure diagnostic algorithm for improving reliability in order to deal with problems and to monitor it automatically when the sensor errors [10].

In order to solve such problems, this paper suggests sensor the Failure-Diagnostic Context-awareness Middleware (FDCM), sets virtual environment of plants factory, and conducts verification tests of sensor failure diagnostic algorithm for supporting highly reliable services of WSN application systems through analysis of context-awareness through sensing information and collected data based on WSN.

2 Related Works

In general, such failure of sensor nodes can be divided into hard failure such as exhaustion of Micro Processor and electricity, and soft failure such as calibration error of sensors and random noise error [11, 12].

As for studies on hard failure of sensor nodes, there was one study which aimed to apply BIST (Built-In-Self-Test) technique widely used for improving productivity of components such as DRAM in order to detect electricity shortage and hardware failure. BIST technique can be applied to detect failures of multiprocessor systems mutually connected to each other guaranteeing hardware redundancy, and accordingly, sensor nodes installed in widespread areas are able to be applied effectively as they also have hardware redundancy. Also, as for an applicable technique for detecting failures in case that there is no hardware

redundancy, consensus algorithm has been suggested [13, 14]. Consensus algorithm is an algorithm which detects fixed nodes through nodes checked to have no failures on sensor network, and detects failures through constant exchange and renewal of information with nodes around it by using Suspect matrix and Fault vector created by information of nodes around it.

As for various techniques for detecting failures of sensor soft, there is J. Chen's LFSD (Localized Faulty Sensor Detection) algorithm which can detect errors of sensor data transported from nodes around it based on difference between sensor values of around nodes attained periodically and its own measured value, and change of the value according to time [15].

3 Architecture of Failure-Diagnostic Context-Awareness Middleware (FDCM) System

This paper is that checking of Sensing Data whether there is error of sensors and devices through real-time analysis of data received from lots of environmental sensors, and composes middleware which can control the system promptly and effectively according to the result. Also, it suggests sensor failure diagnostic algorithm based on sensor network which can minimize the cost of equipment and cost of maintenance by checking data transported from relevant sensors through analysis without applying additional error detection module. In order to improve reliability of the algorithm in virtual plant factory environment, middleware is supposed to be composed as follows (Fig. 1).

3.1 Data Management Module

Data management module plays a role of processing data collected through sensors and storing them in database, and delivering them to the context data provider module which examines whether there are errors of sensors or devices. As it is difficult to figure out the situation of normal operation only with data collected through sensors, it restructures packets by adding needed information for judgment to collected sensing data. Data management module is composed of data parser class, connection class, packet data class, and packet creator class. Data parser class extracts data which will be the basis for checking whether there is real failure or not by processing source data collected from many heterogeneous sensors. It also support the function for removing sync byte and header part from collected sensing data, and the function for converting primarily processed data into data which can be utilized practically. Connection class delivers sensing data collected in sync nodes to context data provider module through serial communication. It sets basic information required for communication, and performs functions such

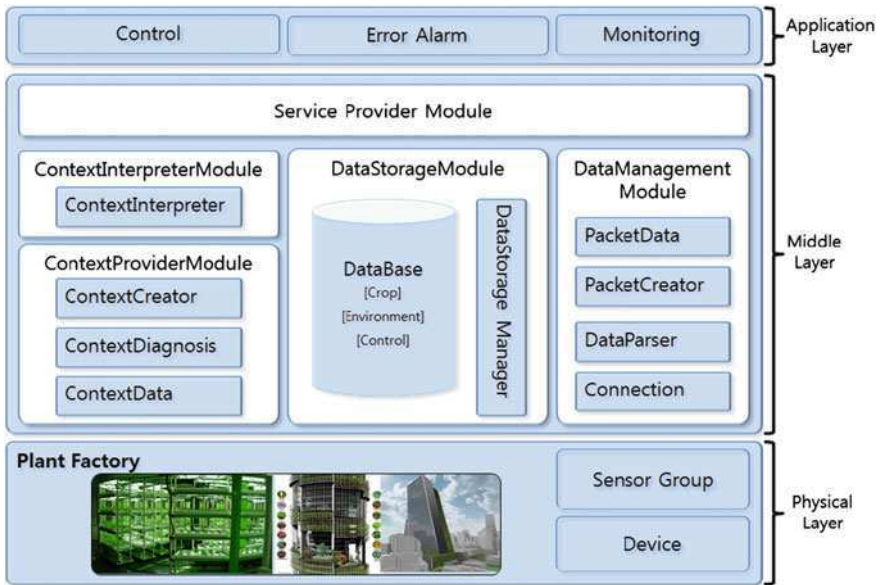


Fig. 1 FDCM architecture

as connection and disconnection to serial communication, and stop of data transmission. Packet data class is a class which defines restructured packets by adding required information for context data provider module to analyze data. Table 1 shows the structure of packets composed in this class. Packet creator class, which supports creation of packet information, performs a function for inputting data of variables managed by packet data class, and for helping data return packets.

3.2 Context Provider Module

Context provider module performs a role of converting restructured data packets transported from data management module into context data, and is composed of context diagnosis class, context data class and context creator class. Context diagnosis class confirms error of sensors and devices by analyzing restructured packets in data management module, distinguishes a situation where it is required to run devices, and restructures this into context data. Context data class is a class defining context data in order to analyze services required by context interpreter module, and includes data for deciding context data. Context data gets data from the context creator class. This class performs a function for setting context data in order to check errors of sensors and devices, and a function for setting ID of

Table 1 Composition of packet data

Variable	Explanation	Note
id	ID of the relevant sensor	
block_id	ID of a block installed with the relevant sensor	
sector_id	ID of a sector including the relevant sensor	
data_CODE	Types of data collected by the sensor	
currentData	Data collected by the current sensor	
beforeData	Sensing data just before collecting the current data	
blockAve	Average value of current data of sensors in the same block	Current sensors excluded
beforeblockAve	Average value of before data of sensors in the same block	
sectorAve	Average value of block data of sensors in the same block	Current sensors excluded

devices related to optimal standard data managed by relevant sectors and performance of application services.

3.3 Context Interpreter Module

Context interpreter module decides which application service will be provided by analyzing context data transported from context provider module. Context interpreter class which composes this module is a class asking for support by analyzing context data transported from context provider module, performs a function for checking whether current sensors or devices work well by analyzing context data, and calls a failure notification service to service provider module when sensors or devices error.

3.4 Service Provider Module

In the service provider module, all services provided by this middleware are implemented. This module provides relevant services when the context interpreter module asks for support based on the result of context analysis. Service provider class which is implemented with failure notification service decides which error message will be performed by analyzing requested data, and provides data of failing sensors and devices to users.

3.5 Data Storage Module

Data storage module stores standard data needed for the application system and environment data collected by sensors. Also, it defines data for use by extracting data from data repository, and defines a method related to data processing. `DataStorageManager` class manages general things such as insert, modification and deletion of data in database, and transports data required by other modules. Database is composed of total 5 tables; `ENVIRONMENT`, `DATA_TYPE`, `DEVICE`, `CONTROL`, and `OPTIMAL`. `ENVIRONMENT` table stores data collected by sensors, and it was set to be able to expand sensors flexibly even when sensors having new data are added by having N vs. 1 relation with `DATA_TYPE` table which stores a list of data types collected by sensors. Also, as for devices, it is possible to expand devices even when new devices are installed inside of the system by designing control table and `DEVICE` table having N vs. 1 relation, too. `OPTIMAL` table stores optimal environment standard data required by relevant systems, and context data provider can refer to the relevant table when controlling or diagnosing.

4 FDCM's Services

Failure diagnosis algorithm is a process which enables to manage facilities effectively and to provide reliable data by diagnosing whether sensors and devices installed in the system operate normally. As for operating process, when the data management module gets data from sensors, it stores relevant data in database through the data storage module. After that process, it makes restructured packets by asking for data required to restructure sensing data to the data storage module. Restructured packets diagnose whether there are any failure during analysis process in the context provider module. If the relevant sensor is checked to failure, it makes context data by asking for data to the data storage module. Relevant context data is transported to the context interpreter module, and it requests which service will be provided through context interpretation. The service provider module provides error data about sensors or devices to users.

The important part of this is a process of composing context data by analyzing packets in the context provider module. Situations which can occur during operation of WSN application system are divided into two categories. Details about failure diagnostic to each situation are as follows.

4.1 When the Device Does Not Work

In case that the device does not work, when difference between `blockAve` and `sectorAve` is not higher than `gap`, and difference between `beforeData` and `currentData` is not higher than `gap`, it can be inferred that a situation shown in

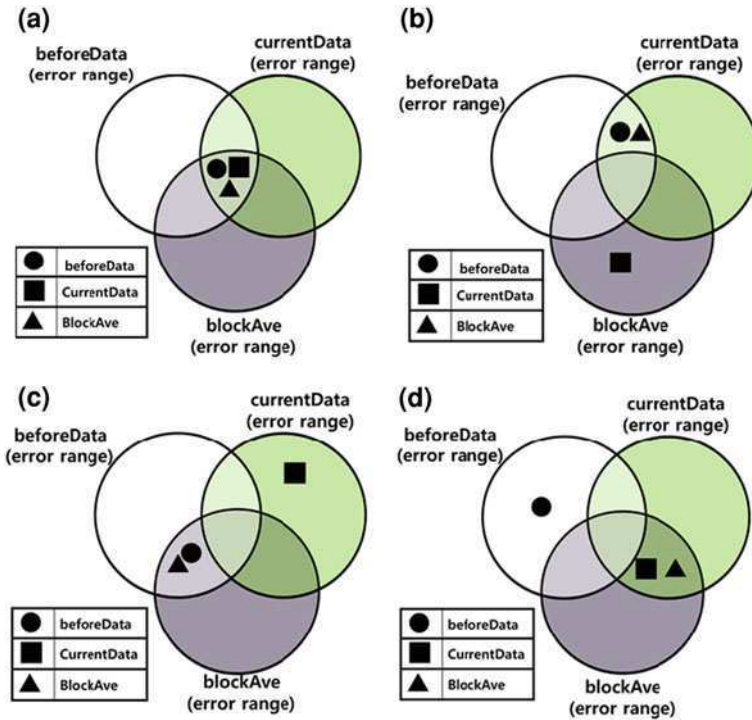


Fig. 2 Can be estimate of situation when stop of control device

(a) and (b) of Fig. 2 through comparison of currentData and blockAve. As a situation shown in (a) of Fig. 2 means that the sensor keeps an average value constantly, it check the result as normal one, and decreases by 1 when sensor-ErrCount is 0 or more. A situation shown in (b) of Fig. 2 is checked to be outside of the average value constantly. As it is possible for the sensor to error, it increases sensorErrCount by 1.

When difference between beforeData and currentData is higher than gap, and difference between blockAve and beforeData is lower than gap, it is a situation where currentData is outside of the average value while operating normally, as shown in (c) of Fig. 2. In case of that, as it can be inferred that failure occurs or external factors influence on the operation, it increases sensorErrCount by 1.

When difference between beforeData and currentData is higher than gap, and difference between blockAve and beforeData is higher than gap, a situation where sensing data has been outside the average value and then returns to the average again, as shown in (b) of Fig. 2. In this case, it decreases sensorErrCount by 1.

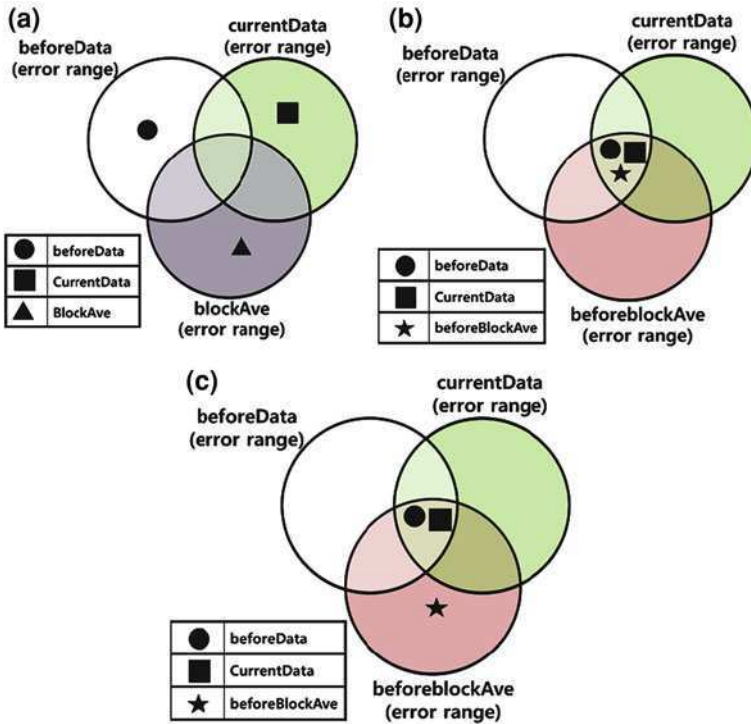


Fig. 3 Can be estimate of situation when run of control device

4.2 When the Device is on Control

In case that the device is in operation, when difference between beforeData and currentData is higher than gap, and difference between currentData and blockAve is higher than gap, a situation shown in (c) of Fig. 3 and (a) of Fig. 3 can be expected. In (b) of Fig. 3, as the device is in operation though currentData is outside the blockAve, value of blockAve and currentData will change continually. This means that it is not clear whether the sensor is failing or it is an instant change generated during control, and it examines later change of data by increasing sensorErrCount by 0.5.

During the device is in operation, when difference between beforeData and currentData is not higher than gap, and difference between currentData and beforeBlockAve is higher, it means that the sensing value does not change but surrounding environment is constantly changing. So, the sensor may error, increasing sensorErrCount by 0.35. On the contrary, when difference between currentData and beforeBlockAve is not higher than gap, it means that the device does not operate well. In this case, it increases deviceErrCount by 1 (Fig. 4).

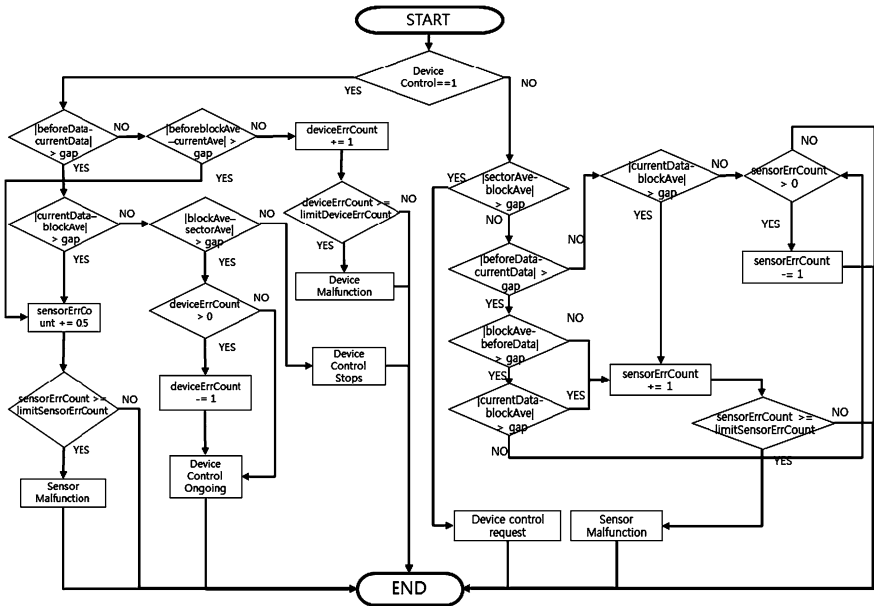


Fig. 4 Flowchart of the FDCM

Such diagnosis and measures are repeated whenever it gets data from sensors installed in a plant factory, and when sensorErrCount or deviceErrCount of relevant sensors exceeds 5, the critical value, each will be checked as a sensor error or device error.

5 Conclusions

This paper suggested failure diagnostic algorithm which can actively diagnose error of sensors or all kinds of devices or environmental problems frequently occurring installation sites in order to operate automated systems applying WSN technology reliably and effectively. In order to check the core function, whether sensors or devices error or not, situations where errors can be generated were defined by comparing currentData, beforeData, blockAve, and beforeblockAve through failure diagnostic algorithm suggested in this paper.

Future studies will supplement this algorithm by verifying and improving standard critical values considering many environmental data and the properties of sensors and devices as well as currently verified temperature through many tests, make it deal with all kinds of situation which can be generated in WSN application systems flexibly by adding context data which this middleware can diagnose, and examine a minute control algorithm for effective operation of devices.

References

1. Yang S, Park S, Lee EJ, Hong Ryu J, Kim B-S, Kim HS (2008) Dual addressing scheme in IPv6 over IEEE 802.15.4 wireless sensor networks. *ETRI J*, 30:627–633
2. Saegun Oh (2007) New IP-USN trend and prospect development. In: National IT Industry Promotion Agency, National IT Industry Promotion Agency Studies Information Technology Weekly Trend No. 1300
3. Liu Y, Liu K, Li M (2010) Passive diagnosis for wireless sensor networks. *IEEE/ACM Trans Netw* 18:1132–1144
4. Akan OB, Isik MT, Baykal B (2009) Wireless passive sensor networks. *IEEE Commun Mag* 47(8):92–99
5. Vasar C, Prostean O, Filip I, Robu R, Popescu D (2009) Markov models for wireless sensor network reliability. In: ICCP 2009. IEEE 5th international conference
6. Kim S (2007) Sensor network R&D and using cases. In: Technology Weekly Trend 2007
7. Kim M, Lee Y, Park C (2007) USN middleware technology development trend. In: Electronics and telecommunications research institute, electronics and telecommunications trend analysis, vol. 22
8. Kim M, Lee E (2005) Sensor database technology in ubiquitous environment. In: National IT industry promotion agency studies information technology weekly trend
9. Hadim S, Mohamed N (2006) Middleware challenges and approaches for wireless sensor networks. In: *IEEE Distributed Systems Online*, vol.7
10. Min H, Lee S, An S (2007) Design and network implementation of sensor node for wireless sensor networks. In: Korea Computer Congress 2007, vol. 34
11. Chen J (2007) Distributed fault detection of wireless sensor networks. Iowa State University Ames, Iowa
12. Yook U, Yun S, Kim S (2007) Development of fault detection algorithm applicable to sensor network system. *Korea Fuzzy Log Intell Syst Soc* 17(6):760–765
13. Ranganathan S, George AD, Todd RW, Chidester MC (2000) Gossip-style failure detection and distributed consensus for scalable heterogeneous clusters. In: HCS Research Laboratory
14. Young M (1989) The technical writer's handbook. Mill Valley, Seoul
15. Lee M, Mun J, Jeong M (1998) A design of multipurpose robust controller for robust tracking control of optical disk drive. *J Control Autom Syst Eng* 4(5):592–599

Livestock Searching System on Mobile Devices Using 2D-Barcode

ChulYoung Park, HoSeong Cho, DaeHeon Park, ChangSun Shin,
Yong Yun Cho and JangWoo Park

Abstract In this article, we have designed and implemented livestock searching system on mobile devices using 2D-Barcode. 2D-Barcodes capacities have risen a hundredfold for 1D-Barcode. In recent years, it can share data, (text, phone number, hyperlink, etc.) by means of application on feature phone and smart devices. Also, QR-Code is suitable for identification of missing livestock, because of QR-Code is easy to print with low-cost and it was designed to withstand external damage. And it can store identity of livestock and provide the location information to livestock-farmer using smart devices. Also, our system generates QR-Code that fit vCard v3.0 format with livestock-farmer. Therefore, the first people to found a missing livestock that can contacts the farmer immediately. In order to do this, we developed mobile application and server program.

C. Park · H. Cho · D. Park · C. Shin · Y. Y. Cho · J. Park (✉)
Department of Information and Communication Engineering,
Suncheon National University, 413 Jungangno, Suncheon,
Jeonnam 540-742, Korea
e-mail: jwpark@sunchon.ac.kr

C. Park
e-mail: naksu21@sunchon.ac.kr

H. Cho
e-mail: thsgk1215@sunchon.ac.kr

D. Park
e-mail: dhpark@sunchon.ac.kr

C. Shin
e-mail: csshin@sunchon.ac.kr

Y. Y. Cho
e-mail: yycho@sunchon.ac.kr

In addition to this, we developed QR-Code generation module and we built database server from the farmer's identity and livestock's identity.

Keywords 2D-Barcode · Mobile application · QR code system

1 Introduction

In recent years, increased the use of mobile devices and offered the more information to general public. Also, 1D-Barcode and 2D-Barcode technique have caught the attention on the mobile environment. 2D-Barcode can share data, (text, phone number, hyperlink, etc.) by means of application on feature phone and smart devices. The increase of mobile device users, thereby increase of the application using 2D-Barcode. This application service is providing through mobile devices.

Barcodes divide into two types. First, 1D-Barcode type is using on logistics, distribution and etc. 1D-Barcode express of information depending on thickness of bar lines that the only vertical. Second, 2D-Barcode type represents of information by way of braille or mosaic. 2D-Barcodes information capacity have risen a hundredfold for 1D-Barcode. And it was designed to withstand external damage. It can be recognition in the direction of 360 degrees is the advantages of 2D-Barcode. And 2D-Barcode is easy to use without database because of high-capacity [1].

2D-Barcode is a service for providing about information of object. In recent years, it can share data by means of application on feature phone and smart devices. QR-Code in either of the 2D-Barcode technique can be easy that read through smart devices application. It can store identity of livestock and farmer's contact. Also, QR-Code is suitable for identification of missing livestock, because of QR-Code is easy to print with low-cost and it was designed to withstand external damage.

Also, we has designed and implemented application service on mobile devices using 2D-Barcode for missing livestock. It can store identity of livestock and provide the location information to livestock-farmer using smart devices. In this article is composed as follows. The [Sect. 2](#) describes related works. The [Sect. 3](#) describes system design and implementation. And the [Sect. 4](#) makes a conclusion.

2 Related Works

2.1 Data Matrix Code

Data matrix is one of the most widely used two dimensional barcodes. It can be broken up into ECC 00-140 and ECC 200 according to the error checking and correction algorithm. The data matrix is a high density code that can be encoding

Fig. 1 The symbol of data matrix code



Fig. 2 The symbol of PDF417



with 3116 numeric chars and 2345 alphanumeric chars in the ASCII character set [2].

Data matrix code is characterized as follow. First, symbol size can use 0.001 to 14 inch. Also, it can be represent alphanumeric chars of up to 2334 characters per symbol in 1.4 inch square. Second, symbol is always a square or rectangular shape. And the finder pattern is surrounded by outline. Third, it can be recognition in the direction of 360 degrees is the advantages of 2D-Barcode using CCD scanner or camera (Fig. 1).

2.2 PDF417

Portable data file (PDF417) is portable data file of high density 2D-Barcode in 1991. PDF417 have error correction capability of the eight step. Half of the data was lost on account of noise; even so, it can be recognize with maximum level. And PDF417 composed to unit of code word (Fig. 2, Table 1).

PDF417 includes start-code and stop-code at both ends. Code word of row indicator includes lane number, each of rows, each of columns and error correction ratio, etc. [3].

2.3 QR Code

QR code is the acronym for quick response code. QR code has the large capacity about 7,089 chars of numeric, 2,396 chars of alphanumeric, 2,953 bytes. It was created by Denso-Wave in 1994. QR Code can include to text, vCard and URL,

Table 1 The difference each mode of PDF417

Mode	Example	Font size and style
Byte	6	1,108 bytes
Text (Alphanumeric/ASCII)	2	1,850 chars
Numeric	3	2,725 chars

Table 2 Specification of QR Code

Code size	21 cell × 21 cell	
Data type and value	Numeric 8 bit or byte binary UTF-8 characters	7,089 2,953 1,817
Error correction functionality (LEVEL)	L M Q H	7% of code word 15% of code word 25% of code word 30% of code word

Fig. 3 The symbol of QR code



etc. QR Code has error correction algorithm. It has 30% restoration ratio by one code word. Also, it can be recognize in the direction of 360 degrees [4] (Table 2).

QR Codes store a lot of information with number of dots. However, it is necessary wide area. Also, QR Code has to include three finder pattern and several alignment patterns (Fig. 3).

QR Code has finder patterns, alignment patterns, timing patterns, and a quiet zone (Fig. 4).

QR Code v.3 use finder pattern for the detection of QR Code position. It can be recognized in the direction of 360 degrees. Timing pattern is used to determine the coordinates of symbol on decoder application. Alignment pattern should be used to for correcting the distortion. Quiet zone is margin for reading the QR Code [5].

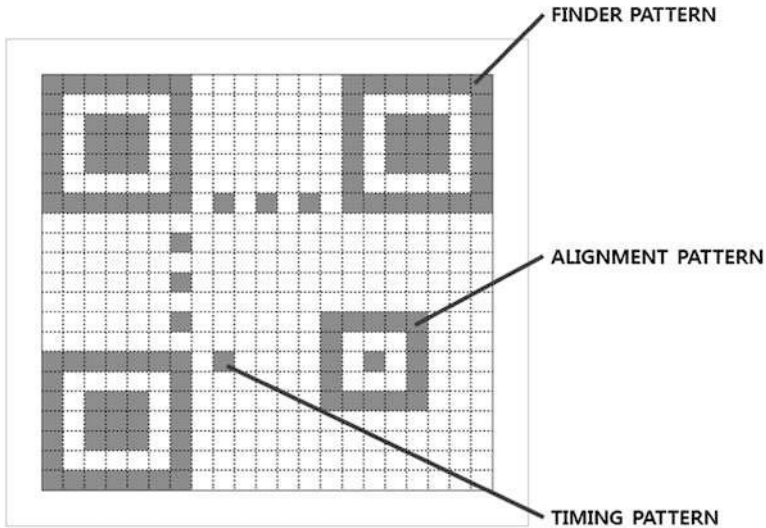


Fig. 4 The QR code structure

Our system was using ZXing library on Objectives-C and JAVA when read the QR Code. ZXing barcode library is open-source. It supports UPC-A, UPC-E, EAN-8, EAN-13, Code 39, Code128, QR Code, Data Matrix, PDF417 and ITF.

3 System Design and Implementation

3.1 System Composition

Figure 5 shows the Livestock searching system composition. Our system is broken into two parts: server part and client part. The server part was composed of web server, remote server, mobile application server and database server. The client part was composed of remote server, client PC, wireless access point, QR code printer and mobile devices.

The web server was included page for generate to QR Code. And the remote server manages for each client group. The mobile application server was designed for iPhone or Android platform.

Figure 6 presents the flow for recognize of QR Code. These procedures get the information from QR Code on mobile device and smart devices application. The information decoded from QR Code is text, vCard and URL, etc.

Our system stored for identity of livestock and farmer’s contact using vCard 3.0 formats. vCard is a business card format. The fields of vCard 3.0 format as follows (Fig. 7).

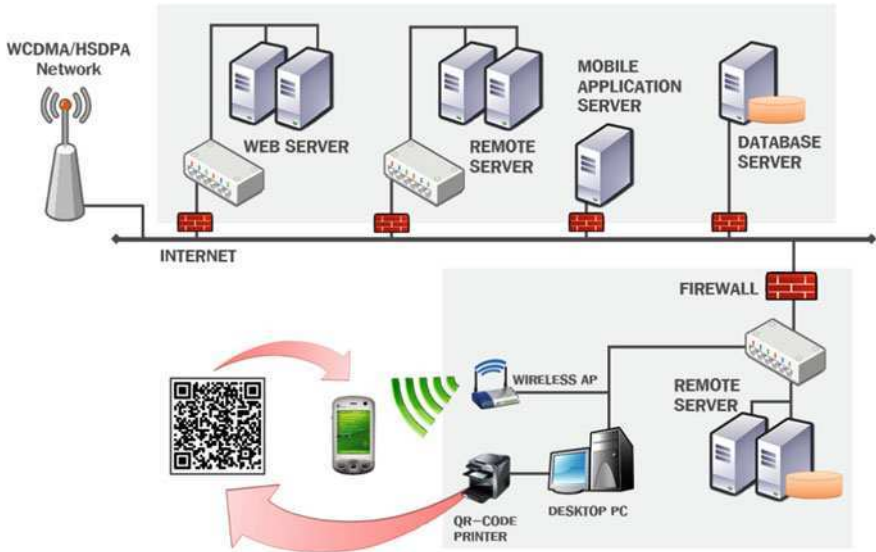


Fig. 5 Livestock searching system composition

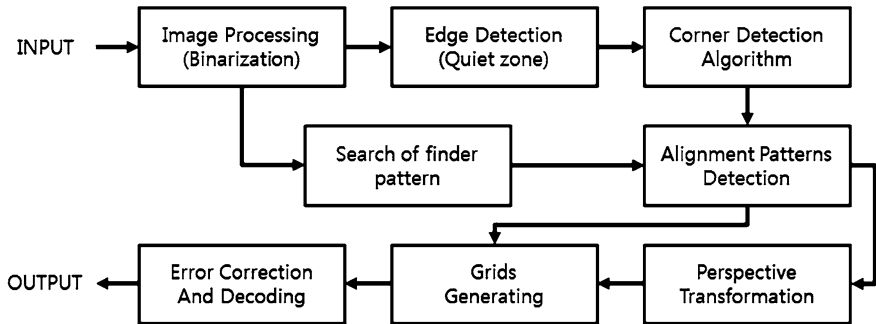


Fig. 6 The flowchart of recognize for QR Code

Fig. 7 The fields of vCard 3.0 format

```

BEGIN:VCARD
VERSION:3.0
N:
FN:
ORG:
TITLE:
TEL;TYPE=WORK,VOICE:
TEL;TYPE=HOME,VOICE:
ADR;TYPE=WORK::
LABEL;TYPE=WORK:
ADR;TYPE=HOME::
LABEL;TYPE=HOME:
EMAIL;TYPE=PREF,INTERNET:
REV:20080424T195243Z
END:VCARD

```

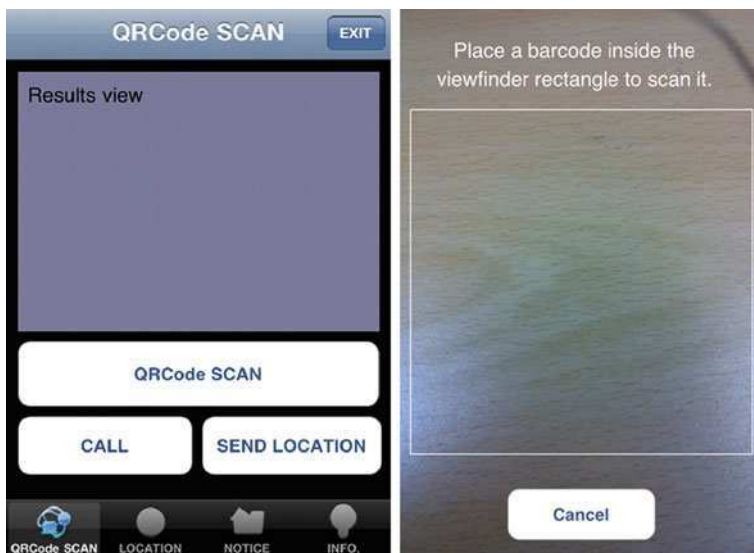


Fig. 8 The application for recognition of QR Code

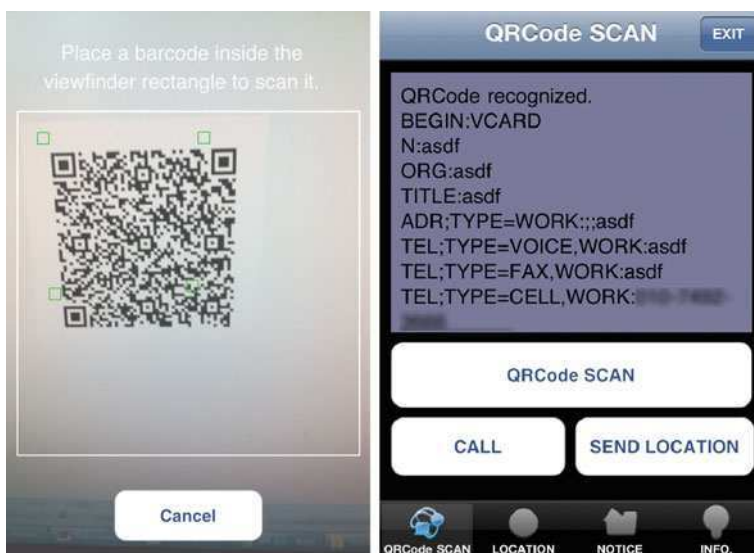


Fig. 9 The results of recognition



Fig. 10 The map information transmitted screen

3.2 Result of System Implementation

Figure 8 shows the application for recognition of QR Code. This application stored to farmer’s contact of livestock.

vCard parse module find “TEL;TYPE=VOICE,WORK:” field and “TEL;TYPE=CELL,WORK:” field in results.

Figure 9 shows the results of recognition. “Send location” button pressed in this application; send the message with map information using GPS module mounted on smart devices. And our system was used to the Google static maps API (Fig. 10).

4 Conclusion

In this article, we designed for Livestock searching system and implementation. QR Code has large capacity and high-density on the paper of small size. For this reason, it is suitable for the livestock out to pasture. Also, our system was used to QR Code and GPS module of smart devices for missing livestock less out of range. The first people to found a missing livestock that can contacts the farmer immediately. Also, Livestock searching system can be implementation with low-cost.

It is for this reason that QR Code can print on the paper.

Acknowledgments “This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC (Information Technology Research Center) support program supervised by the NIPA (National IT Industry Promotion Agency)” (NIPA-2011-(C1090-1121-0009)) and This work was supported by the Industrial Strategic technology development program (10037290, Development of Smart Growth Management System) funded by the Ministry of Knowledge Economy (MKE, Korea).

References

1. Gao JZ, Prakash L, Jagatesan R (2007) Understanding 2D-BarCode technology and applications in M-commerce—Design and implementation of A 2D barcode processing solution. In: COMPSAC 2007. 31st Annual international, vol 2, pp 49–53, 24–27 July 2007
2. Lisa S, Piersantelli G (2008) Use of 2D barcode to access multimedia content and the web from a mobile handset. In: GLOBECOM 2008. IEEE, pp 1–3, 30 Nov 2008
3. Yong Y, Shunying Z, Wang H (2009) An integrated system of freeway toll and traffic data investigation: PDF417 two-dimensional bar code system. In: ICMTMA '09, vol 3, pp. 466–470, 11–12 Apr 2009
4. Chang Y-H, Chu C-H, Chen M-S (2007) A general scheme for extracting QR code from a non-uniform background in camera phones and applications. In: ISM 2007, pp 123–130, 10–12 Dec 2007
5. Yue L, Ju Y, Mingjun L (2008) Recognition of QR code with mobile phones. In: CCDC 2008, pp 203–206, 2–4 July 2008
6. Parikh D, Jancke G (2008) Localization and segmentation of a 2D high capacity color barcode. In: WACV 2008, pp 1–6, 7–9 Jan 2008
7. Kato H, Tan KT, Chai D (2008) Development of a novel finder pattern for effective color 2D-barcode detection. In: ISPA'08, pp 1006–1013, 10–12 Dec 2008
8. Tan KT, Chai D (2010) JPEG compression of monochrome 2D-barcode images using DCT coefficient distributions ICIP pp 2169–2172, 26–29 Sep 2010
9. Huaqiao H, Wenhuan X, Qiang H (2010) A 2D barcode extraction method based on texture direction analysis. In: ICIG'09, pp 759–762, 20–23 Sep 2010
10. Dong H, Jianfu T, Zhaoxuan Y, Yanwei P, Meng W (2010) 2D barcode image binarization based on wavelet analysis and Otsu's method. In: ICCASM 2010, vol 5, pp 30–33, 22–24 Oct 2010
11. Kato H (2010) Performance of a color 2D barcode as a pervasive computing tool. In: ISPACS 2010, pp 1–4, 6–8 Dec 2010
12. Chang Y-H, Chu C-H, Chen M-S (2007) A general scheme for extracting QR code from a non-uniform background in camera phones and applications. In: ISM 2007, pp 123–130, 10–12 Dec 2007

Towards a Context Modeling for a Greenhouse Based on USN

Daeheon Park, Kyoungyong Cho, Jangwoo Park and Yongyun Cho

Abstract Recently more than ever, because of further global warming, violent climate changes, environmental pollution and food problem researchers have taken a lot of interests in greenhouses and vertical farms in agricultural environments. Generally, works in greenhouses are based on situation information from various sensors based on USN. In this paper, we propose a context modeling method for a lot of situation information which can arise in agricultural environments. The suggested context modeling method can define a various data generated in greenhouses based on USN as structural contexts based on a rule-based RDF. Through the suggested modeling method, various sensed data in a greenhouse can be regenerated as high-level context sets. Because the high-level contexts can be usefully used as important service execution conditions, the suggested context modeling method can be widely applied to various smart agricultural service applications or platforms for greenhouses based on USN and IT technologies and raise efficiency in development of a context-aware system and a context-based service automation system.

Keywords Agriculture · Greenhouse · USN · Context modeling

D. Park · K. Cho · J. Park · Y. Cho (✉)
Information and Communication Engineering, Suncheon National University,
413 Jungangno, Suncheon, Jeonnam 540-742, Korea
e-mail: yycho@sunchon.ac.kr

D. Park
e-mail: dhpark@sunchon.ac.kr

K. Cho
e-mail: jkl@sunchon.ac.kr

J. Park
e-mail: jwpark@sunchon.ac.kr

1 Introduction

In ubiquitous computing environments, contexts are one of very important data attributes as service execution conditions [1]. Like in the many fields of ubiquitous computing, there are many of IT technologies such as RFID/USN, computing devices, and databases in greenhouses or vertical farms. So, to develop smart services in greenhouses or vertical farms, developers can use contexts from sensors. Until now, a little of researches introduce a context modeling method which can be applied in greenhouses or vertical farms. Therefore, we need a context modeling method to define low-level situation data related with the various sensors and the agricultural environments in greenhouses or vertical farms as unified contexts.

This paper introduces a context modeling method can redefine a lot of situation information, which can arise in greenhouses or vertical farms with RFID/USN and various IT technologies, as structural high-level contexts. To do that, the suggested context modeling method defines various situation conditions in greenhouses with rule-based RDF contexts, and categorizes the various data as two types of contexts, which are situation contexts and profile contexts. The former is for data sensed from such real sensors in a greenhouse as a humidity sensor, soil temperature sensor, leaf temperature sensors, and so on. The latter is for profile conditions predefined in such database or data files as service schedule information or harvest schedule information needed for automatic agricultural services in a greenhouse. Through the two-categorized context model, the low-level data can be efficiently divided according to the features of sensors or profiles. With the rule-based context composition, the situation contexts and the profiled contexts can be easily redefined as a high-level context. Therefore, with the suggested context modeling method, various data related with a greenhouse can easily be redefined with RDF-based contexts, and can be widely used in development of context-aware or smart applications for greenhouses or vertical farm.

2 Related Works

Until now, there have been many researches for context models or context modeling technologies in various fields [2–6]. The current researches for context models may be focused on how to redefine real data as ontology based on RDF/OWL [7–9]. Recently, [2] introduces a formal context model based on ontology languages for the Semantic Web. The introduced model consists of four ontologies, which are independently categorized as like users, devices, environment and services. Through the systematically divided ontology structure design supports flexibility of context model in changes of specific service domains and ontologies.

As another recent research for context model, there is [2]. The model is designed as service execution conditions for automation of agricultural works in

agricultural environments. The model is based on a RDF/OWL-based ontology as conditions to choice service execution in context-aware workflow service model, and is used a rule-based reasoner for reasoning of high-level contexts from various low-level contexts which can be occurred in an agricultural service domain.

Context models in agricultural environments protected from outside such as a greenhouse and a vertical farm have to consider various situation conditions from crops to user's conditions, growing schedule or harvest schedule. However, because the existing approaches do not include any module for profiled information or predefined data in their context models, they are not enough to adopt in greenhouse or vertical farm environments equipped with various sensors and networked each others with IT technologies and RFID/USN. So, we need a context model to describe not only sensed data but also profiled or predefined situation information as contexts.

3 The Suggested Context Modeling for Greenhouses Based on USN

3.1 A Context Layer in Greenhouses

Commonly, situation data in greenhouses bring out from sensors, devices and databases, and are redefined as contexts of values and types about profile, location, time, and so on. Because, the situation data is low-level, it can not be directly used in context-aware applications or platforms as meaningful data. Therefore, it needs to be transformed as meaningful high-level data, which is a context. Further, situation data in greenhouses includes not only the sensed data but also profiled data from predefined in databases or file systems. So, contexts for greenhouse environments have to define both situation data and profiled data.

Figure 1 illustrates a brief conceptual context layer in greenhouses based on USN.

In Fig. 1, the suggested context model consists of two context layers, which are the situation context layer and the profile context layer. First, the situation context is for low-level situation data transmitted from the sensors layer and the device layer. The sensors layer on USN can periodically monitor various situation changes in greenhouses, and the RDF-based entity rule in Fig. 1 translates the situation data with low-level individual entities.

In this time, the situation data can be soil humidity/temperature values or leaf humidity/temperature values, according to the types and function of the sensors. And, the RDF-based entity rule redefines the various real sensed data as entities with the data type and value. The devices layer in Fig. 1 is for context of situation data from physical networks based on USN, computing devices, and various devices in greenhouses such as cultivating machines or collecting machines. The entities can be ontology-based constraints according to the rule-based context

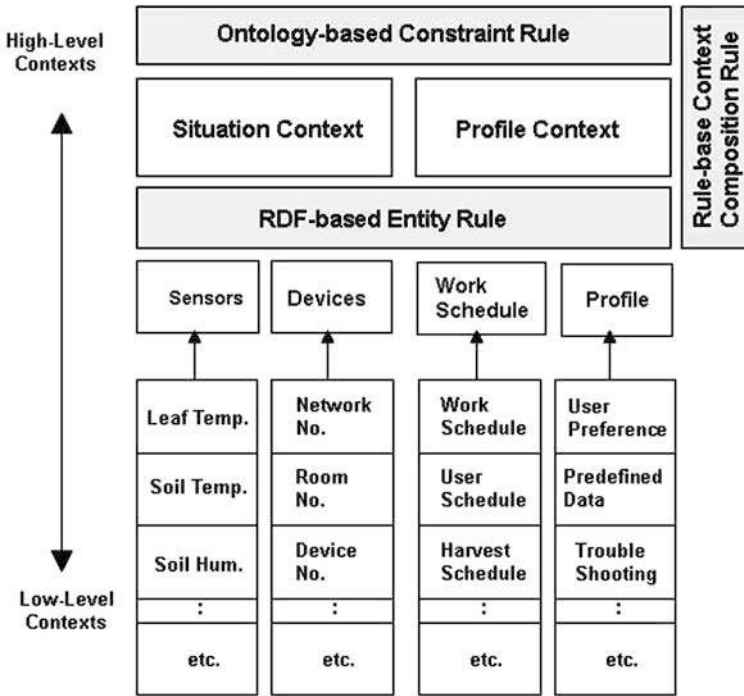


Fig. 1 A conceptual context layer in greenhouses based on USN

composition rule. The individual entities of situation contexts or profile contexts can be an ontology-based constraints through rule-based composition process, using ontologies for greenhouses domains.

3.2 Contexts for a Greenhouse Domain

Figure 2 shows a part of a possible sample context model for the sensors and devices of the situation context layer in a greenhouse domain according to the context modeling method described in Fig. 1.

A context instance for a greenhouse domain can be variously generated according to a designed context model and an ontology. Therefore, it is very important to design a systematic context model automatically to transform sensed data and profiled data as contexts.

As shown in Fig. 2, the low-level sensed data layer shows a data section, in which RDF-based entities of real data sensed from sensors are. And, the high-level context model layer describes a context model section, through which ontology-based contexts consisting of the entities are generated. For example,

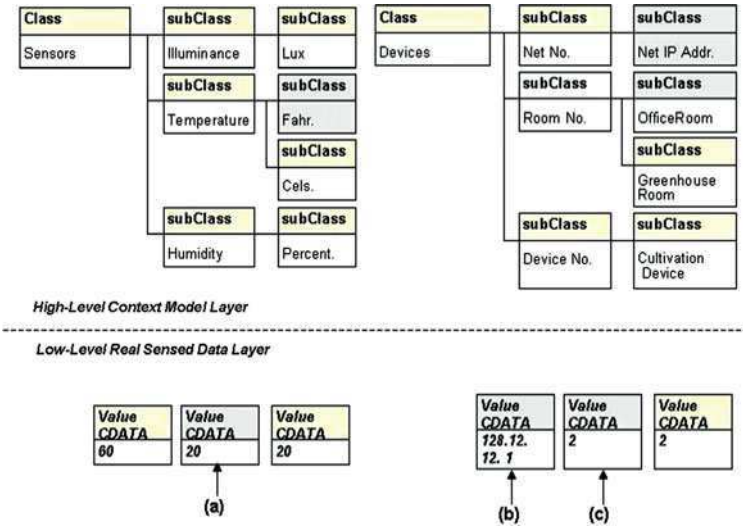


Fig. 2 A part of a context model for a greenhouse domain

there are three sensed values with a gray-colored box, which are (a), (b), and (c) in Fig. 2. The (a) is a entity means that the current temperature value transmitted from a temperature sensor in a greenhouse is Fahr. 20. The entity is a data set, which includes type and value, which are Fahr. and 20 individually. Then, the entity (a) can be composed as a higher-level context through the ontology-based context model for the subclass Fahr., the subclass of the class sensors. The (b) and (c), which are entities, can be composed as a higher-level context through the subclasses of the class devices in Fig. 2. The values may be specific constant values, mainly numbers, to identify devices, not sensed data.

4 Conclusion

In this paper, we introduced a context modeling method for such equipped agricultural environments as greenhouse or vertical farm based on USN and IT technologies. The suggested context modeling method uses a rule-based context composition rule, an RDF-based entity rule, and an ontology-based constraint rule. The RDF-based entity rule is for low-level real data and the ontology-based constraint rule is for composition a constraint with entities. And, the rule-based context composition rule is for composing of higher-level contexts, using constraints and entities. The suggested modeling method categorizes possible situation data in greenhouse domain into specific contexts, which are situation contexts and profile contexts. And, this paper showed possibility of the suggested context

modeling method for generating high-level contexts by designing a sample context model for sample sensed data in a greenhouse domain. Therefore, through the suggested modeling method, various sensed data in a greenhouse can be regenerated as high-level contexts. And, it can be very helpful to develop agricultural service automation, smart service applications or platforms for greenhouses based on USN and IT technologies and raise development efficiency of a context-aware application related with greenhouse environments.

Acknowledgments This work was supported by the industrial Strategic technology development program(10037290, Development of Smart Growth Management System) funded by the Ministry of Knowledge Economy(MKE, Korea) and This work (Grants No. R00045044) was supported by Business for Cooperative R&D between Industry, Academy, and Research Institute funded Korea Small and Medium Business Administration in 2011.

References

1. Dey AK (2001) Understanding and using context. *J Pers Ubiquitous Comput* 5(1):4–7
2. Hervas R, Bravo J, Fontecha J (2010) A context model based on ontological languages: a proposal for information visualization. *J Univers Comput Sci* 16(12):1539–1555
3. Cho Y, Park S, Lee J, Moon J (2011) An OWL-based context model for U-agricultural environments. *LNCS* 6785:452–461
4. Hsien-Chou L, Chien-Chih T (2008) A RDF and OWL-based temporal context reasoning model for smart home. *Inf Technol J* 6(8):1130–1138
5. Dejene E, Marian S, Lionel B (2007) An ontology-based approach to context modeling and reasoning in pervasive computing. *PerComW'07*
6. Strang T, Linnho-Popien, C (2004) A context modeling survey. *UbiComp 2004*
7. W3C (2004) RDF/XML Syntax Specification. W3C Recommendation
8. McGuinness DL, Harmelen FV (eds) (2004) OWL web ontology language overview. W3C Recommendation
9. Matthew H, Protege OWL Tutorial, <http://owl.cs.manchester.ac.uk/tutorials/protegeowltutorial/>

Ad-Hoc Localization Method Using Ranging and Bearing

Jang-Woo Park and Dae-Heon Park

Abstract In Ad-hoc sensor networks, it is very essential for sensors to know their own positions exactly which provide the context to sensed data. Sensors in Ad-hoc sensor networks enable to locate their positions from a relatively small number of landmarks that know their coordinates through external means (e.g., GPS). In this paper, we assume that sensor nodes can measure the distances and relative bearings to neighboring nodes within their transmission range. The proposed method will utilize the distances and relative bearings to find the locations of nodes. Firstly, sensors nearest landmarks will locate their position and then in order nodes more hops far from landmark will. We utilize many landmark coordinates and multiple paths to a landmark to improve the accuracy of the position. Simulation results under the various conditions have been obtained and especially compared with the results using DV-hop and DV-distances.

Keywords Localization · AOA · Ad-hoc sensor network

J.-W. Park (✉) · D.-H. Park
Department of Information and Communication Engineering,
Suncheon National University, 413 Jungangno, Suncheon,
Jeonnam 540-742, Korea
e-mail: jwpark@sunchon.ac.kr

D.-H. Park
e-mail: dhpark@sunchon.ac.kr

1 Introduction

Knowledge of positions of sensors can provide the context to the information which has been collected by sensors in wireless sensor network (WSN). And then, many attractive applications such as routing, tracking assets and et al., can be available through knowing the locations of sensors in WSN.

Positioning algorithms are classified in either centralized or distributed. Centralized method is that the calculation is performed by a server whereas in distributed algorithms all the nodes are able to calculate their own position. There have been lots of researches on the localization. DV-hop and DV-distance proposed by Niculescu [1] have been well known. This paper also is owed by their papers.

WSN means the set of sensor nodes which are deployed ad-hoc. Every sensor node in WSN have ability to communicate other nodes within their own transmission ranges. Sensor nodes are able to measure mutual distances to their adjacent nodes with time of arrival (TOA), time difference of arrival (TDOA) or received signal strength (RSS) and relative angles using angle of arrival (AOA).The localization problem usually means estimating positions of the nodes in WSN based on a mixture of mutual distance, angle or proximity. Existing methods exploit a variety techniques including iterative triangulation [1–4], multidimensional scaling [5], convex programming [6].

In this paper, WSN consists of sensor nodes and landmarks which are exactly same the general sensor nodes except for having their coordinates. The coordinates of landmarks can be given by global positioning system (GPS) or directly by man. Also, it is assumed that the sensors have ability to measure the distance and relative angles to their neighbors. And then sensors except landmarks don't know their absolute reference bearing such as north so that they should infer the reference bearing from the bearings of landmarks. First, positions of nodes within one hop of landmarks will be calculated and then nodes near the nodes which know their positions can calculate their positions so on. There is possibility for nodes to calculate their positions utilizing coordinates of many landmarks. That is, because some nodes having connection to many landmarks can have many coordinates so that those nodes can utilize the many landmarks to improve the accuracy of their coordinates.

Section 2 describes the method for nodes to obtain relative angles to their neighbors and their own azimuths. Determining nodes' coordinates based on the measured angles and distance will be explained in Sect. 3. And then Sect. 4 shows the simulation results and Sect. 5 summarizes conclusion.

2 Measuring the Angle to Neighbor Nodes

All nodes in WSN will be assumed to be possible to communicate their neighboring nodes within their transmission range. Especially landmarks have already known their coordinates and had their own reference bearings (for example, East). It will be assumed that all nodes have ability to measure the distance to their neighbors and relative angles based on their own axis (called heading). Figure 1 shows definitions of angles used in this paper. The node's heading which is used for measuring the angle to neighbors is different among nodes. Node's headings are shown by thick arrows. In Fig. 1, \widehat{ab} is the measured angle at node A to node C. θ shows the incident angle from neighbors with reference of East. For example, θ_{AB} is the incident angle from node A measured at node B. Then, \widehat{a} is the azimuth of node A which means the angle of A's heading measured from East.

Figure 2 show the details for relation of the incident angle to azimuth. Because the landmarks have their azimuths, the calculations will be started from the nodes near landmarks. First of all, assume that node B has been aware of its own azimuth and the angle to node A. Node B is able to calculate the incident angle, θ_{BA} using its own azimuth and the angle to node A

$$\theta_{BA} = \begin{cases} \widehat{ba} - \widehat{b} & \text{for } \widehat{ba} \geq \widehat{b} \\ 2\pi + (\widehat{ba} - \widehat{b}) & \text{for } \widehat{ba} < \widehat{b} \end{cases} \quad (1)$$

Figure 2a is for $\widehat{ba} \geq \widehat{b}$ and Fig. 2b is for $\widehat{ba} < \widehat{b}$. The obtained incident angle θ_{BA} will transferred to node A and then will be used for calculating the angles related to node A. That is, node A will calculate the incident angle θ_{AB} from node B as follows,

$$\theta_{AB} = 2\pi - \theta_{BA} \quad (2)$$

Also from θ_{AB} from Eq. 2 and the measured angle to node B at node A, \widehat{ab} the azimuth of node A can be obtained,

$$\widehat{a} = \begin{cases} \widehat{ab} - \theta_{AB} & \text{for } \widehat{ab} \geq \theta_{AB} \\ 2\pi + (\widehat{ab} - \theta_{AB}) & \text{for } \widehat{ab} < \theta_{AB} \end{cases} \quad (3)$$

The same calculation will be performed all nodes from the neighbors of landmarks so that all nodes with any connection to landmarks can have their own azimuths. And then nodes will also calculate all incident angles from their neighbors from their azimuths and the measured angles to their neighbors. Although nodes with at least one connection path to landmarks is possible to calculate the information related to angles, for nodes far from many hops from landmarks angle error accumulation will be profound. So, we will restrict the hop counts within 5.

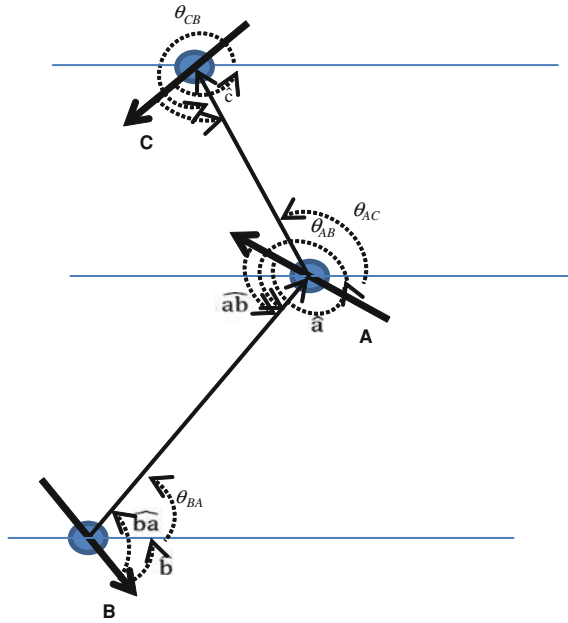


Fig. 1 The typical networked sensors in WSN with the definition of useful angles

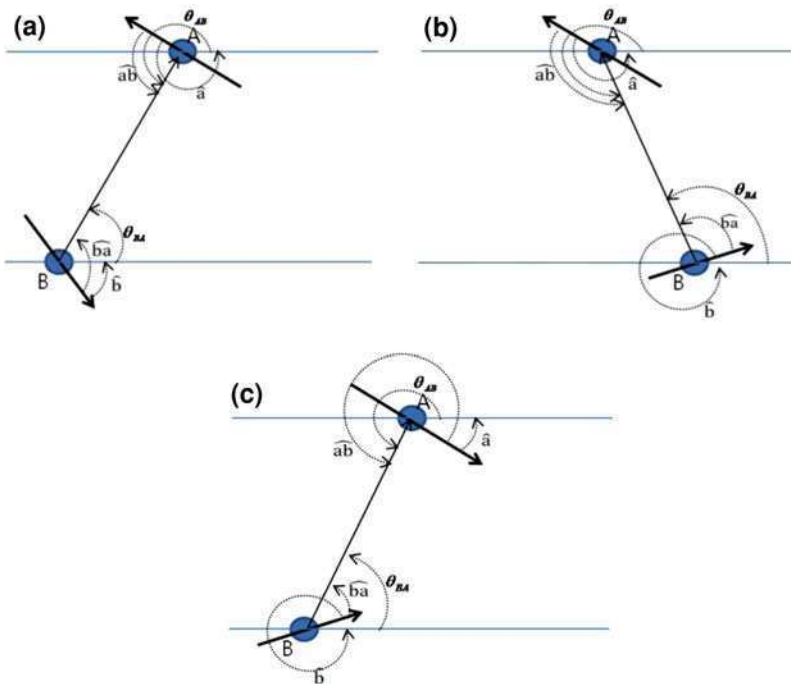


Fig. 2 Details for calculating useful angles

3 Localization

The proposed algorithm consists of three steps. In the first step, all sensor nodes find their adjacent nodes within transmission range and then simultaneously measure the distances and angles to their neighbors. The next step begins at the landmarks. Adjacent nodes to landmarks calculate their azimuths and the incident angles from their nearest landmarks using the method described in previous section. The procedure will continue until all nodes with connection path to landmarks within allowed hop counts (in this case 3 or 5) are aware of their angle information. At the third step, the distances to neighbors and calculated angle of nodes allow the nodes to calculate their position. The calculated location of nodes will broadcast to their neighbors. Then, the node location will be calculated from the coordinates of its neighbors which have already known their coordinates. The calculated position of nodes will be transferred to their neighbors. This process will be repeated. So, all nodes can have their positions as long as there are any connections to them to landmarks.

For example, when the coordinate of node A, (x_A, y_A) is known, the coordinate of node B is simply calculated as follows,

$$\begin{pmatrix} x_B \\ y_B \end{pmatrix} = \begin{pmatrix} x_A + r_{AB} \cos(\theta_{BA}) \\ y_A + r_{AB} \sin(\theta_{BA}) \end{pmatrix} \quad (4)$$

where r_{AB} is the measured distance between node A and B. As known, knowing at least one landmark's coordinate, the node can calculate its own position. If it knows coordinates of more than one landmark, the node can utilize the many landmarks to improve its position. Figure 3 shows the connected paths of one arbitrary node to landmarks. In Fig. 3 green squares are landmarks and the small circles are nodes. And the red lines show the linked path from a node to landmark. This Figure is now showing the paths to landmarks only having minimum hop. Here, node 120 knows the positions of 9 landmarks so that it is possible for node 120 to have more than 9 coordinates. And also, there can be multiple paths from a node to a landmark. This situation is shown in Fig. 4. The sensor node 96 has two paths to a landmark 2 far from two hops. It means that node 96 will have two coordinates calculated based on landmark 2. The choice for a coordinate among multiple coordinates based on one landmark will contribute improving accuracy of the position.

This paper has proposed the algorithm to update or determine the position in the situation described. The first algorithm is called "minimum hop method." This method calculates coordinate of a node from coordinate of the landmark having least hops to a node. So this method needs the information of hop counts to landmarks but will be most simple. The second method is the "minimum distance method." This method is similar to the first method, which is considering the landmark coordinate closest to a node. In the next article, the results using the second algorithm will not be shown because the accuracy is not good. The third method is called "simple mean." Simple mean method will average the

Fig. 3 Paths from arbitrary node (120) to landmarks with minimum hops

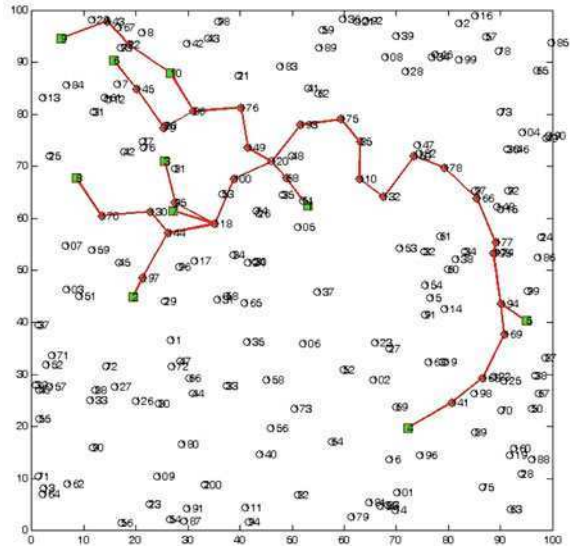
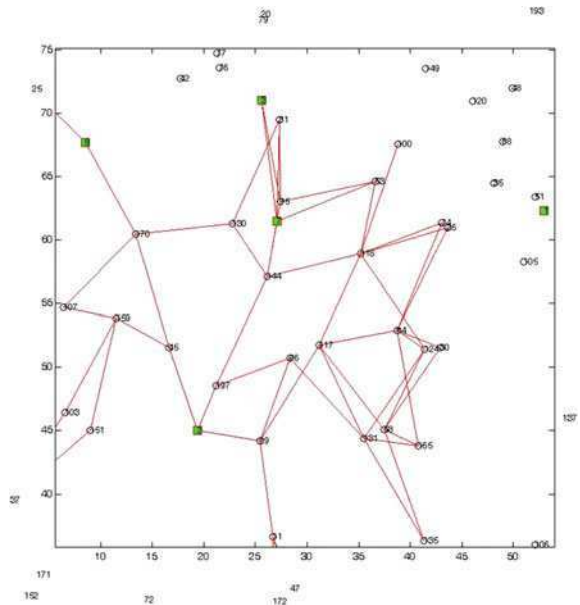


Fig. 4 Multiple paths of a node to a landmark with the same hop count



coordinates calculated using connected landmarks. The fourth method is called “multipath mean.” This method will take advantage of multiple paths to landmarks. As explained before, a node has possibility to calculate multiple position coordinates even based on one landmark because of multiple paths. Therefore, multiple coordinates will be averaged in this method.

4 Result and Discussion

This article will show the simulation results and compare with the results of DV-hop and DV-distance method [1]. The proposed method will be susceptible to angle error because the node's azimuth is calculated from a landmark far apart. This becomes severe when a landmark is far apart from. This paper, measured distance is assumed to be Gaussian. And it is assumed the measured angle is also Gaussian.

$$r_{\text{meas}} = r_{\text{exact}}(1 + \sigma_r N(0,1)) \quad (5)$$

$$\theta_{\text{meas}} = \theta_{\text{exact}} + \sigma_\theta N(0,1) \quad (6)$$

where $r_{\text{meas}}(\theta_{\text{meas}})$ is the measured distance (angle), $r_{\text{exact}}(\theta_{\text{exact}})$ is the true value of the distance(angle), $\sigma_r(\sigma_\theta)$ is a specific constant [7], and $N(0,1)$ is a normally distributed random variable. Therefore, the noise error in measured values is modeled as additive and can be varied by changing the specific constant $\sigma_r(\sigma_\theta)$ where in Matlab[®] program, $\sigma_r(\sigma_\theta)$ in Eqs. 5 and 6 roles standard deviation of normal distribution, so we will simply call it a standard deviation.

Figure 5 shows the location error and coverage with the node density, where coverage means the ratio of nodes calculating the coordinate to all nodes. Node density [8, 9] can be calculated,

$$d = \frac{N(\pi R^2)}{A} \quad (7)$$

where N is the total number of sensor nodes deployed in sensor field, A is the area of sensor field, R means the transmission range of nodes which is the same among all nodes. Localization error obtained in the simulation is defined by

$$L_{\text{error}} = \frac{\sum_{i=1}^N |r_{\text{calc}} - r_{\text{real}}|}{N} \quad (8)$$

where N is the total number of nodes, r_{calc} is the calculated coordinate of a node, and r_{real} is the real coordinate of a node.

In this simulation, sensor field is considered to be square with length 100 m. The transmission range is 10 m. As shown in Eq. 7, node density can be changed according to varying the number of nodes or transmission range when the area of sensor field is fixed. The results in Fig. 5 are obtained from varying the number of nodes.

In this simulation, we restrict the hop count to landmarks of 5 in order to prevent the propagation of angle error. That is, when calculating the coordinates of nodes, only landmarks within 5 hops far from a node will be considered. The larger hop counts, the larger coverage but the larger localization error. Figure 5 shows the results using the method proposed in this paper are better than those from DV-hop or DV-distance. DV-hop or DV-distance method show the localization

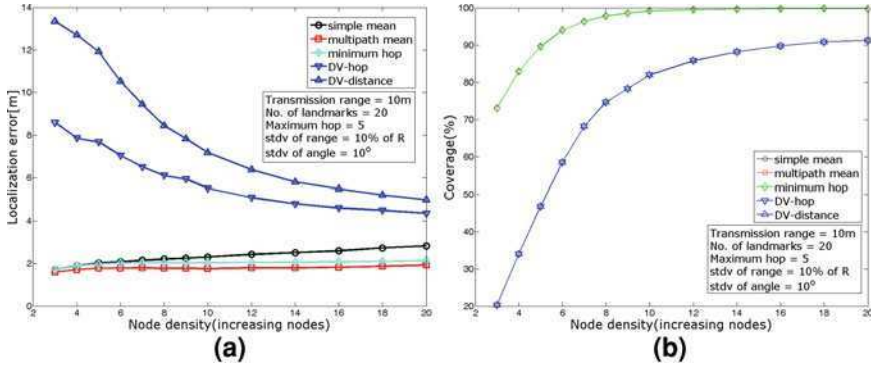


Fig. 5 Localization error and coverage with the variation of node density (node density is calculated by changing the number of nodes)

error goes lower according to increasing the node density. The coverage, however, does not reach 100% even increasing the node density to 20. In the case of our method, multipath mean shows best localization error and the localization error increased slightly with the node density. Next good result is obtained using the simple mean. And then, minimum hop shows relatively bad result. On the other hand, the coverage obtained from our method reaches nearly 100% at node density of about 10 but the DV-methods do not. This is because in this simulation we determine the nodes find their coordinates only when the localization error is smaller than the transmission range (R). Due to the algorithm’s simplicity, DV-methods result in relatively large error so that coverage from these methods is seemed not to reach 100%.

Figure 6 is the simulation result obtained by varying the transmission range. The trend of result is similar to that of Fig. 5. The coverage, however, are sharply increased in DV-methods. This is, as explained before, fully related to the method determining the found node. In this case, increasing the transmission range causes more nodes to find their coordinates with larger error and then leads to the increased coverage. On the other hand, coverage from DV-methods reaches nearly 95% at over 12 of node density but our method shows 95% at node density of 6. This results from merit of angle measurement as well as distance measuring [10]. Other difference from Fig. 5 is the localization error obtained using “minimum hop method.” Other than “simple mean” or “multipath mean”, this method shows improving the localization error with increasing the transmission range.

The effect of increasing the number of landmarks on the localization error and coverage is shown in Fig. 7. Increasing the number of landmarks can make hops to a node small so that the accumulation of angle error can be reduced and allow the node to choose the landmark with the small error. And the distance from a node to a landmark will be guessed with small error. Figure 7a shows such result but in case of coverage, because even though the number of landmarks is increased, all nodes cannot find the landmarks. In this simulation, because we restrict the hop

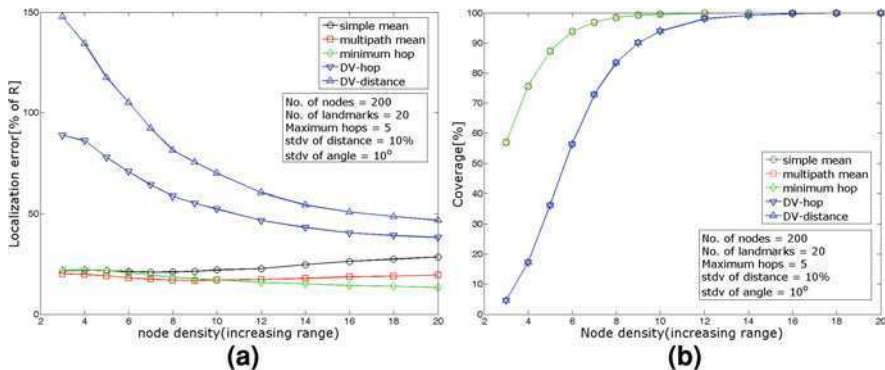


Fig. 6 Localization error and coverage with the variation of node density (node density is calculated by varying transmission range)

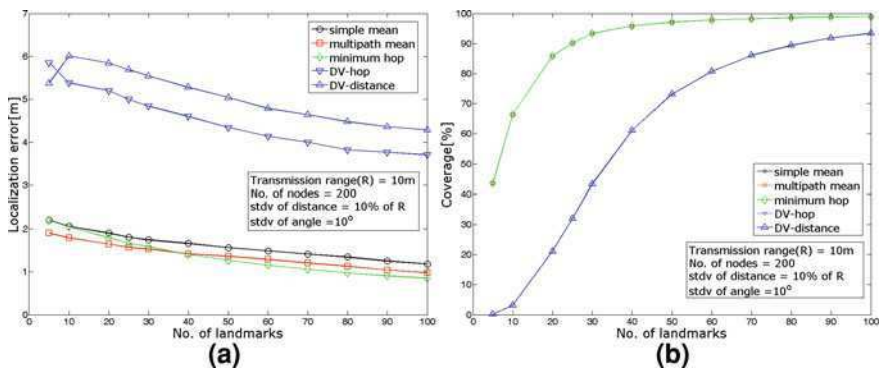


Fig. 7 Localization error and coverage with the number of landmarks

counts to 5 to find landmarks, there are still isolated nodes from landmarks even though increasing the number of landmarks.

5 Conclusion

In this paper, we introduce the localization method of sensor nodes in ad-hoc wireless sensor network (WSN). This method assumed that the nodes have the ability to measure the mutual distance and relative angle to their neighbors within transmission range. The proposed algorithm starts at measuring the angle and distance at the landmark and their neighbors and then finding their azimuths and incident angles. Obtained incident angle and measure mutual distance allows a node to calculate it's coordinate. This process will continue until all nodes with connections to any landmarks calculate their coordinates.

Then, sensor nodes are able to have a number of coordinates from many connected landmarks and multiple paths to landmarks. To utilize this, we propose four methods to determine the position among a number of coordinates. The proposed methods show better localization error and coverage than DV-methods. But our methods seem to be affected by the propagation and accumulation of measured angle error. So we restrict the hop counts to reach the landmark from a node with 5. That is, when calculating the position of a node, only coordinates of landmarks within five hops are considered. Validity of our method is confirmed by showing the simulation results with various conditions.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2010-0013411) and this work was supported by the industrial Strategic technology development program (10037290, Development of Smart Growth Management System) funded by the Ministry of Knowledge Economy (MKE, Korea).

References

1. Niculescu D, Nath B (2001) Ad-hoc positioning system (APS). In: Proceedings IEEE global telecommunications conference (GLOBECOM'01), vol 5. San Antonio, USA, Nov 29, pp 2926–2931
2. Niculescu D, Nath B (2003) Ad Hoc positioning system (APS) using AOA. In: Proceedings of the 23rd conference of the IEEE communications society (INFOCOM'03), April 2, pp 1734–1743
3. Stefano GD, Petricola A (2008) A distributed AOA based localization algorithm for wireless sensor networks. *J Comput* 3(4):1–8
4. Savvides A, Han C, Strivastava MB (2001) Dynamic fine-grained localization in ad-hoc networks of sensors. In: *Mobile computing and networking*, pp 166–179
5. Shang Y, Ruml W (2004) Improved MDS based localization. In: Proceedings of the 23rd conference of the IEEE communications society (INFOCOM'04), March, pp 2640–2651
6. Biswas P, Ye Y (2004) Semidefinite programming for ad-hoc wireless sensor network localization. In: Proceedings of the third international symposium on information processing in sensor networks, pp 46–54
7. Biswas P, Aghajan H, Ye Y (2005) Integration of angle of arrival information for multimodal sensor network localization using semidefinite programming. In: Proceedings of 39th Asilomar conference on signals, systems and computers, pp 1–9
8. Chintalapudi KK, Dhariwal A, Govindan R, Sukhatme G (2004) Ad-hoc localization using ranging and sectoring. In: Proceedings of the 23rd conference of the IEEE communications society (INFOCOM'04), pp 2662–2672
9. Liu K, Wang S, Zhang F, Hu F, Xu C (2005) Efficient localized localization algorithm for wireless sensor networks. In: Proceedings of the fifth international conference on computer and information technology (CIT'05), pp 517–523
10. Chintalapudi KK, Dhariwal A, Govindan R, Sukhatme G (2003) On the feasibility of ad-hoc localization systems. Technical report, Computer Science Department, University of Southern California, Los Angeles

Part IV
Intelligent Robotics, Automations,
Telecommunications Facilities,
and Applications

An Improved Localization Algorithm Based on DV-Hop for Wireless Sensor Network

Long Chen, Saeyoung Ahn and Sunshin An

Abstract Localization information is necessary and has become more and more important with the tremendous applications in the wireless sensor network. DV-Hop is a kind of range-free localization algorithms. Since the sensor nodes position is determined by hop-size estimation. Reducing the estimated hop-size error can improve sensor nodes position accuracy. As a result, we provide a novel solution for locating the sensor nodes using weighted value according to distance influence, so that we will get the estimated position will be much closer to its real position without additional hardware support in this paper. Simulation results demonstrate that the performance of the proposed algorithm is better than that of the DV-Hop algorithm.

Keywords Wireless sensor networks · Localization · DV-Hop

1 Introduction

A Wireless Sensor Network (WSN) consists of a large number of sensor nodes which are deployed in a monitor area, and they form a self configuring multi-hop network by the way of wireless communication [1]. WSNs distinguish themselves

L. Chen (✉) · S. Ahn · S. An
Department of Electronics Engineering,
Korea University, Seoul, Korea
e-mail: kulcn5032@gmail.com

S. Ahn
e-mail: syahn@korea.ac.kr

S. An
e-mail: sunshin@dsys.korea.ac.kr

from other traditional wireless or wired networks through sensor and actuator based interaction with the environment. Such networks have been proposed for various applications including search and rescue, disaster relief, target tracking, and smart environments. In these applications, data are collected by nodes must combine with location information to explain where an event has happened. Thus, the location information of nodes plays a very important role in WSN [2].

Based on whether it is required to measure the distance between two nodes or not, we divide localization protocols into two categories [3] range-based and range-free. Range-based protocols need to measure point-to point distance or angle for calculating location, and Range-free protocols calculate location by estimating hop count and hop size. Because of the hardware limitations of WSN devices in such as those outlined above applications, solutions in range-free localization are being pursued as a cost-effective alternative, the classical range-free algorithms include: APIT algorithm [3], Centroid algorithm [4], Amorphous algorithm [5] and DV-Hop algorithm [6].

In Centroid algorithm, nodes can use the centroid of composition created from their proximate reference beacons for positioning. The method depends entirely on the network connectivity, so the method can only achieve low-accuracy positioning. Besides, we require higher density of beacon nodes in this algorithm. DV-Hop algorithm is a distributed localization algorithm using the distance vector routing, which has the advantages of higher precision, and the method is simple. In Amorphous algorithm, nodes position is calculated by using beacon node communication radius instead of average hop distance, although this algorithm is an improvement of DV-Hop algorithm, the estimated hop distance value is so large that the enhancing localization accuracy is unobvious.

In this paper, we propose an improved DV-Hop algorithm to decrease localization error of the original DV-Hop algorithm. In the presented approach, the hop size is corrected by using weighted value. Our estimated position can be closer to the actual position. Comparing to the original algorithm, simulation results show that greater localization accuracy can be achieved in the improved algorithm.

The rest of this paper is organized as follows. [Section 2](#) has a description of related work about DV-Hop algorithm. [Section 3](#) presents the improved algorithm. [Section 4](#) compares and evaluates the improved algorithm performance by simulations. Finally, we draw our conclusion in [Sect. 5](#).

2 Related Work

Niculescu and Nath proposed DV-Hop localization algorithms that there is no need to directly measure the distance between nodes, the original DV-Hop algorithm is a sort of the calculation of hops number based on distance-vector algorithm [6, 7]. The basic principle is to use the product of the average hop size and the number of hops (hop count) between unknown node and beacon node to represent the distance between them, and then obtaining the unknown node location information by using Trilateration positioning method [8].

The process realization of DV-Hop algorithm can be divided into the following three steps.

2.1 First Step: Calculating the Minimum Number of Hops Between Unknown Node and Each Beacon Node

Each beacon node broadcasts a message, including their location information and the hop count value initialized to zero $(x_i, y_i, 0)$. When all neighbor nodes receive the message with hop count value from the surrounding beacon nodes, saving and broadcasting a new message with information of hop count value plus one $(x_i, y_i, 1)$ to their neighbors. And this process will continue till the whole network nodes obtain the messages from their adjacent beacon nodes. Receiving node will only save the minimum hop count among many hop counts has received from beacon nodes, and moreover, ignoring the message of a larger hop count value from a same beacon node.

2.2 Second Step: Estimating the Average Single Hop Size

When each beacon node $i (x_i, y_i)$ receives enough hop information from the other beacon node $j (x_j, y_j)$, and their distances can be shown as $\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$, the single hop size $HopSize_i$ can be estimated, and further, the message will be broadcasted throughout the network.

$$HopSize_i = \frac{\sum_{i \neq j} \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{\sum_{i \neq j} h_{ij}} \tag{1}$$

where h_{ij} is hop count value from the node j to node i .

2.3 Third Step: Calculating the Unknown Node Location

Unknown node receives the estimated distance to a beacon node value, and estimates their location using the trilateration position method. As shown in the following formula

$$\begin{cases} (x_1 - x)^2 + (y_1 - y)^2 = d_1^2 \\ \vdots \\ (x_n - x)^2 + (y_n - y)^2 = d_n^2 \end{cases} \tag{2}$$

where n denotes beacon node number, and (2) can be expressed as

$$\begin{cases} x_1^2 - x_n^2 + 2(x_n - x_1)x + y_1^2 - y_n^2 + 2(y_n - y_1)y = d_1^2 - d_n^2 \\ \vdots \\ x_{n-1}^2 - x_n^2 + 2(x_n - x_{n-1})x + y_{n-1}^2 - y_n^2 + 2(y_n - y_{n-1})y = d_{n-1}^2 - d_n^2 \end{cases} \quad (3)$$

And (3) can be also represented using linear equation

$$AX = B, \quad (4)$$

where

$$A = \begin{bmatrix} 2(x_n - x_1) & 2(y_n - y_1) \\ \vdots & \vdots \\ 2(x_n - x_{n-1}) & 2(y_n - y_{n-1}) \end{bmatrix} \quad (5)$$

$$B = \begin{bmatrix} x_1^2 - x_n^2 + y_1^2 - y_n^2 + d_n^2 - d_1^2 \\ \vdots \\ x_{n-1}^2 - x_n^2 + y_{n-1}^2 - y_n^2 + d_n^2 - d_{n-1}^2 \end{bmatrix}, \quad X = \begin{bmatrix} x \\ y \end{bmatrix} \quad (6)$$

The solution can be got by using standard least-squares approach, we have

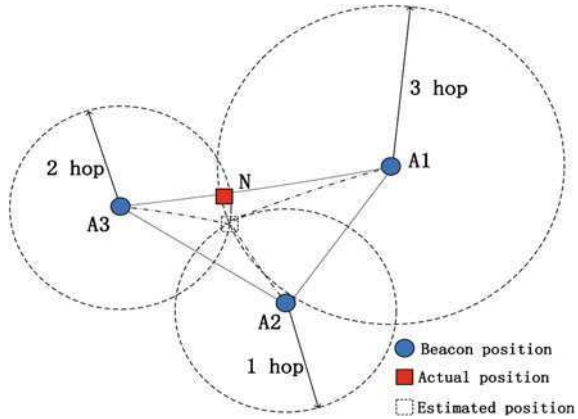
$$X = (A^T A)^{-1} A^T B. \quad (7)$$

3 Improved DV-Hop Algorithm

In the DV-Hop algorithm description, it can be concluded that beacon node distribute, beacon amount, node density and the average hop size for one hop has remarkable effect on the accuracy of estimation. The unknown nodes compute the distance to the beacon nodes based on the minimum hop size and the number of hops to the beacon nodes. However, an error might occur during the process, especially when there are greater than or equal to two of hop count, the node actual position will be difficult to determine accurately. In other words, the error increases as a result of the increase in the number of hops. Hence the accuracy of selected beacon node hop-size will directly affect the precision of localization. For this problem, we improve DV-Hop algorithm focus on correcting the hop-size based on weighted value.

Unfortunately, the calculated location of the node is often not the actual position. From Fig. 1, we describe the position errors of unknown nodes using DV-Hop algorithm, and present a way to reduce error that enhancing our positioning accuracy.

Fig. 1 Location estimation error by trilateration algorithm



Where A1, A2 and A3 are beacon nodes, N is unknown node. The dashed circle indicates the hop count value from beacon node to unknown node. As shown in Fig. 1, there are three hops from node N holds to node A1, one hop to node A2 and two hops to node A3. The estimated position of node N may be existed in the intersection of three dashed circles. However, because of the existence of hop size error, there is a large probability that the node actual position is not in our estimated position as indicated in Fig. 1.

In order to solve the problem by improved DV-Hop algorithm, we consider the following example is shown in Fig. 2.

We propose an improved approach of calculated hop size with weighted value at the first step, aiming at the error reduction between estimated position and actual position. Since all nodes are not completely uniform distribution in the area, an unknown node may receive several hop size messages from the surrounding beacon nodes, computing the average hop size of them received from different beacon nodes. The closer a beacon node is to the unknown node, the greater effect the beacon node has on average hop size calculated. As shown as in the following defined equation

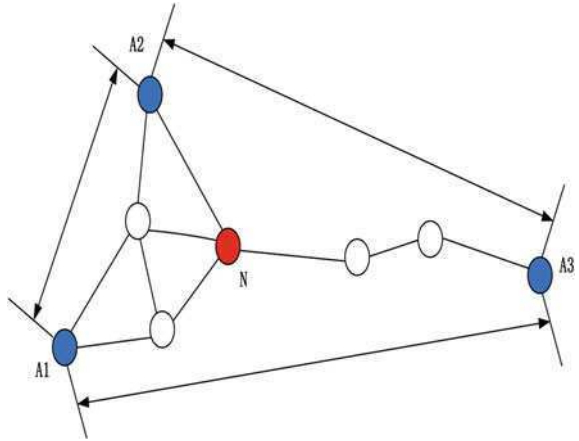
$$d_b = \frac{\left\{ \sum_{i=1}^n \frac{H_i}{\sum H_i} HopSize_i \right\}}{n} \tag{8}$$

where db is the average hop size of the number that n beacon nodes, H_i is hop count and $HopSize_i$ is obtained using (1).

In the improved algorithm, a new correction value cv will be introduced at the second step. The average hop size error correction value between beacon nodes is given by

$$cv = \sum \{ |d_i - d_b \times hop(i,j)| / hop(i,j) \} \tag{9}$$

Fig. 2 An example of improved DV-Hop algorithm



In the equation, $d_t = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ denotes the true distance between two nodes; $hop(i, j)$ is the hop count from node i to node j ; $|d_t - db \times hop(i, j)|$ is the absolute value of difference between a true distance and an estimated distance. Therefore, an improved hop size value $HopSize_c$ is given by:

$$HopSize_c = HopSize_i + cv. \tag{10}$$

At the final step, we can obtain the estimated position using the following equation

$$HopSize_c \times hop(i, j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \tag{11}$$

In this way, the node estimated position result is much closer to the actual position than the original DV-Hop algorithm.

Based on this scheme, we can efficiently decrease the effect of hop size error on node localization and improve the localization accuracy.

4 Simulation Results

Our performance evaluation focuses on the localization accuracy. For this purpose, we conducted simulations using MATLAB [9] to compare and analyze the performance of the algorithms proposed with the original DV-Hop localization algorithm. In the initial simulation experiments, 200 sensor nodes were randomly distributed in a 100 m × 100 m square area, we assume that sensor nodes are in the isotropic dense network which can achieve relatively reasonable localization accuracy, and the sensor node have the same maximum radio range R is set to 30 m. To evaluate the effectiveness and the availability of the improved algorithm,

Fig. 3 Localization error

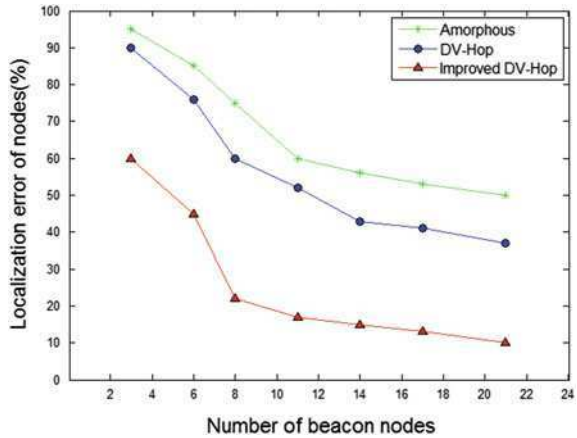
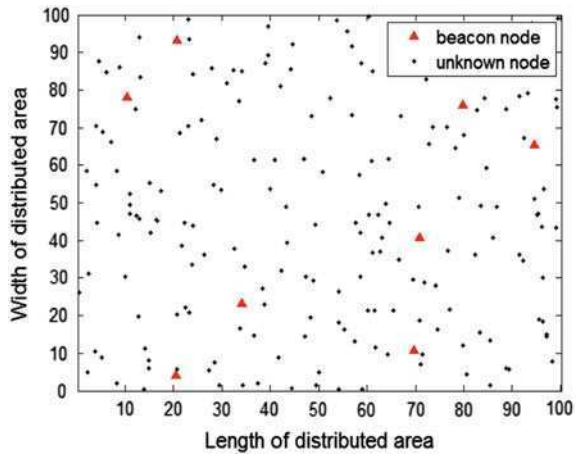


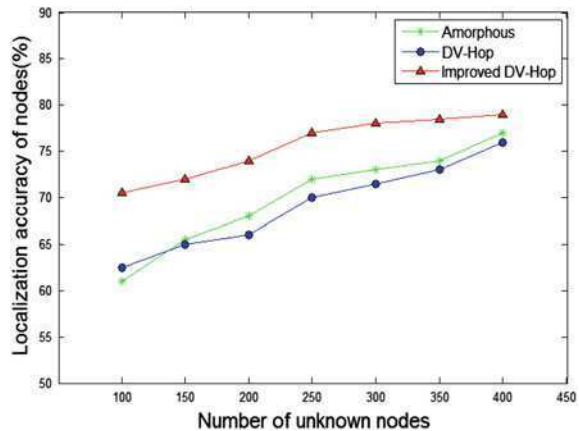
Fig. 4 Node distribution



we have made 50 times different experiments and obtained the average of results, analyzing the localization error depending on the number of beacon nodes.

As shown as in Fig. 3, there are only a smaller number of beacon nodes in the case. The improved algorithm has no significant improvement of localization performance, since there is comparatively little data on correcting the average hop size. However, the average hop size for estimating node position can be corrected well with an increase in the number of beacon nodes. Comparing with the original DV-Hop algorithm and Amorphous algorithm, the node reduction of localization error shows significant advances in the improved algorithm, according to the simulation results, we can get the reduction of total average localization error rate that is 30%. Especially, we can observe a reduction of localization errors from

Fig. 5 Localization accuracy



60 to 21% when there are eight beacon nodes in the simulation result, the improved algorithm offers a significant reduction of localization error of nodes.

In the next simulation, as shown as in Fig. 4, we consider the case that the number of beacon nodes is fixed to 8 and the number of unknown nodes is fixed to 200. Then we compare the performance of algorithms and analyze the relationship between localization accuracy and the number of the unknown nodes, the simulation results are shown in Fig. 5 that the localization accuracy of the algorithms are all improved as the number of unknown nodes increase. Since the node density is directly proportional to the number of unknown nodes, in other words, the connectivity of whole network is increased. Moreover, Simulation results indicate that the average localization accuracy can rise by about 10% from the DV-Hop algorithm to the improved algorithm. On the whole, the improved algorithm has a better performance than the original DV-Hop algorithm in localization accuracy of nodes.

5 Conclusion

In this paper, we have proposed a novel localization method for improving the original DV-Hop algorithm with weighted value. The improved method introduces a corrected value to obtain a corrected average hop size value so as to decrease the localization error, and the estimated position is much closer to the actual position using the new hop size value. Besides, the method is not influenced by the nodes distribution. When we need to achieve the same localization accuracy, the number of beacon nodes required in the improved algorithm is less than that in original DV-Hop algorithm. In other words, the cost of whole Wireless Sensor Network can be reduced, which is able to be selected as a practical locating scheme. The simulation results show that our improved method can increase the localization

accuracy of the original DV-Hop algorithm. But the drawback of the algorithm is that we can achieve relatively reasonable localization accuracy only in the isotropic dense network. Therefore, designing an algorithm that can be extended to more general network as improving the node localization accuracy is our future research objective.

References

1. Estrin D, Govindan R, Heidemann J, Kumar S (1999) Next century challenges: scalable coordination in sensor networks. In: Proceedings of the ACM MobiCOM 1999, Seattle, pp 263–270
2. He T, Huang C, Blum BM, Stankovic JA, Abdelzaher T (2003) Range-free localization schemes for large scale sensor networks. In: Proceedings of the ACM MobiCom 2003, San Diego, pp 81–95
3. Bulusu N, Heidemann J, Estrin D (2000) GPS-less low cost outdoor localization for very small devices. *systems. IEEE Pers Commun Mag* 28(5):28–34
4. Nagpal R, Shrobe H, Bachrach J (2003) Organizing a global coordinate system from local information on an ad hoc sensor network. In: The 2nd international workshop on information processing in sensor networks (IPSN'03), Palo Alto, April 2003
5. Niculescu D, Nath B (2001) Ad Hoc positioning system (APS). In: Proceedings of the IEEE GLOBECOM 2001, San Antonio, pp 2926–2931
6. Niculescu D, Nath B (2003) DV based positioning in ad hoc networks. *J Telecommun Syst* 22(1/4):267–280
7. Chan YT, Ho KC (1994) A simple and efficient estimator for hyperbolic location. *IEEE Trans Signal Process* 42:1905–1915 August 1994
8. Doherty L, Pister K, Ghaoui LE (2001) Convex position estimation in wireless sensor networks. In: IEEE INFOCOM 2001, Anchorage, AK
9. Hanselman D, Littlefield B (1996) *Mastering MATLAB*. Prentice Hall, Upper Saddle River

A Design of Intelligent Smart Controller for Object Audio-based User's Active Control Service

Jong-Jin Jung and Seok-Pil Lee

Abstract The intelligent smart controller introduced in this paper is a kind of smart remote controller for providing various an audio playing information and user interface interacted with DSP-based main platform. It is implemented in shape of web-application and android application. The Object-based audio service provides the combination of multi-object audio source (e.g. vocal, guitar, drum, etc.) for user. This property of object-based audio service enables users to actively play the music (e.g. object add or remove, object position change, recreate own music) during presenting audio. For audio channel mixing and band filtering, TMS320C6727 DSP and several peripheral devices are used. And for user interface of the control of audio, web-application and android app that are working in iPad, android-based smart phone is developed.

Keywords Smart controller · Object-based audio service · User interactive audio control

1 The Object-Based User's Active Audio Service

So far audio service is audio developer-oriented music service, that is, the vocal and all instruments are mixed into single audio source is given to user. User just controls simple audio control such as volume control, track move, skip timeline, etc.

J.-J. Jung (✉) · S.-P. Lee

Digital Media Research Center, Korea Electronics Technology Institute,
9FL, Electronics Center, #1599 Sangam-dong, Mapo-gu,
Seoul 121-835, Korea
e-mail: Jong-JinJung@keti.re.kr

S.-P. Lee

e-mail: Seok-PilLee@keti.re.kr



Fig. 1 The Object audio-based user's active Control User Interface Using iPad web-application

Thus cannot help hearing passively music. But in case of object audio-based user's active control service, user can actively control object-based audio with own musical tastes. User can add, remove or control the each objects (e.g. vocal track, guitar track...), add sound effect, change of the optimized hearing position (Sweet-spot), change of object audio source and speaker position. And instead of singer's vocal track, user can add their own recorded vocal to instrumental music that is played by professional session, as if user were a single. And then so user can generate 2ch music with this combination of objects. Finally, user can recreate own creative music, copy to own mobile multimedia devices and can enjoy them. Figure 1 shows the object audio-based user's active control service using smart user interface that is implemented in web-application of iPad and iPhone. The left side of Fig. 1 shows positions of object audios. The center point is a user's positions and the icon is a position of each instruments. Thus user can change a position of an arbitrary instrumental to the left-bottom side. If so, that object sound is heard from southern west direction. The right side of Fig. 1 shows user's control about objects. Using right side interface, user can sound on/off, add/remove, volume up/down an arbitrary objects. Also user can save the current configuration set and export user's music to own mobile multimedia devices.

2 The Object-Audio-Based Presentation Platform

2.1 The Implementation of the Object-Audio-Based Presentation Platform

The object-audio-based presentation Platform act in object audio control and rendering, channel mixing, band filtering, interacting with user interface, 3D-sound

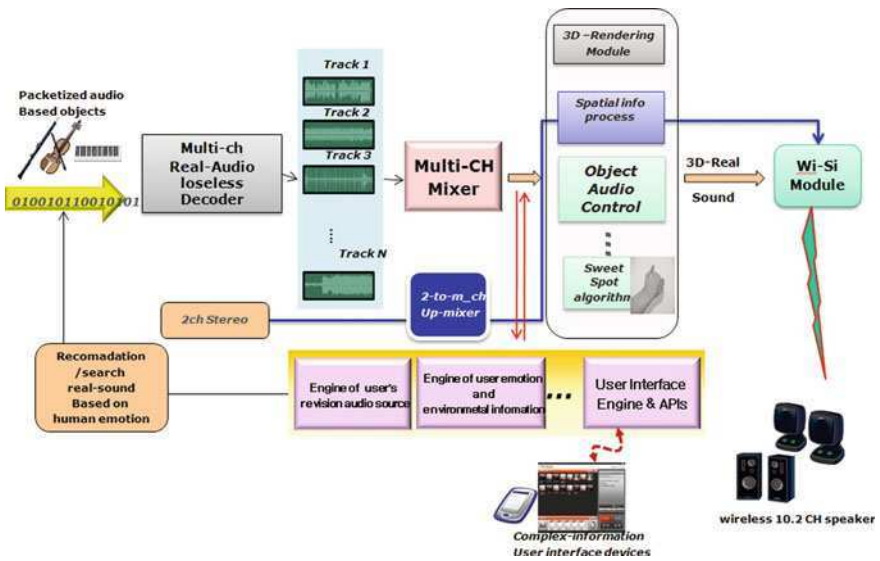


Fig. 2 Block diagram of the object audio-based presentation platform

effect rendering, search the optimized user’s hearing position(sweet-spot search), spatial information of object audio source and speaker position, up-mixing 2ch-to-multi-ch, processing user information data, and so on. The detail functions of this platform are illustrated in Fig. 2. For the implementation this platform, 2 DSP chips (TMS320C6727 DSP chip), DAC, USB 2.0, etc. are used. The 1st DSP block executes decoding of object-based audio, object audio mixing, object audio processing based on positions and state (e.g. mute on/off, volume level, combination of active objects, etc.) of each object audio. The 2nd DSP block processes the various realistic 3D-rendering, spatial information processing, sweet spot search, etc. And DAC is used to output the final user’s mixed audio to multichannel speakers. Figure 3 shows the hardware architecture of this platform.

In addition to these hardware chips, this platform has the web-server engine for user-friendly smart graphical interface (Actually, iPad UI is implemented for user’s smart remote control) and wireless modem module with which we can hear real-sound with wireless speakers.

2.2 The Implementation of Smart Remote Controller Using Web-Application and Android Application

The smart remote controller provides a graphical interface with which user can get the various information of audio playing information, and control the process of object audio rendering. It is designed in web-application and android application.

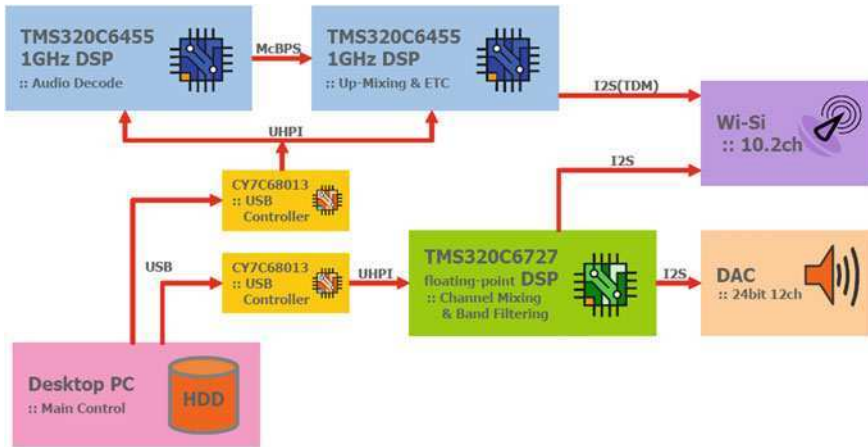


Fig. 3 The Hardware architecture of the object audio-based presentation platform

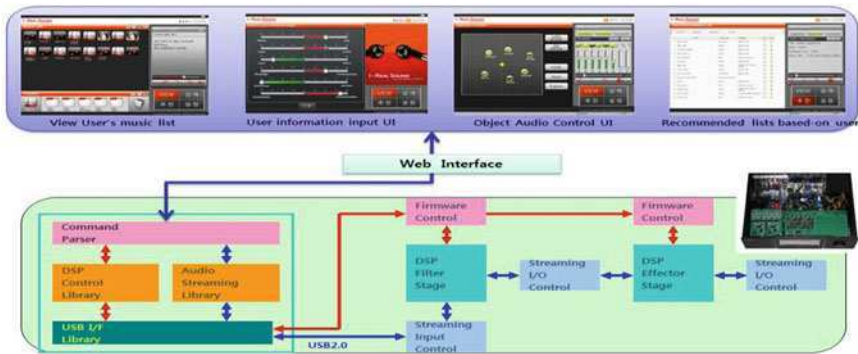


Fig. 4 The interface platform and remote controller and example of graphical user interface

In case of web-application, web-server engine is implemented in the DSP platform and web-client engine is implemented in smart remote device (iPAD, iPhone). Thus user can easily utilize smart remote controller using own smart phone or smart pad that must support web browser, if user access web-server IP address or wireless access-pointer. Figure 4 show the interface smart controller with DSP platform, and several example of smart controller user interface.

For design of object-based audio play module in android platform, 3 mobile android devices (android OS2.1) that are Nexus-one of HTC, Galaxy-S and Galaxy-Tab of Samsung are used in this paper. The structure of each module is like Fig. 5. In case of android application, user can easily enjoy that service while user is moving. Figure 6 shows a designed android application. User inputs own information for recommending lists from server and receive the recommended lists and can select one of them. During playing, user can add or remove an arbitrary track (object), thus new mixed music is played.

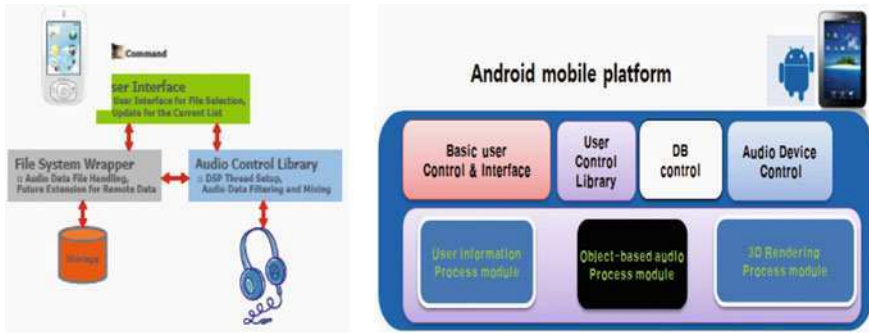


Fig. 5 The Structure of object audio-based user's active control modules in Android Platform



Fig. 6 Android App of object-based real sound service in android mobile devices

Real sound control library structure implemented in android platform is explained in Fig. 2. “Basic user control & Interface” module provides user for basic user interface for interaction, such as a user input, an audio information and audio control.

2.3 Structure of User Control Library

User control algorithm for real sound play is designed based on object-based audio control. User control library designed by user control algorithm has major 3 functions. The first is an object-based audio control module, the second is a user management system interoperated user information-based recommendation server and the last is recommendation interoperation system. An object-based audio control module provides a basic audio play, pause, stop, volume control, add or remove object, object position change, and so on. The user management system sends all user information to server, and receives the recommended lists from server.

2.4 Structure and Control of Object-Based Real Sound

The function of object-based audio control has core three parts that are an object control, a preset management and a user information-based recommendation system. The object audio control manages not only a basic audio play, pause, stop, volume control, but also a preset track add or remove, object add or remove, object position change. Its basic function is control of object mixing by calling audio control command. The basic track information of object audio is “preset”. The preset is a kind of configuration set, and all information for playing object-based real sound is stored in preset. Preset is expressed in XML metadata defined TV-Anytime Specification. User can add an arbitrary object to “preset” or remove an object from “preset” and save the current configuration set (object combination, objects hearing position, the current user information and etc.) to another new “preset”.

3 The Implementation Results of Real Sound Play in Android Platform

For design of object-based audio play module in android platform, this paper uses three android mobile devices that adopt android OS 2.1. They are a Nexus-one of HTC, Galaxy-S and Galaxy-Tab of Samsung. The Fig. 6 shows a designed application in android. User inputs own information for recommending lists from server and receive the recommended lists and can select one of them. During playing, user can add or remove an arbitrary tracks (object), thus new mixed music is played.

4 Conclusion

This paper introduced the object audio-based user’s active control service in DSP and android platform. In this service, user can actively control the object-based audio, change track source position to arbitrary direction, make a combination with only user’s selected track and recreate own creative music. Especially, the service implemented in mobile devices make user enjoy this service without restrict of places. This service may provide not only an interesting audio service, but also the momentum of audio industry. In the future, more study on this service is not only beneficial to audio content provider, service provider, device manufacture and user, but also can cause a great, creative and innovative change to audio market.

References

1. Jang I, Seo J, Kang K, Kim HY (ETRI) Kevin Seung Chul Ham (Audizen Inc), MPEG2008/M15626. A proposal for technical specification of interactive music AF
2. ISO/IEC 14496-14:2003, Information technology—coding of audio-visual objects—Part 14: MP4 file format, Nov 2003
3. ISO/IEC 21000-9, Information technology—multimedia framework (MPEG-21), Part 9: file format, July 2005
4. Adobe flash. <http://www.adobe.com/products/flash>
5. Adobe systems (2007) Flash player for mobile devices delivers high-impact video and dynamic web content white paper, Oct 2007
6. Cho CS, Kim JW, Choi BH (2008) A low complexity MPEG-4 ALS coding for high quality object audio system. In: IEEE transactions CE, Dec 2008 submitted
7. Park K, Seo J, Wee J, Jeon W, Paik J, Wireless audio transmitting apparatus, speaker and system and controlling method, EP08173036.8 meters in Table 1

The Method of Main Vocal Melody Extraction Based on Harmonic Structure Analysis from Popular Song

Chai-Jong Song, Seok-Pil Lee, Kyung-Hack Seo
and Hochong Park

Abstract In this paper, we propose the method of main vocal melody extraction based on harmonic structure analysis technique from polyphonic music signal. It is the most important part of contents based music retrieval method which has mainly three parts. The first part is pitch estimation from humming signal, the second one is the melody extraction from polyphonic music signal and the last one is the matching engine which measure the distance between two vectors. The accuracy of melody extraction affects the overall system performance rather than any other parts. Human vocal track makes the harmonics like most musical instruments. This is one of the most important things that we have considered to utilize. So, we might extract the main vocal melody from the complicated mixed signal with musical instruments. We utilize harmonic structure analysis and track pitch sequence during three frames include current frame. The proposed method contains three major blocks named preprocessing, multi-pitch extraction with peak picking, fundamental frequency detection and the last part with pitch tracking, predominant melody detection. We have started this project with aiming for supporting commercial service for music portal provider, KARAOKE system and mobile devices.

Keywords QbSH · Multi-F0 · Melody extraction · Pitch contour

C.-J. Song (✉) · S.-P. Lee · K.-H. Seo
Digital Media Research Center, KETI, #1599, Sangam-dong,
Mapo-gu, Seoul, South Korea
e-mail: jcsong@keti.re.kr

S.-P. Lee
e-mail: lspbio@keti.re.kr

H. Park
Department of Electronics Engineering, Kwangwoon University,
Seoul, Republic of Korea
e-mail: hcpark@kw.ac.kr

1 Introduction

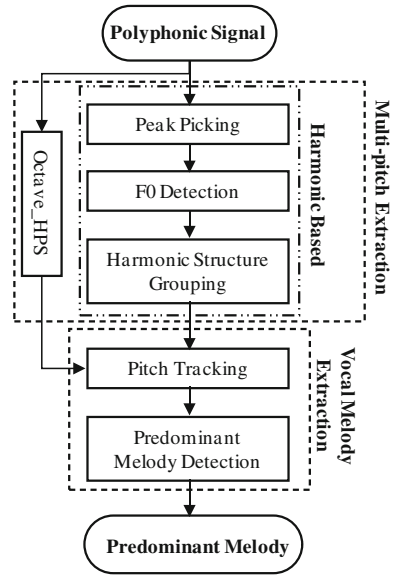
The best way of looking for the contents is tag based query method using metadata like title, singer, lyrics, something like that. It is also the most powerful tool we have ever experienced. However, tagging the contents is also a laborious and time-consuming work. So, contents based query technique has been considered as a complemented way. Especially in digital music domain Query by Singing/Humming (QbSH) has been researched for a long time with various methods [1, 2]. The most of those researches have been remained at monophonic data like humming or MIDI signal. In these cases it has a critical problem of data sparseness. It is necessary to build reference database from polyphonic music dataset to avoid this problem and provide commercial service. So, we propose the advanced main vocal melody extraction method based on harmonic structure analysis. This article draws the proposed method starting with briefly description of overall architecture of the QbSH system. This system is one of the traditional server and client model. At the client side, it takes humming signal from the input device during 12 s and judges that the input signal is tainted with background noise. If it does, it suppresses background noise and estimates pitch. After that, it sends estimated pitch to server after formatting it with MPEG Query Format (MP-QF) international standard. Server parses queried data and measures similarity between queried data and reference data which is taken from feature database, and then recommends top 20 candidates having highest similarity score to the client.

2 Proposed Melody Extraction Algorithm

Main vocal melody extracted from polyphonic music signal is used as the reference vector set of this QbSH system. Multiple fundamental frequencies as called multi-F0 have to be calculated before estimating main melody from polyphonic music signal. It has been mixed very complicated with various musical instruments and vocal sound at the same time. So, extracting main melody from complicated mixed signal is very hard work. This topic has been researched for so long time, but there is not any outstanding result; especially as the accompaniment is stronger than main vocal sound [3–6]. Having stronger beat patterns is one of the trends of recently popular music in some genres like dance, rock, and heavy metal, etc. Human vocal and most musical instruments except percussion instruments make harmonic structure. So, we utilize the harmonic structure analysis technique in order to make decision of multi-F0 candidates and track main melody sequence from calculated multi-F0 candidates.

Figure 1 depicts the procedure of proposed method. It can be divided by three main parts. The first one is pre-processing for emphasizing vocal sound using modified speech enhancement module taken from IS-127 Enhanced Variable Rate Codec (EVRC). The second one is multi-pitch extraction based on harmonic

Fig. 1 The procedure of main melody extraction



structure analysis technique. The last one is vocal melody extraction by tracking pitch sequence.

3 Pre-Processing

The input signal from the music database is sampled at 44.1 kHz with 16 bits per sample at stereo. This is down-sampled at 8 kHz and down-mixed into the mono channel before pre-processing in order to emphasize vocal sound. Every part is frame based processing which is windowed by 16 ms with Hanning window and has one frame look-ahead. Pre-processing makes this frame having harmonics or not by using Zero Crossing Rates (ZCR), frame energy, and deviation of spectral peaks. We introduce the vocal enhancement module based on the multi frame processing and noise suppression algorithm to improve accuracy of vocal pitch. It is modified from adaptive noise suppression algorithm of IS-127 EVRC speech coder which has the advantage of enhanced performance with relatively low complexity [7]. Windowed signal is transformed into frequency domain with Short Time Fourier Transform (STFT), and then grouping frequency signal into 16 channels. The gain is calculated with Signal to Noise Ratio (SNR) between input signal and noise level predicted by pre-determined method at each channel. Input signal is rearranged with this gain at each channel respectively. The noise suppressed input signal is obtained by inverse transformation. This article assumes the input signal as “vocal melody + accompaniment” while EVRC assumes the input signal as “voice + background noise”. This method improves accuracy rate up to maximum 10.7% for the melody extraction.

4 Multi-pitch Extraction

The multi-F0 candidates are estimated from the predominant multiple pitch calculated by the harmonic structure analysis. The multi-F0 is decided by grouping the harmonics into several sets by checking validation of its continuity and Average Harmonic Structure (AHS). The melody is obtained by tracking the estimated F0. Voiced or unvoiced frame is determined on the pre-processing stage as I mentioned at the previous section. If the current frame is judged to unvoiced frame, the algorithm assumes that F0 does not exist, otherwise does harmonic analysis. Multi-F0 is estimated through three processing module like peak picking, F0 detection and harmonic structure grouping. There are some peak combinations with F0 because polyphonic signal is mixed with several musical instrument sources. F0 having several harmonic peaks is evaluated by (1).

$$\begin{aligned} |X[k] > |[k - 1]| \text{ and} \\ |X[k] > |X[k + 1]| \text{ and} \\ |X[k]| \text{PTH}_{(l,h)} \end{aligned} \quad (1)$$

Here, $\text{PTH}_{l,h}$ is low and high band Peak Threshold (PTH) for local peaks. Because average energy is not same between low and high band of polyphonic music signal in general, it is divided at point of 2 kHz. Skewness (SK) of frequency envelop make decision of PTH adaptively. If $\text{SK} = 0$ then energy is symmetric, if $\text{SK} > 0$ then energy is leaned to low band, if $\text{SK} < 0$ then high band has the more energy than low band.

$$\begin{aligned} \text{IF SK} = 0, \text{ Then } \text{PTH}_{-1}, \text{PTH}_{-h} &= (\overline{X_a}) \\ \text{IF SK} < 0, \text{ Then } \text{PTH}_{-1} &= X_a - \sigma_a, \text{PTH}_{-h} = X_h - \sigma_h/2 \\ \text{IF SK} > 0, \text{ Then } \text{PTH}_{-1} &= X_a - \sigma_a/2, \text{PTH}_{-h} = X_h - \sigma_h \end{aligned} \quad (2)$$

Here, $\overline{X_a}, \overline{X_h}, \sigma_a, \sigma_h$ is mean value and standard deviation for full band and high band respectively. For example, F0 is limited from 150 to 1 kHz. Three F0 of 300, 400 and 150 Hz is shown as Fig. 2. It does not have the first peak of harmonics of 150 Hz. There are 3, 2, 5 peaks per each F0 respectively, but there are only 7 peaks because A0 and C1, A1 and C3, A2 and C5 are overlapped. You can obtain 21 peak distances for every two peaks. The harmonic relation is calculated between peak [v] and every F0 candidates (Fig. 3).

5 Vocal Melody Extraction

If all of the F0 satisfies the ideal harmonic structure, real frequency peak will be at the harmonic peak which they must be. Following this process, you can take 5 F0 candidates at 150, 200, 300, 400, and 450. F0 is assumed as the maximum

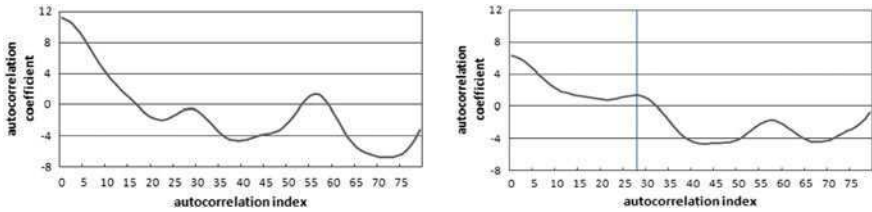


Fig. 2 Before vocal enhancement (left), after vocal enhancement (right)

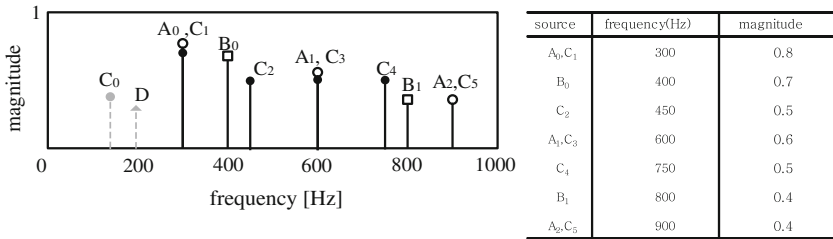


Fig. 3 Harmonic relationship of different F0 (left) and it's table (right)

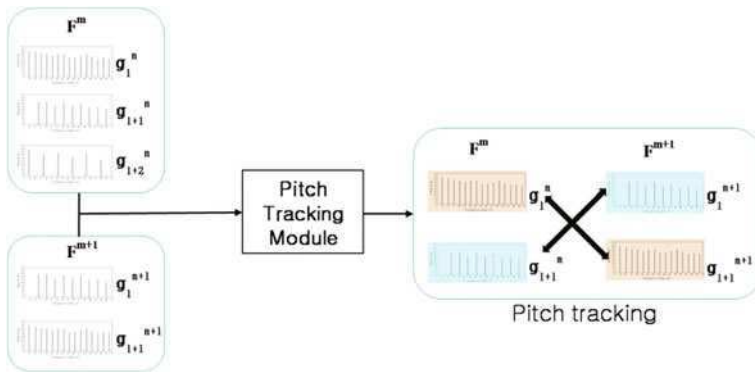


Fig. 4 Pitch tracking module

spectrum peak. AHS determine F0 significant degree by calculating the average energy of harmonic peaks. AHS is calculated by Eq. 3.

$$\bar{h} = \frac{2h_c}{N} \sum_{i=0}^{\frac{N}{2h_c}} h_a[i] \tag{3}$$

Here, \bar{h} is AHS and h_a is harmonics magnitude. Vocal melody extraction module is tracking estimated F0 candidates of three frames include current frame.

Table 1 Multi-pitch extraction result (*left*), MIREX 2009 melody extraction result (*right*)

	Participant	RPA(%)	RCA(%)
	Cao and Li	85.625	86.205
	Durrieu and Richard	86.96	87.398
	Hsu, Jang and Chen	63.11	74.101
	Joo, Jo and Yoo	81.959	85.798
	Dressler	85.969	86.424
	Wendelboe	83.135	86.593
	Cancela	86.962	87.545
	Rao and Rao	81.446	88.038
	Tachibana, Ono, Ono and Sagayama	59.768	72.129
	Proposed Method	90.418	92.27

For this tracking we use three factors as frequency, amplitude and phase. Tracking pitch is done by Eq. 4.

$$\begin{aligned}
 g_1^m &= \{f_0^m, f_1^m, \dots, a_1^m, \dots, \phi_1^m\} \\
 F^m &= \{g_1^m, g_2^m, g_3^m, \dots\} \\
 a_1^m &= \text{udr_}a_1^m
 \end{aligned} \tag{4}$$

Here, g_l^m means harmonic group of m th frame and l th F0, frequency, amplitude and phase. F^m means predominant frequency of m th frame a_1^m means amplitude of m th frame and l th candidates (Fig. 4).

6 Conclusions

We evaluate the melody extraction algorithm with two methods as Mean Reciprocal Rank (MRR) used on TREC Q&A and Raw Pitch Accuracy (RPA) and Raw Chroma Accuracy (RCA) used on Music Information Retrieval EXchange (MIREX) contest for melody extraction task [8]. MRR is calculated by Eq. 5. We evaluate multi-pitch extraction performance.

$$\text{MRR} = \frac{1}{N} \sum_{n=1}^N \frac{1}{\text{rank}_n} \tag{5}$$

Here, N is total frame and rank_n is the rank of extracted F0 against reference F0 at n th frame. We take the Audio Description Constest (ADC) 2004 dataset for evaluating the algorithm because the Korean dataset does not have the groundtruth. We evaluate two different methods as RPA and RCA. RPA is the accuracy between extracted and reference melody. RCA ignore octave errors from extracted melody. The result of evaluation is shown as Table 1.

References

1. Orio N (2006) Music information retrieval: a tutorial and review. *Found Trends Inf Retr* 1:1–90
2. Downie JS (2008) The music information retrieval evaluation exchange (2005–2007): a window into music information retrieval research. *Acoust Sci Tech* 29:4
3. Poliner G, Ellis DP, Ehamann AF, Gomez E, Streich S, Ong B (2007) Melody transcription from music audio: approaches and evaluation. *IEEE Trans Audio Speech Lang Process* 15(4):1066–1074
4. Eggink J, Broown GJ (2004) Extracting melody lines from complex audio, ISMIR
5. Klapuri AP (2003) Multiple fundamental frequency estimation by summing harmonic amplitude. *IEEE Trans Speech Audio Process* 8:6
6. Goto M (2004) A real-time music scene description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Commun* 43(4):311–329
7. TIA-EIA-IS-127, Enhanced Variable Rate CODEC
8. Audio melody extraction results. <http://www.music-ir.org/mirex/2009/index.php/>

The Fusion Matching Method for Polyphonic Music Feature Database

Chai-Jong Song, Seok-Pil Lee, Kyung-Hack Seo
and Kang Ryoung Park

Abstract This article proposes the fusion matching method for polyphonic music feature database which are extracted from music signal. The best way looking for the song is the tag based retrieval method using metadata like title, singer, lyrics, etc. This is very convenient and powerful way if you have already known about information of contents what you are looking for. But if you do not have any information of the contents, contents based query method might be a plan-B. Query by Singing/Humming (QbSH) is the powerful tool and the best supplemental method looking for song or music over the internet or among huge database. This topic has been researched for a so long time with various solutions. But, there have not been any outstanding solution so far. So we propose the fusion matching method with three matchers against polyphonic music signal in order to improve matching performance. Proposed method is based on Dynamic Time Warp (DTW), Linear Scaling (LS) and Quantized Binary Code (QBcode) and then combines them with fusion score based PRODUCT rule.

Keywords MIR · QbSH · DTW · LS

C.-J. Song (✉) · S.-P. Lee · K.-H. Seo
Digital Media Research Center, KETI, #1599, Sangam-dong,
Mapo-gu, Seoul, South Korea
e-mail: jcsong@keti.re.kr

S.-P. Lee
e-mail: lspbio@keti.re.kr

K. R. Park
Division of Electronics and Electrical Engineering,
Dongguk University, Seoul, South Korea
e-mail: parkgr@dongguk.edu

1 Introduction

The content based query method has been suggested the supplemental way of tag based query method. Query by Singing/Humming (QbSH) among contents based query methods is the thing we are interested. QbSH has evolved from note based matching method to frame based method. Note based matching method reached limitation because it is hard to extract notes from the polyphonic music precisely. To overcome this problem, Up-Down-Repeat (UDR) method is proposed from many researches but it is not also outstanding. Frame based method is used recently. So we are also using this way in this article. We propose the fusion matching method with three matchers as Dynamic Time Warp (DTW), Linear Scaling (LS) and Quantized Binary Code (QBCode) [1, 5]. Melody sequence extracted from polyphonic music is reference vector and pitch sequence from humming signal is test vector in this method. The overall system is described briefly as follows: At the client side, it records user humming signal from input device during 12 s and then judges that the input signal is tainted with background noise. If it does, it suppresses background noise, and then estimates pitch from this signal and finally it queries estimated pitch sequence to the server after formatting it with MPEG Query Format (MP-QF) international standard. The server parses queried data and measures similarity between queried data and reference data, and then recommends top 20 candidates with highest score to the client. We have three steps to develop this algorithm. At the beginning point, we have built up this algorithm with Roger Jang's corpus that is one of datasets for QbSH task of Music Information Retrieval Exchange (MIREX) 2005 [3, 4]. This dataset has 2,898 manuscript humming pitch vectors represented by semitone at every 32 ms. 48 Musical Instrument Digital Interface (MIDI) sequences are contained in this dataset. At the next stage, we have improved this algorithm with MIDI dataset which contains 1,200 humming clips corresponding to 100 K-pop songs. At the final stage, we have optimized matching algorithm with polyphonic music dataset having 2,000 K-pop MP3 s from various genres.

2 Proposed Matching Algorithm

The matching algorithm takes two vector sequences as test and reference pattern. In this paper test pattern is the pitch sequence estimated from humming signal. Reference pattern is the melody contour extracted from 2,000 Korean popular songs. Pitch sequence is estimated by algorithm that is based on time-frequency domain autocorrelation method. In order to get rid of pitch doubling and halving problem two domain autocorrelation function is used in this algorithm. To build up reference database the harmonic structure based vocal melody extraction method is used too. Figure 1 depicts block diagram of this method. It is starting with eliminating the silent duration on pitch and melody contour because it does not

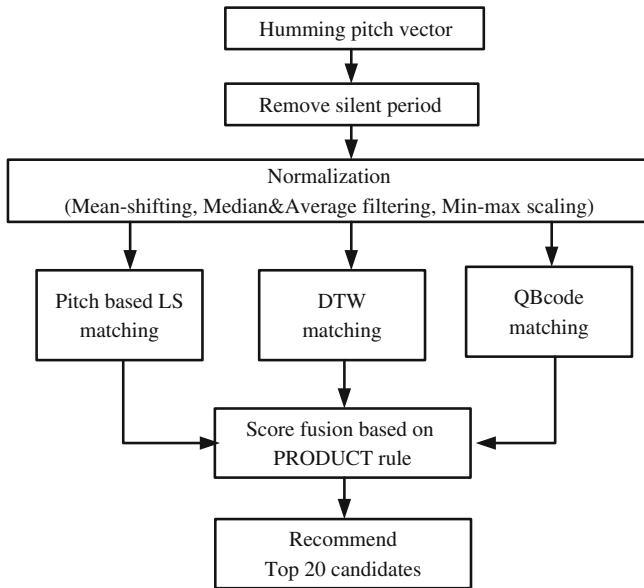


Fig. 1 Matching engine flow gram

have any information and it can be possible to reduce computational complex. Then it normalizes two patterns using Mean-shift, Median & Average filter and Min-max scaling. Mean-shift filter adjusts level of humming signal to the level of reference because humming signal might be located at higher or lower rather than original level of music. Median & Average filter with 5-tap is adopted to remove the shot noise and over shoot caused by surround noise, shivering or vibration of vocal tone. Min-max scaling is applied to compensate the gap of amplitude between two vector sequences [6]. Similarity is measured by three matchers simultaneously after normalizing two vectors. Scores from three matchers are combined with weighting factor into single fusion score. The matching engine measures similarity between test and reference pattern with three matchers. It recommends top 20 candidates having higher scores.

2.1 Dynamic Time Warping

The advanced DTW is the main matcher of three of them. DTW is one of Dynamic Programming (DP) that measures distance between two patterns with different length. DTW has several important constraints like start-and-end point constraint that must align start and end point between two vectors, local region constraint that must grow one by one grid, and three more constraints. It is also sliding matching

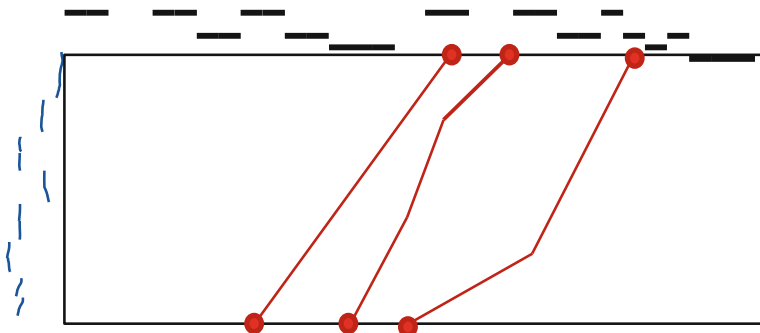
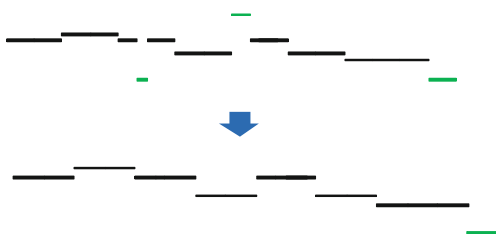


Fig. 2 Proposed advanced DTW

Fig. 3 Chroma representation of reference data



method with fixed window and hopping size. We figure out that the block window sliding method causes the critical problem against QbSH because it does not guarantee that people would be humming at specific point of music (Fig. 2).

To overcome this problem advanced DTW is proposed in this paper. It does not have any constraints and fixed window size for sliding matching. So it is possible to start or end at any point and any path. There are two more things that we have considered. The first one is way of measuring similarity between two patterns. Euclidian distance which is linear metric is very useful tool in order to calculate distance. We have figured out that there are a lot of pairs of having nearly same distance in case of using linear metric. Linear metric is not suitable for these cases. We introduce the log scale metric instead of Euclidian distance. Here is an example. There are two pairs having same distance of 5 using linear metric, $\vec{a} - \vec{b} = [1, -1, 1, -1, 1]$ and $\vec{a} - \vec{b} = [0, 0, 0, 0, 5]$. Which pair you might choose? We want to pick up the pair that has more same elements rather than similar elements between two vectors. So we use log metric instead of linear metric. Another one is chroma representation of reference data. Chroma is gathering pitch values of overall octaves into one octave. So it can solve the pitch doubling and halving problem more easily (Fig. 3).

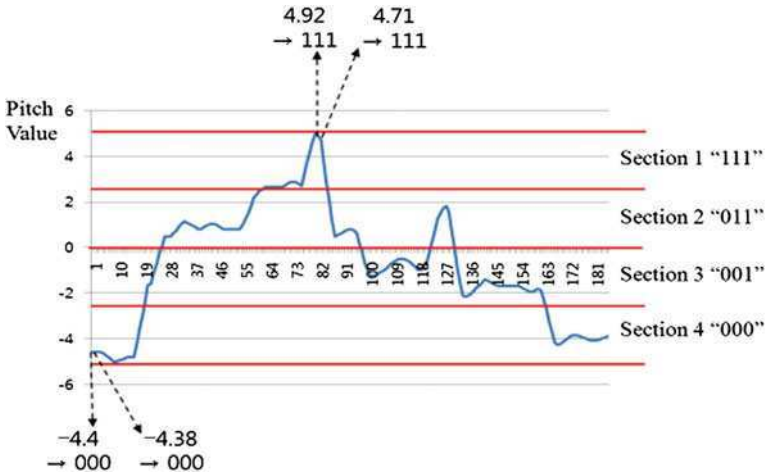


Fig. 4 Quantized Binary code

2.2 Quantized Binary code

Another matcher is the Quantized Binary code called QBcode. It has the 4 section of normalized vector and different binary codes are assigned to each section as ‘000’, ‘001’, ‘011’ and ‘111’. Every pitch value is reassigned to each section code they are located. Before this reassignment, input pitch values are normalized from -6 to 6. The similarity is calculated with Hamming Distance (HD) as shown in Eq. 1.

$$HD = \frac{\|BPA \otimes BPB\|}{T} \tag{1}$$

Where, BPA and BPB represent the extracted QB codes of test and reference patterns respectively, and \otimes mean the Boolean Exclusive-OR operator between corresponding pairs of two QB codes. And T denotes the total number of the QB codes of the vectors. By using the HD, the processing speed is fast compared to the use of other kinds of distances, such as the Euclidean distance [5] (Fig. 4).

2.3 Pitch Based Linear Scaling

LS algorithm is the simplest and quite effective one to match two patterns having different length each other. The main idea is compressing or expanding input data along time axis with several different lengths. LS is very suitable against QbSH method because length of humming signal depends on who is humming.

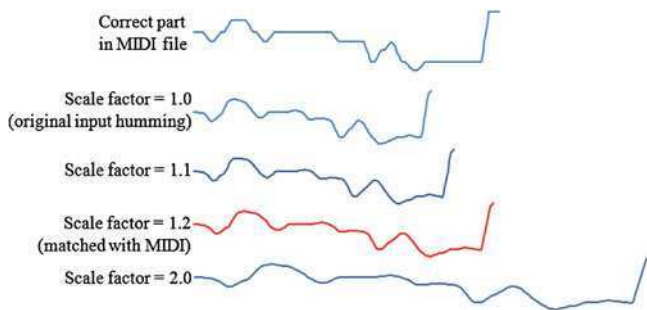


Fig. 5 An example of linear scaling matching

So, humming data should be compressed or stretched before measuring similarity. Test vector is expanded by scale factor from $\times 1.0$ to $\times 2.0$ with 5 steps. The distance is calculated with the log metric for the same reason of DTW (Fig. 5).

2.4 Score Level Fusion

The three scores from the above different matching algorithms are merged into the one fusion score. There are many methods for score level fusion as MIN rule, MAX rule, SUM rule and so on. The fusion score is calculated with the PRODUCT rule which multiply two scores. Basically, Proposed DTW carries out the most important role on the matching stage, and LS and QBCode is complement for DTW. So it gives the weight as 0.5, 0.2 and 0.3 to DTW, LS and QBcode respectively. The matching engine recommends top 20 candidates with higher fusion scores.

3 Dataset

The reference dataset contains vector sequence and segmentation from 2,000 MP3. Humming test set is consisted of 1,200 humming vector sequences against 100 songs as called AFA100 which is among 2,000 songs referred to MNet music chart that is the most popular one of Korean music portal services. We also have 2,000 MIDI data from KARAOKE system to verify our matching algorithm. We include this MIDI data into our system for Korean KAROKE service at implementation phase. The music dataset covers 7 different genres with ballad, dance, children song, carol, R&B (Rhythm and Blues), rock, trot and well-known American pop. The 1,200 humming clips with 12 s duration are recorded against AFA100 to evaluate the algorithms because it is hard to hum at every time of testing the

Table 1 Evaluation of matching engine

	Top1 (%)	Top10 (%)	Top20 (%)	MRR	Time (s)
32 ms	74.90	89.20	92.20	0.793	12.4
64 ms	71.10	80.10	83.70	0.738	4.7

algorithms. It is consisted with almost same ratio of sing and humming and recorded from 29 persons. Three among them have the experience of music related study at university and others not. We analysis and classify that into 3 groups as beginning, climax part and others. We figure out that beginning part is a slight over 60% and climax part is about 30%. It did not expect that the beginning part is almost twice of climax. We evaluate the performance with this humming set.

4 Conclusion

We evaluate the matching algorithm with MRR method which was widely used in the MIREX contest with the recorded 1,200 humming clips [2]. We have the two input steps as 32 and 64 ms. The evaluation condition is as followed: 1,200 humming clips for test vector, 2,000 polyphonic songs with from 3 to 6 min duration for reference vector on Intel i7 973 with 8 MB memory. Table 1 shows the performance of the proposed algorithm.

References

1. Orio N (2006) Music information retrieval: a tutorial and review. *Found Trends Inf Retr* 1:1–90
2. Ghias A et al (1995) Query by humming-musical information retrieval in an audio database. In: *Proceedings of ACM Multimedia*, pp 231–236
3. Roger Jang's corpus, <http://neural.cs.nthu.edu.tw/jang2/dataSet/childSong4public/QBSh-corpus/>
4. Jang JSR, Lee HR (2008) A General framework of progressive filtering and its application to query by singing/humming. *IEEE Trans Speech, Audio Language* 2(16):250–258
5. Downie JS (2008) The music information retrieval evaluation exchange (2005–2007): a window into music information retrieval research. *Acoust Sci Technol* 29(4):247–255

Towards an Autonomous Indoor Vehicle: Utilizing a Vision-Based Approach to Navigation in an Indoor Environment

Edward Mattison and Kanad Ghose

Abstract We introduce a vision system based approach to autonomous navigation and mapping in an indoor environment. Our goal is to eventually create a common processing system that can direct both aerial and ground vehicles. This paper presents our initial results for controlling a ground vehicle as well as demonstrates the potential of a vision-based system controlling a completely airborne solution or a solution that combines aerial vehicles with ground vehicles. The autonomous ground explorer (AGE) described here serves as a test vehicle to validate the algorithms that will be integrated into the control of our aerial vehicle that is under development. The ground explorer uses only vision (three basic web cameras) to efficiently move within building corridors. The ground explorer navigates hallways and makes directional decisions based on the following processes: tracking the corridor's visual vanishing point, anticipating intersections using odometry and fiduciary markers, identifying and classifying intersections based on vanishing points and fiduciary markers, and completing controlled turns based on a confidence factor in its location on a calculated route within a simplified topological map.

Keywords Autonomous ground aerial vehicle · Indoor navigation · Robotic vision

E. Mattison (✉) · K. Ghose
Computer Science Department, State University of New York,
Binghamton, NY 13902, USA
e-mail: Emattis1@binghamton.edu

1 Introduction and Related Work

There is significant value in having ground vehicles that can navigate autonomously in an indoor environment for many applications such as search and rescue operations in an indoor environment, environmental sampling and testing in an indoor hazardous material situation, routine building surveillance or monitoring, and indoor military reconnaissance in a hostile environment. This paper presents the design and implementation of such an autonomous ground vehicle *system*, where the main sensory mechanism used for navigation and vehicle steering is vision-based. Our vehicle processes the images captured by on-board cameras to determine vehicle heading and to steer adaptively within the corridors of the building. The use of a vision-based navigation system eliminates the need for a separate set of sensors for *indirectly* sensing the physical proximity of the vehicle from the walls of the corridors, to detect crossings and to negotiate the turns and crossings—all using a single image processing system. Our main motivations for using such a vision-based autonomous navigation system is driven the need to have a solution that works equally well for small form factor ground vehicles as well as for indoor aerial vehicles. Both are limited in their payload and energy supply capacities. Our system uses three on-board wireless cameras that capture images and send them down on a wireless link to a base station, which is either a stationary platform or a larger ground vehicle that follows the smaller form factor vehicle(s). The base station analyzes the captured images and processes them to derive the appropriate steering and control signals, which are sent upstream to the autonomous vehicles on a radio link to control their movement. The ground vehicles controlled in this manner are not autonomous by themselves, as the processing necessary for their steering and control, although automated, are done off-board. However, the system comprising of the ground vehicle and the off-board processing system essentially provides the autonomous navigation capabilities. We demonstrate the basic principles of our system using a large form factor vehicle and large cameras. It is worth noting that the same system can be implemented with currently available smaller form factor wireless cameras and remotely controlled smaller ground and aerial vehicles.

The area of autonomous vehicle navigation has been a fertile research area for many years. While the ideal autonomous vehicle navigation system is yet to be perfected, several types of ground vehicles have reasonable proficiency at navigating autonomously. For outdoor applications, global positioning systems (GPS) based autonomous navigation systems exist [MicroPilot Autopilot at micropilot.com] for both ground and aerial platforms to move autonomously from one waypoint to the next. For indoor applications, simultaneous location and mapping (SLAM) [1] has been the dominant technique used for autonomous navigation for many ground vehicles. Many others, such as [2, 3] have refined SLAM-based autonomous navigation techniques for ground vehicles. SLAM-based techniques are very accurate but a limitation has been their reliance on very expensive, heavy and high power consuming solutions, including the use of laser-scanning devices. Furthermore,

these systems are very complex and require a very complex and computationally intensive system to process sensory data and derive the vehicle control signals. We believe that this type of data collection is excessive for basic indoor vehicular navigation and particularly so for indoor navigation systems using small form factor ground vehicles and airborne indoor vehicles. Our goal was to develop a uniquely simple system, with low computational complexity, low weight, low cost, and low power consumption that could navigate accurately in an indoor environment.

Our solution relies on decoupling the system components across two platforms the sensory (image capturing) and steering systems are located on the autonomously steered and remotely controlled vehicle, while the sensor data processing (image processing), autonomous navigation system and the derivation of the signals for remote control are implemented on a stationary platform (or a heavier ground vehicle). This arrangement makes it possible for us to make the autonomous steering system and vehicle control capabilities as sophisticated a necessary, without burdening the actual small form factor vehicle in terms of weight, form factor or power requirements. Furthermore, using only one type of sensor (namely, image capture) to reduce the number of on-board sensors and thus simplify the interfaces on the mobile platform. The image processing system permits us to maintain vehicle bearings/heading, avoiding the walls of the corridors; it also enables us to negotiate intersections and turns. Taken together, the overall system architecture permits us to autonomously steer small form factor ground vehicles as well as reasonably lightweight aerial platforms. There are several examples of airborne vehicles that have multiple on-board cameras, have stable flight characteristics, make use of a wireless link to send down captured images to a base device, and receive directional steering commands from the base [AR.Drone at ardrone.parrot.com, CyberQuad at cybertechuav.com.au].

With such targets vehicles in mind in the long run, we have implemented a prototype vision-based fully autonomous steering system for a *ground vehicle*. This paper describes that implementation. We are in the process of porting a similar approach for our prototype air vehicles. The main objective behind this paper is to show that a completely vision-based navigation and steering system is feasible for autonomous indoor navigation and explorations. All processing is relegated to a laptop computer that is part of the ground vehicle. For the air vehicles, the processing device could either be on a following ground vehicle or at a static base station. It uses the same processing and control algorithms that will be eventually used for the airborne vehicle. We will now describe the details of the prototype ground system and its assessment in the rest of this paper.

2 Autonomous Ground Vehicle Components

Our autonomous ground explorer prototype is based on the ER1 robotics platform [Evolution Robotics at evolution.com]. The ER1 is composed of an extruded aluminum frame, three wheels, two stepping motors, and a control module.

The ER1 can carry a standard laptop. This permits the relatively small robot local access to significant processing power. The ER1 has a three-wheeled configuration, with two wheels powered by extremely accurate stepping motors and the third wheel acting as a swivel point. The vehicle has three cameras, one looking forward and two looking sideways.

Our system is based on image analysis and a control program written in Java. The control program provides the following functions: A mapping module represents the map in terms of a topological graph and visually presents a real time map with continuously updated vehicle location. A routing module determines the shortest path from current location to destination. At each intersection, the module determines the vehicle status from the following three choices: destination reached, intermediate point along route, or the vehicle is lost. A movement module guides the vehicle as directly as possible from one intersection to the next. A Java API acts as an interface to issue movement commands to the ER1 robotic platform. A Java API acts as an interface with RoboRealm image analysis scripts. The RoboRealm image processing pipeline software is an integral part of our navigation system. There are three instances of RoboRealm running, one for each camera. Many functions are programmed within RoboRealm, such as visual vanishing point, corner detection for intersection identification, fiduciary image matching for landmark detection. The final piece of software is the JGraph Java graphing package. We use this package to represent graph data and to provide visual representation of real time map.

3 Topological Map

Our mapping system is a simple topological graph made in Java with the JGraph package API. The system displays the current graph and updates the current vehicle position in real time. Several papers have incorporated topological map into their navigation systems. The nodes represent the intersections and the edges represent the hallways or corridors. Additionally, landmarks as well as the vehicle itself are also represented as nodes on the graph. However, they are not connected to any edges, they maintain their position in the graph based on their X and Y coordinates. Ranganathan [4] and Filliat [5] both utilized a topological map approach, but included many more connected nodes on their graphs, to include landmarks and regular interval sensed data points. Roberts [6] represented the intersections of underground mines as the nodes and the curving underground mine pathways as the topological graph edges. The simplistic nature of his approach led me to translate his technique to the indoor building environment.

4 Vision-Based Approach to Navigation

Since we target a ground vehicle in this effort, we are essentially unconstrained in terms of weight, battery capacity, processing power, and sensor availability. However, our ultimate goal is to validate our system and replicate it on our hybrid

airship. The airship will be very constrained in terms of weight, battery capacity, and processing power, so we chose to constrain our ground vehicle. This led us to a single sensor approach: vision-based navigation. Our system is composed of algorithms contained within system modules. At this point in time, our system includes modules that control mapping, routing, distance measuring, vanishing point detection and following, intersection detection and identification, and controlled turning. The control algorithm continuously repeats a relatively simple sequence of events. I want to highlight several of the process and explain our approach.

4.1 Vanishing Point Identification and Following

Vanishing point identification and following is a critical process in our navigation. The forward camera vanishing point is used to direct the ground vehicle's forward movement. Side camera vanishing points are used in the detection and classification of intersections. Our approach is simple. We capture video from the forward facing webcam. Then we perform an analysis on the images. Our image processing consists of a canny edge detection algorithm and a line convergence algorithm to determine the vanishing point of the current scene (see Fig. 1). The camera has a field of view of 640 pixels wide by 480 pixels high, the focal point of the camera is coordinate (320,240). If we are pointing directly towards the hallway vanishing point, the vanishing point coordinate would have an x-axis value close to 320. If our vehicle were to drift left or right during its travel down a corridor, we correct this using the vanishing point. If we drift to the right, the x-axis coordinate of the vanishing will decrease. We do not want azimuth corrections to cause constant oscillation, so we do not make any corrections until the vehicle is greater than 10% off from center.

4.2 Anticipate, Identify and Classify Intersections

The ground vehicle navigation system relies heavily on its ability to anticipate, identify, and classify intersections. These intersections represent the nodes on our topological map (graph). It is critical that our system recognizes when the vehicle is located at a graph node with as close to absolute certainty as possible. This is required because major directional decisions are only made at intersections. Our system uses four pieces of information to increase the level of confidence that the vehicle is located at an intersection: odometry data, corner detection using image matching, and left and right vanishing points, and laser distance measurements.

We can anticipate intersections using odometry and a corner detection algorithm. The stepper motors on the ground vehicle are very accurate. We use the odometry data from the motors to help us anticipate upcoming intersections based on known

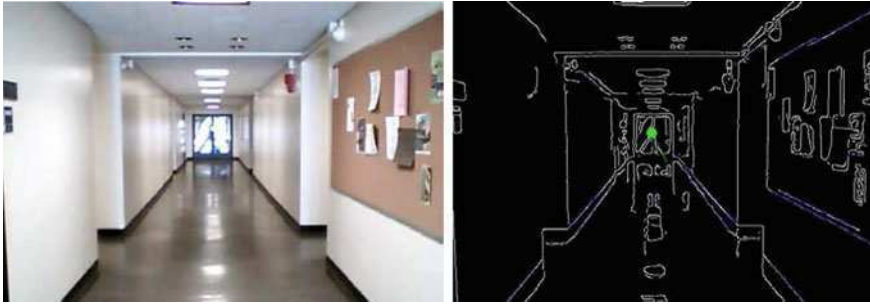


Fig. 1 Raw hallway image and image after edge and vanishing point detection

map distances. Odometry error is minimized by resetting all odometry data as we reach each intersection. We use an image analysis technique to identify corners as we approach them. We process forward images with a canny edge detection (already being used in the vanishing point calculation) and then we image match certain features that repeatedly appear in specific areas of the image as we approach intersections. The odometry and corner detection together provide us great early warning of approaching intersections. As we approach an impending intersection, we use the data collected to classify the intersection. Intersections can be classified as one of seven types based on the edges that are connected to that node of the graph: four-way intersection, three-way left, three-way right, T-intersection, left-turn, right-turn, and dead end. As an example, if we are entering a “T” intersection, I would expect the following information to be collected: Left distance goes to infinity; Left vanishing point appears; Right distance goes to infinity; Right vanishing point appears; Front vanishing point disappears. The use of three independently calculated vanishing points (front, left, and right) proves to be very accurate at identifying and classifying intersections on our ground vehicle. In ideal building conditions, we have a 97% identification rate for known intersections. We continue to refine our approach to provide the highest possible confidence factor to our routing module, which is tasked with directional decisions.

4.3 Fiducial Markers as Landmarks

In a future version of our navigation system, we plan to add the capability to recognize as well as capture unconstrained landmarks from the environment. However, in this iteration of the system, we have added limited landmark recognition by placing fiducial markers around the environment at strategic locations. Fiducial detection and recognition differs from that of generic object recognition as some assumptions are made on the type of object being detected.

For example, a black and white fiducial can be detected without needing color, and can be easily separated from the background due to its high contrast nature. Because Fiducials are planar objects that have very distinct corners and shapes, they can be placed on the floor, ceiling or walls and be detected correctly without any recalibration of the camera. For this reason, we chose to use them in our navigation and localization applications.

5 Experimental Results

The ground explorer has proven to be a capable platform to demonstrate and validate our vision-based navigation algorithms. As a whole, the current system can calculate a route, safely navigate down the center of a corridor, identify and classify intersections, make directional decisions, and efficiently navigate from point A to point B on a known map with greater than 90% accuracy.

The system has been tested many times and experimental results are very consistent. Based on a test run we calculate several data points to track our progress. We identify the planned location and the actual location approximately every 2 m, which allows us to calculate the average distance off planned path (ADOPP) and the maximum distance off planned path (MDOPP). In the example test run provided (see Fig. 2), the average distance from the planned path was 0.22 m and the maximum deviation from the planned path was 0.57 m. These distances are well within our expected operating range for the vehicle and were anticipated for several reasons. First, the planned path is shown as the exact middle of a hallway. However, as soon as the ground explorer identifies an anticipated turn, it initiates the turning procedure. The experimental results show that the vehicle always turns before the “planned” path then slowly corrects itself to the center of the hallway. Secondly, in order to reduce oscillation, the vehicle does not correct its azimuth until its vanishing point is more than 10% off center. Therefore, the ground explorer is performing as expected in terms of actual path. Small adjustments could be made to our algorithms to reduce this deviation if desired.

However, the system is not without its problems. The ground explorer currently functions well only in a controlled indoor environment. The building must have good lighting characteristics for the vanishing points to be reliably calculated. The system is greatly aided when the wall and floor colors have a stark contrast as well. If these conditions are not met, the accuracy of the system suffers greatly. Additionally, even in ideal conditions, corner detection using image matching (without fiducial markers) has been successful only 65% of the time. Using fiducial markers, intersection anticipation improves to 78% (fiducial placed at edge of corner to be viewed with front-facing camera) and 95% (fiducial placed on corridor wall before a corner to be viewed with side-facing cameras) respectively. Therefore, most intersection anticipation is determined from odometry data and the use of fiducials. We identify, verify, and classify the intersections using the

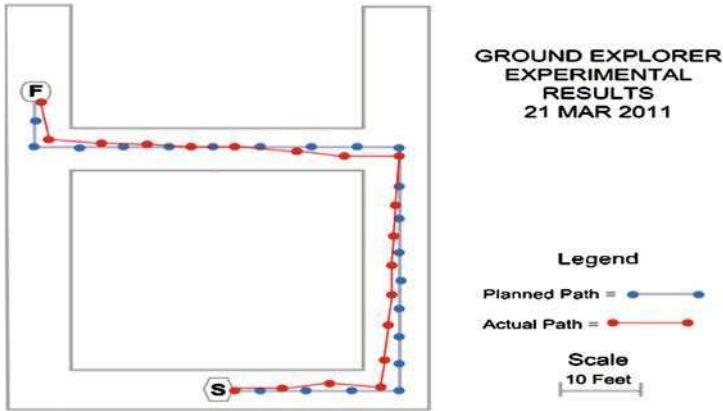


Fig. 2 Results depicting planned path versus actual path

front and side-looking camera vanishing points, but we do not want to rely so heavily on odometry data and fiducial markers for anticipating corners.

6 Aerial Vehicle Integration

As we progress towards our goal of creating an Adaptive Indoor Aerial/Ground Navigation and Mapping System, we have begun to migrate and test many of the ground vehicle algorithms to our aerial test platform. We spent a year experimenting with and building several aircraft. These designs ranged from a blimp, to a helicopter, to a hybrid airship, and a quadcopter. However, the emergence of an extremely low cost, stable quadcopter in the form of the AR.Drone has led us to abandon our airship design aspirations and focus entirely on our navigation algorithms. This aerial vehicle allows us to move slowly, in a controlled manner, within a building hallway structure. We have the capability to avoid obstacles, but we can also make mistakes without a catastrophic failure occurring. No current system has demonstrated significant indoor autonomy in the air. Our goal is to create such a system. Our approach to an autonomous indoor aerial navigation system has two main components. The hardware centers on the quadcopter, while the software is focused on our vision-based autonomous navigation algorithms. The project is largely a programming and integration endeavor, using many existing components. The pieces have been integrated and are controlled by a Java application, which serves as the main system controller. Images are sent from the airship to the base and control commands are sent back to the airship via wireless network. We have integrated the aerial platform to the point that it can take-off and land autonomously, maintain a level flight within building corridors, and navigate hallways by following a vanishing point.

7 Future Work

We have demonstrated the effectiveness of our vision-based navigation system on our ground vehicle. Our follow-on actions encompass three segments. First and foremost, we plan to translate and adapt many of our algorithms from the ground vehicle to the aerial vehicle, as discussed in the previous section. Secondly, we plan to augment the ground vehicle with additional capabilities such as adding the ability to capture landmark images in lieu of fiduciary markers, the capability to map unknown areas, the ability to avoid obstacles within hallways, dynamic map changes and route re-calculation, and incorporating RFID technology into the navigation system. The third step in our process will be to add coordination mechanisms between the ground and aerial vehicle. The ground and aerial vehicles have different strengths and weaknesses, so depending on the challenge encountered by the system, one vehicle may be better suited to accomplish a given task. Coordination between the vehicles will allow cooperative decisions and task assignment based on predetermined criteria.

References

1. Smith R, Cheeseman P (1986) Estimating uncertain spatial relationships in robotics. SRI International Press, Menlo Park, pp 1–26
2. Durrant-Whyte H, Baily T (2006) Simultaneous localization and mapping: part I. *IEEE Robot Autom Mag* 13:99–108
3. Thurn S, Wolfam B, Fox D (2005) Simultaneous localization and mapping, Chapter 10. In: *Probabilistic robotics*, MIT Press, Cambridge, pp 309–330
4. Ranganathan P, Hayet JB (2002) Topological navigation and qualitative localization for indoor environments using multi-sensory perception. *Robot Auton Syst* 41:137–144
5. Filliat D, Meyer J (2003) Map-based navigation in mobile robots: a review of localization strategies. *Cogn Syst Res* 4:243–282
6. Roberts J, Duff E, Corke P (2003) Reactive navigation and opportunistic localization for autonomous underground mining vehicles. *Info Sci Int J* 145:127–146
7. Ryu B, Yang H (1999) Integration of reactive behaviors and enhanced topological maps for robust mobile robot navigation. *IEEE Trans Syst Man Cybern Part A Syst Hum* 29(5):474–485
8. Tovar B, Gomez L, Murrieta-Cid R (2006) Planning exploration strategies for simultaneous localization and mapping. *Robot Auton Syst* 54:314–331
9. So A, Chan W (2002) LAN-based building maintenance and surveillance robot. *Autom Constr* 11:619–627
10. Rekleitis I, Meger D, Dudek G (2006) Simultaneous planning, localization, and mapping in a camera sensor network. *Robot Auton Syst* 54:921–932

Artificial Pheromone Potential Field Built by Interacting Between Mobile Agents and RFID Tags

Piljae Kim and Daisuke Kurabayashi

Abstract In this study, the concept of the chemical substance pheromone is utilized for the robotic tasks. This paper first illustrates the model of pheromone-based potential field. The field is constructed through the interaction between mobile robots and data carriers, such as RFID tags. The stability analysis of the pheromone potential field is carried out also aiming at the implementation on a real robotic environment. The comprehensive analysis on stability provides the criteria for how the parameters are to be set for the proper potential field, and has led to a new filter design scheme called pheromone filter, which satisfies both the stability and accuracy of the field. The unique structures of both the revised mobile robot and the designed filter show that the proposed method facilitates a more straightforward and practical implementation.

Keywords Pheromone potential field · Mobile robots · RFID tags · Stability analysis · Pheromone filter

1 Introduction

It has been known for some time that social insects such as ants and bees communicate with each other through a process which is generally called stigmergy, and perform given tasks effectively by using the chemical substance

P. Kim (✉) · D. Kurabayashi
Department of Mechanical and Control Engineering,
Tokyo Institute of Technology, 2-12-1 Ookayama,
Meguro-ku, Tokyo 152-8552, Japan
e-mail: pjkim@irs.ctrl.titech.ac.jp

D. Kurabayashi
e-mail: dkura@irs.ctrl.titech.ac.jp

pheromone [1]. Inspired from these biological characteristics, researchers have been recently motivated to undertake studies on pheromone-based robotics [2, 3]. In addition to this, there are other studies that make use of the radio-frequency identification (RFID) technology [4] for realizing digital or artificial pheromones [5, 6], wherein the agents communicate with other agents by updating a pheromone trail through the RFID tags distributed in an environment. In particular, interests in RFID technology for navigation of the mobile robot have been currently growing. For instance, Vorst et al. focused on simultaneous localization and mapping (SLAM) techniques that map static tags' locations [7], and Kodaka et al. [8] tried to build a navigational entropy map using RFID tags distributed on the floor. However, few researches have suggested guidelines on stability when applying RFID tags for these real robotic tasks. This study utilizes the idea of the artificial pheromone. The study emphasizes that the stability analysis is fundamental not only to secure the simple implementation, but also to improve the scheme in both stability and accuracy.

2 Modeling Based on Pheromone Deployment

In this section, we outline the model and present the real platform that is being developed. Let us first introduce the ant colony as the most popular biological model that makes entire use of the pheromone. Figure 1a shows the shortcut-producing process observed from the black garden ant (*Lasius niger*) colony. When a colony is offered a food source, a scouting ant discovers the source and returns to the nest, laying a pheromone trail which dissipates over time and distance. Since the trail is reinforced, the shortcut between the food source and the nest is formed by the strengthened pheromone trail. Likewise, if there are sufficient number of robots and data storing devices such as RFID tags, the robots may also form and follow the shortest path between home and goal as shown in Fig. 1b. How, then, can the numbers of the robots and RFID tags be determined so that the system works properly? Also, what kinds of functions are required for each agent? This study has been launched in order to provide an answer for these questions.

In the framework of this study, the system illustrated in Fig. 1b can be realized via the interaction between mobile agents with RFID transceivers and RFID tags, on which the digital equivalent of pheromones is laid by the agents. For mobile agents, we use revised e-puck robots that have RFID readers and writers on their forward and backward side, as shown in Fig. 2a. The passive RFID tags, which are the white rectangular tags in Fig. 2a, are adopted for this study in the same way as our previous work [6]. The RFID tags are wireless and battery-free, and each tag is marked with a unique identifier and is equipped with a small memory that allows it to store data. The data consists of the tags' own IDs and scalar pheromone values. The passive nature of the RFID tags implies that pheromone can be diffused only via the interaction between robots and RFID tags, which is illustrated in Fig. 2b.

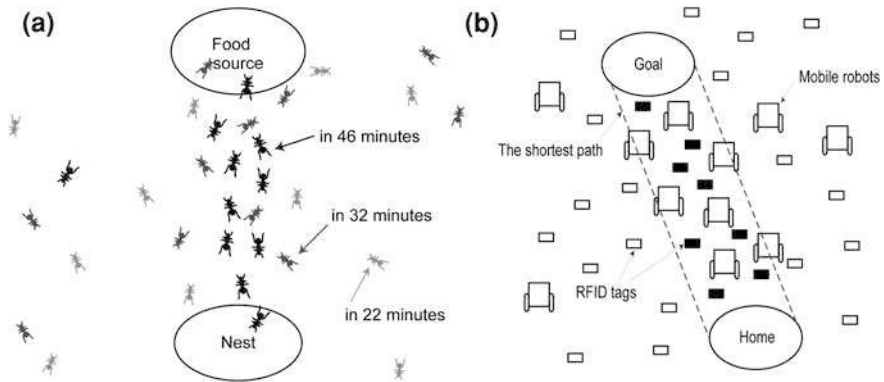
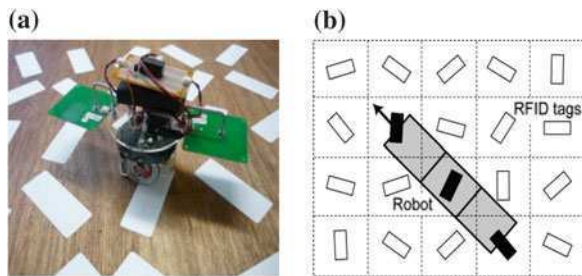


Fig. 1 The shortcut on which pheromone is accumulated in the *Lasius niger* colony (adapted from Camazine et al. [9]) and its realization in robotic platform. **a** The shortcut in ant colony, **b** the shortest path in robotic platform

Fig. 2 The revised robot and the illustration of the interaction in the experimental field. **a** Revised e-puck with RFID reader and writer and distributed RFID tags, **b** concept of the interaction between a mobile robot and RFID tags



3 Stability of Pheromone Potential Field

The mathematical pheromone model in one-dimensional space x is

$$\frac{\partial u(t, x)}{\partial t} = D \frac{\partial^2 u(t, x)}{\partial x^2} - Ku(t, x), \tag{1}$$

where $u(t, x)$ denotes the pheromone density, and the coefficients D and K represent diffusion and evaporation rate respectively [10]. We can easily extend the same idea to the two-dimensional space without losing generality. This paper, therefore, focuses on the one-dimensional case for reasons of simplicity. Equation 1 can be discretized as following difference equation through using the forward time central space (FTCS) scheme

$$\begin{aligned} & \frac{1}{\Delta t} \{u(t_n + \Delta t, x_i) - u(t_n, x_i)\} \\ &= \frac{D}{(\Delta x)^2} \{u(t_n, x_i - \Delta x) - 2u(t_n, x_i) + u(t_n, x_i + \Delta x)\} - Ku(t_n, x_i), \end{aligned} \tag{2}$$

where we calculate the time derivative in a forward manner and calculate the space derivative in a centered manner. The numerical stability is estimated with relation to the growth or decrease of the rounding error in the calculation scheme of the finite difference method. In this study, we consider the perturbation stability analysis [11], which is, in our opinion, the simplest and most straightforward.

Let us slightly simplify the equations by using the notation that temporal indices are represented by a superscript and spatial indices are represented by a subscript, such that the value of the function U at the time t_n and at the point x_i is expressed as U_i^n , i.e., $u(t_n, x_i) = U_i^n$. The pheromone equation using this notation becomes

$$\frac{1}{\Delta t} \{U_i^{n+1} - U_i^n\} = \frac{D}{(\Delta x)^2} \{U_{i-1}^n - 2U_i^n + U_{i+1}^n\} - KU_i^n. \tag{3}$$

If we add the perturbation ε_i^n to around U_i^n , the difference equation is written

$$\frac{1}{\Delta t} \{U_i^{n+1} - (U_i^n + \varepsilon_i^n)\} = \frac{D}{(\Delta x)^2} \{U_{i-1}^n - 2(U_i^n + \varepsilon_i^n) + U_{i+1}^n\} - K(U_i^n + \varepsilon_i^n). \tag{4}$$

Having rearranged the equation, we get

$$U_i^{n+1} = \underbrace{(1 - K\Delta t)U_i^n + \frac{D\Delta t}{(\Delta x)^2} \{U_{i-1}^n - 2U_i^n + U_{i+1}^n\}}_{\hat{U}_i^{n+1}} + \underbrace{\varepsilon_i^n \left\{ 1 - \frac{2D\Delta t}{(\Delta x)^2} - K\Delta t \right\}}_{\varepsilon_i^{n+1}}, \tag{5}$$

where \hat{U}_i^{n+1} represents the unperturbed U . For the stable behavior of the system, the perturbation should be decreased without overshoot, which means the condition is

$$0 \leq \left| \frac{\varepsilon_i^{n+1}}{\varepsilon_i^n} \right| < 1. \tag{6}$$

From Eqs. 5 and 6, the stability condition without overshoot leads to

$$\left(\frac{D}{(\Delta x)^2} + \frac{K}{2} \right) \Delta t \leq \frac{1}{2}. \tag{7}$$

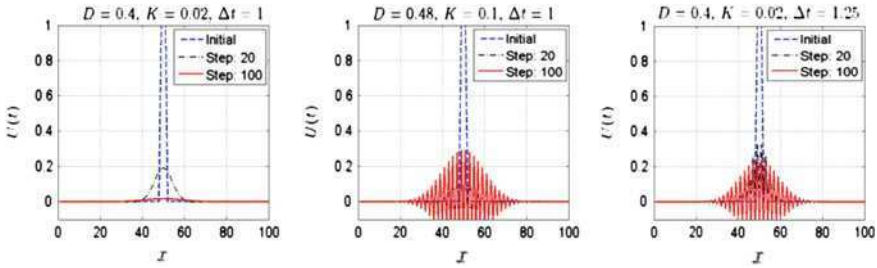


Fig. 3 The outputs of the pheromone equation for varying parameters. **a** Stable case, **b** unstable for large K , **c** unstable for large Δt

From the derived stability condition, it is noted that all varying parameters, i.e., $K, D, \Delta t$ and Δx , are coupled together for the stable solution, thus we performed a numerical simulation in a leave-one-out cross validation way. Throughout the simulations presented in this paper, we assign constant boundary conditions at the left-hand and right-hand edges, i.e., the Dirichlet boundary conditions are imposed.

Figure 3 shows the outputs of the pheromone equation. Unstable behaviors are observed for the larger K and Δt . It is obvious that Δx and Δt are critical coefficients, because updating the frequency of the potential field depends mainly on the number of RFID tags and the speed of the robots. Unfortunately, however, it is not possible to precisely configure these parameters in advance because of the highly-coupled properties of each parameter. From this point on, we use $D = 0.4, K = 0.02, \Delta t = 1$ and $\Delta x = 1$ as typical stable parameters.

4 Stable Solution for the Pheromone Model

When applying a pheromone model to the real robotic system, as we have examined, all parameters consisting of stability condition need to be carefully designed. To relax these restrictions, we have noticed the fact that the implicit time stepping can improve or even eliminate stability limitations, which suggests that combining a backward scheme in time with a central difference approximation in space, i.e., the so-called backward time central space (BTCS) scheme, may make a pheromone potential field more stable. The BTCS scheme can be written

$$\frac{1}{\Delta t} \{U_i^{n+1} - U_i^n\} = \frac{D}{(\Delta x)^2} \{U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}\} - KU_i^{n+1}. \tag{8}$$

If we define diffusion and evaporation number as $d = D\Delta t/(\Delta x)^2$ and $k = -\Delta tK$, respectively, the equation is written

$$\{U_i^{n+1} - U_i^n\} = d\{U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}\} - kU_i^{n+1}. \tag{9}$$

Note here that the Fourier coefficient of the solution of Eq. 9 can be written

$$U_i^n = V^n e^{j\omega_x(i\Delta x)}, \tag{10}$$

where $j = \sqrt{-1}$, and V^n is the amplitude of the n th harmonic. If we set phase angle $\theta = \omega_x \Delta x$, U_i^n and $U_{i\pm 1}^{n+1}$ can be described as the following Fourier series.

$$U_i^n = V^n e^{ji\theta}, \quad U_{i\pm 1}^{n+1} = V^{n+1} e^{j(i\pm 1)\theta} \tag{11}$$

Let us substitute Eq. 11 into Eq. 9 and divide by $e^{ji\theta}$, then

$$V^{n+1} - V^n = dV^{n+1}(e^{j\theta} + e^{-j\theta} - 2) - kV^{n+1} \tag{12}$$

is derived. Using the Euler’s formula, and if we define an amplification factor G , the equation is rearranged as

$$V^{n+1} = \frac{1}{(1 + 2d(1 - \cos \theta) + k)} V^n = GV^n, \tag{13}$$

thus the condition

$$|G| = \left| \frac{1}{(1 + 2d(1 - \cos \theta) + k)} \right| \leq 1 \tag{14}$$

must apply for the converging solution. This inequality is called the von Neumann stability condition. From the fact that $d > 0$, $(1 - \cos\theta) \geq 0$ and $k > 0$, the condition (14) is always the case. This means that the BTCS scheme is unconditionally stable for any parameters. If we define diffusion matrix \mathbf{D}_m , and represent Eq. 9 in matrix form, the scheme is written as the following simplified form.

$$(\mathbf{I} - d\mathbf{D}_m + k\mathbf{I})\mathbf{U}^{n+1} = \mathbf{U}^n \tag{15}$$

Note that the rank of $(\mathbf{I} - d\mathbf{D}_m + k\mathbf{I})$ is never reduced, hence we are definitely able to get a solution for the pheromone equation. We also anticipate that the computational cost is not that expensive, because $(\mathbf{I} - d\mathbf{D}_m + k\mathbf{I})$ forms a tri-diagonal matrix. However, when it comes to the implementation of Eq. 15, we encounter with a crucial difficulty, due to the fact that the inverse of a tri-diagonal matrix is no longer tri-diagonal. That is, most of the elements are required in order to retrieve the temporal and spatial information from an inverse matrix. On the basis of these considerations, the Crank–Nicolson method was deemed to be the most appropriate solution for this study. It is second-order and implicit in time, and is numerically stable [12]. The method was proposed mainly to improve the accuracy of the BTCS scheme, but we are, rather, interested in the structure produced by the method. It has the averaged form of the FTCS and BTCS schemes at n and $n + 1$, respectively as following.

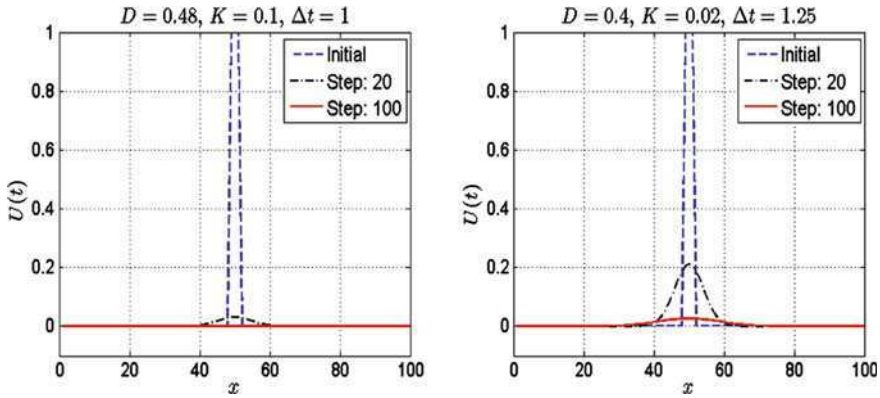


Fig. 4 The outputs of the pheromone equation for varying parameters using the Crank–Nicolson scheme. **a** Stable for large K , **b** stable for large Δt

$$\frac{1}{\Delta t} (U_i^{n+1} - U_i^n) = \frac{D}{2} \left(\frac{U_{i-1}^n - 2U_i^n + U_{i+1}^n}{(\Delta x)^2} + \frac{U_{i-1}^{n+1} - 2U_i^{n+1} + U_{i+1}^{n+1}}{(\Delta x)^2} \right) - \frac{K}{2} (U_i^n + U_i^{n+1}) \tag{16}$$

The von Neumann stability condition is obtained as

$$V^{n+1} = \frac{(1 - 2d(1 - \cos \theta) - k)}{(1 + 2d(1 - \cos \theta) + k)} V^n = GV^n \tag{17}$$

using the same technique that was used for the BTCS scheme. From the above condition, it is obvious that the value $|G|$ will not go higher than 1, and thus the Crank–Nicolson scheme is unconditionally stable for all varying parameters. The stability is validated in Fig. 4. Let us finally represent the method as a matrix form in order to clarify the mathematical structure.

$$\mathbf{U}^{n+1} = (\mathbf{I} - d\mathbf{D}_m + k\mathbf{I})^{-1}(\mathbf{I} + d\mathbf{D}_m - k\mathbf{I})^{-1}\mathbf{U}^n = \mathbf{W}\mathbf{U}^n \tag{18}$$

It is worthwhile to mention here that the system (18) is stable when $\|\mathbf{W}\| < 1$, where $\|\mathbf{W}\|$ represents Euclidean norm.

5 Pheromone Filter

Having recognized the result of the stability analysis, the useful design scheme for a smoothing filter is shaped in this section. We call a developed filter pheromone filter.

Fig. 5 The calculated state transition matrix for an initial delta function: the matrix was derived from the Crank–Nicolson scheme with parameters $\Delta t = 1$, $\Delta x = 1$, $D = 0.4$ and $K = 0.02$

$$\begin{pmatrix} 0.44 & 0.21 & 0.03 & 0.00 & \dots & 0.00 \\ 0.21 & 0.47 & 0.22 & 0.03 & \dots & \vdots \\ 0.03 & 0.22 & 0.47 & 0.22 & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & 0.03 & 0.22 & 0.47 & 0.21 \\ 0.00 & \dots & 0.00 & 0.03 & 0.21 & 0.44 \end{pmatrix}$$

To design a filter, we first observed a generated matrix \mathbf{W} from Eq. 18 in detail, which is generally called the state transition matrix. The calculated matrix is presented in Fig. 5, which was sampled after running ten steps for an initial delta spike under the one of the typical conditions. In the matrix, interestingly enough, each row looks similar to the one-dimensional Gaussian kernel. Recall that our revised robot has the similar structure to the one-dimensional filter, which suggests that the pheromone deployment by mobile robots can be modeled straightforwardly from the state transition matrix. As a first trial, we directly applied the following representative 1×3 kernel to the pheromone model, which is called raw filter.

$$\boxed{\beta \quad \alpha \quad \beta} \Leftrightarrow \boxed{0.22 \quad 0.47 \quad 0.22} \tag{19}$$

Figure 6a shows the enlarged shape of the produced potential field when the filter is applied. The figure was sampled at time step 20 and magnified in order to be precisely observed. To reduce the differences with other methods, we carried out the approximating operation as shown in the following equation,

$$\hat{\mathbf{W}} = \begin{pmatrix} \alpha_1 & \beta_1 & \dots & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \dots \\ \dots & \ddots & \ddots & \ddots \\ 0 & \dots & \beta_{m-1} & \alpha_m \end{pmatrix} \approx \begin{pmatrix} \alpha_1 & \beta_1 & \dots & 0 \\ \beta_2 & \alpha_2 & \beta_2 & \dots \\ \dots & \ddots & \ddots & \ddots \\ 0 & \dots & \beta_m & \alpha_m \end{pmatrix} \tag{20}$$

$$\begin{pmatrix} 0.44 & 0.21 & 0.03 & 0.00 & \dots & 0.00 \\ 0.21 & 0.47 & 0.22 & 0.03 & \dots & \vdots \\ 0.03 & 0.22 & 0.47 & 0.22 & \dots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & 0.03 & 0.22 & 0.47 & 0.21 \\ 0.00 & \dots & 0.00 & 0.03 & 0.21 & 0.44 \end{pmatrix}$$

and adjusted the kernel value of 1×3 filter while ensuring that the total sum of the filter is equal to or less than one. As a result, the following pheromone filter was finally designed.

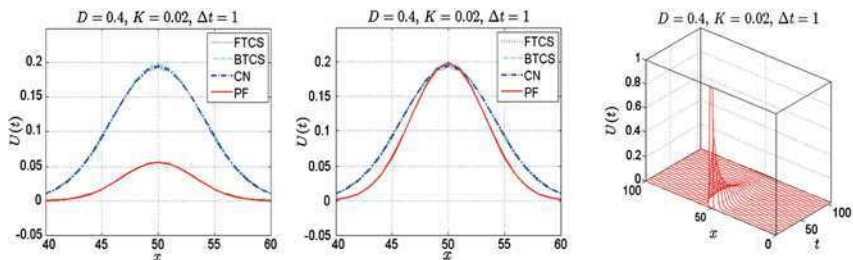


Fig. 6 Pheromone potential field generated using the 1×3 raw filter (for **a**) and modified 1×3 pheromone filter (for **b** and **c**). **a** Raw filter, **b** modified filter, **c** temporal propagation

Table 1 Standard deviation for given steps: PFR and PFM represent pheromone filter raw (Eq. 19) and pheromone filter modified (Eq. 21), respectively

Method	Stdev. for steps ($\times 100$)				
	10	20	30	40	50
FTCS	7.14	4.85	3.53	2.64	2.02
PFR	3.89	1.27	0.44	0.16	0.06
PFM	7.32	4.54	3.00	2.04	1.41

$$\boxed{\hat{\beta} \quad \hat{\alpha} \quad \hat{\beta}} \Leftrightarrow \boxed{0.24 \quad 0.49 \quad 0.24} \tag{21}$$

Using this finally designed filter, the initial pheromone of delta spike was deployed in order to observe the propagation of the potential field. Figure 6b shows this result, where the shape agrees reasonably well with the other well-known numerical solvers. Figure 6c finally represents the temporal propagation when the pheromone filter is applied. The pheromone is propagated in a stable and gradual way. In the figure, CN and PF represent the Crank–Nicolson and pheromone filter, respectively.

From these results, the designed filter seems to provide an efficient way of producing a stable potential field. It should be noted that each scalar value of the pheromone filter is virtually identical to each element of our model shown in Fig. 2, which implies that the revised mobile robot could play the role of the diffusion filter of the pheromone through the RFID tags distributed in an environment. To evaluate the proposed algorithm more quantitatively, the standard deviation of each potential field is compared in Table 1. From the table, it is again observed that the proposed pheromone filter (PFM) works in nearly the same way as the FTCS scheme.

6 Conclusions

This paper has formulated the stability condition for the pheromone potential field. The established criterion can provide a general guideline for researchers in the relevant field. Based on the result of the stability analysis, we further presented a new methodology of making a smoothing kernel, called pheromone filter. The developed method demonstrated stable and accurate performance through the numerical simulation; the stability is guaranteed from the implicit structure and the high-level accuracy was achieved by modification of the kernel elements with observing the output behavior. It is expected that the proposed scheme could provide a practical technique for designing a filtering system wherein the stability of the system is secured. Despite these advantages, the proposed filter has yet to be evaluated in a batch manner. We are developing a sequential filtering algorithm, which is required in order to identify and update the kernel value online. We are now planning to implement the presented method on real robots.

References

1. Wilson E, Holldobler B (1990) *The ants*. Springer, Heidelberg
2. Payton D, Daily M, Estowski R, Howard M, Lee C (2001) Pheromone robotics. *Auton Robots* 11:319–324
3. Parunak HVD, Brueckner SA, Sauter J (2007) Pervasive pheromone-based interaction with RFID tags. *ACM Trans Auton Adapt Syst* 2(2):1–28
4. Want R (2006) An introduction to RFID technology. *IEEE Pervasive Comput* 5(1):25–33
5. Mamei M, Zambonelli F (2007) Pervasive pheromone-based interaction with RFID tags. *ACM Trans Auton Adapt Syst* 2(2):1–28
6. Herianto H, Kurabayashi D (2009) Realization of an artificial pheromone system in random data carriers using RFID tags for autonomous navigation. In: *Proceedings of IEEE international conference on robotics and automation*, pp 2288–2293
7. Vorst P, Schneegans S, Yang B, Zell A (2008) Self-localization with RFID snapshots in densely tagged environments. In: *Proceedings of IEEE/RSJ international conference on intelligent robots and systems*, pp 1353–1358
8. Kodaka K, Niwa H, Sugano S (2009) Active localization of a robot on a lattice of RFID tags by using an entropy map. In: *Proceedings of IEEE international conference on robotics and automation*, pp 1193–1199
9. Camazine S, Deneubourg J-L, Franks NR, Sneyd J, Theraulaz G, Bonabeau E (2001) *Self-organization in biological systems*. Princeton University Press, Princeton
10. Sugawara K, Kazama T, Watanabe T (2004) Foraging behavior of interacting robots with virtual pheromone. In: *Proceedings of IEEE/RSJ international conference on intelligent robots and systems*, pp 3074–3079
11. Roache PJ (1998) *Fundamentals of computational fluid dynamics*. Hermosa Publishers, Albuquerque
12. Crank J, Nicolson P (1996) A practical method for numerical evaluation of solutions of partial differential equations of the heat conduction type. *Adv Comput Math* 6:207–226

Proposed Network Coding for Wireless Multimedia Sensor Network (WMSN)

A. A. Shahidan, N. Faisal, Nor-Syahidatul N. Ismail
and Farizah Yunus

Abstract Wireless sensor network has caused a lot of interest in many applications especially in real-time multimedia data transfer over wireless network which is known as wireless multimedia sensor network (WMSN). Even though such application is quite complex and very challenging to be realized, the availability of CMOS camera and microphones with low cost, low power and small size has reduce the complexity of developing the WMSN. Apart from this, there are several subjects that may also contribute toward the realization of WMSN such as routing protocol, source coding and channel coding. This research is focusing on network coding which is one of the most important research interests since a few years ago due to its ability to increase the throughput and reduce energy consumption of the wireless network system. Many ideas have been proposed involving several network coding schemes based on applications in certain wireless network standard. This project is aimed for multimedia application in IEEE 802.15.4 wireless sensor network. Therefore, the constraint of this standard must be taken into consideration in the propose network coding scheme.

Keywords Network coding · Wireless multimedia sensor network · Transmission delay

A. A. Shahidan (✉) · N. Faisal · N.-S. N. Ismail · F. Yunus
UTM-MIMOS Center of Excellent, Faculty of Electrical Engineering,
Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia
e-mail: ashahidan@fke.utm.my

1 Introduction

The research on wireless sensor network (WSN) technology has begun since a few years ago. Research communities are developing the technology for many applications such as environmental monitoring, hazardous environment exploration, and military tracking [1]. Wireless sensor network consists of many self-organized sensing nodes [2] where every single node has the capability of data monitoring and collecting, data processing and data sharing among the adjacent nodes.

The usage of WSN technology has evolved rapidly from simple application which involves small information size such as water level and temperature to the more complex one which is involving multimedia data such as video and sound. As for multimedia data application, real time response is one of the demanding issues in the research of WSN technologies. Therefore, one of the aims in the research on WSN is to find the possible ways to ensure reliability during data transmission so that higher volume of information can be transferred through the network and the responses of the system is approaching real-time characteristic.

This is to compromise with WSN constraint such as miniature, low cost and low power consumption. Due to such constraint, IEEE 802.15.4 compliant radio is chosen to be widely used in WSN [3]. The maximum achievable rate is up to only 250 kbps [4]. Therefore, only small size of data is usually transmitted. The WSN data rate is considered too slow for multimedia and hence the data should be compressed or coded in such a way that the transmission speed can be compromised. Even though the transmission delay is one of the main issues for multimedia in WSN, the processing delay is a more significant issue.

In WSN, the network layer and transport layer play an important role in achieving data transmission reliability. The network layer offers a best effort service and the transport layer is responsible for achieving reliable, provide congestion control and flow control [5]. At transport layer, both of reliability and congestion control algorithm should be taken into consideration to ensure that the data reach at the destination successfully and achieving high reliability data delivery as proposed in [6]. At network layer, reliability also needs to be considered as an important requirement for data transfer.

Thus, we consider the problem of reducing the energy consumption and increasing the throughput of the network. This can be handled by a coding method known as network coding which has become interesting in the research of WSN in the last few years [7]. In this paper, we investigate the benefit of the network coding in reducing the number of transmissions over the network that operated in erasure channel model. Performance of the network coding used for the purpose of multimedia data transfer is analyzed using simulation experiment.

The remainder of this paper is organized as follows. Related works in wireless multimedia network are reviewed and summarized in [Sect. 2](#). Network coding concept is described in [Sect. 3](#). The details of system model, result and discussion are described in [Sects. 4](#) and [5](#) respectively. Lastly, [Sect. 6](#) presents the conclusion and recommendation for future works.

2 Related Works

There are several works that apply network coding for various applications in wireless sensor networks. In [8], the authors used raptor codes like network coding for multimedia streaming. The throughput of the networks is increased using multicast routing algorithm with network coding as introduced in [9].

Meanwhile, network coding is also realized using fountain approach in order to have low-complexity characteristic networks as described in [10]. The work is meant for data collection in wireless sensor networks which is done through simulation experimented. In [11], the authors used hexagonal lattice topology in wireless sensor network to achieve the maximum possible energy benefit using network coding. This scheme is applied for multiple unicast networks.

In this project, network coding will be applied in multimedia application. There are several research that focus on this application in wireless network such as in [12–14]. In Yakubu Suleiman Baguda et al. [12], introduce cross layer design to transmit H.264 video standard over IEEE 802.11e wireless local area network (WLAN). They take the advantages of EDCA mechanism in IEEE 802.11e to prioritize the frame.

In WSN, the feasibility to transmit multimedia application at low rate and low power using IEEE 802.15.4 has already been proven in [13]. The author used COTS CMOS Camera together with TelG mote platform to transmit JPEG image. Thus, this work has inspiring the author in [14] to propose the transmission of MPEG-4 in the same medium. The priority frames are applied at application layer to improve the quality of video transmission. However, all of this research does not applied network coding for their multimedia transmission.

3 Network Coding Concept

Network coding has become a new research area in the field of information theory which is used in practical networking system [7]. The main assumption of the network coding is; instead of just forwarding the data, the node can combine several received packet to form a new packet or several new packets through certain mechanism [7].

The following is the example of a simple network coding scenario. Let X and Y be two nodes that are located separately and cannot make a direct connection between each other. Let S be an intermediate node between X and Y where the data will be passing through. The connection of these three nodes is illustrated in Fig. 1. Assume that X and Y want to exchange data a and data b respectively through node S . Based on traditional method, four transmissions are required in order to exchange both data. The process is illustrated in Fig. 2. Transmission $T1$ is for sending data a from node X to node S followed by the transmission $T2$ for transmitting data b from node Y to node S . The transmission $T3$ is to send data

Fig. 1 Node *S* as an intermediate node for node *X* and *Y*

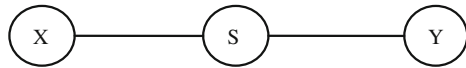


Fig. 2 Four transmissions are required for exchanging data between *X* and *Y*

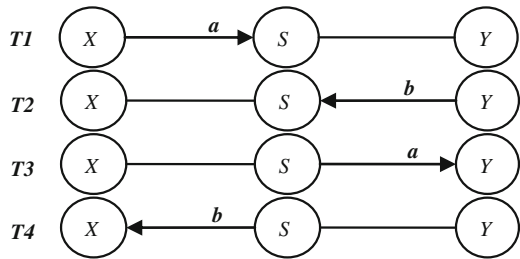
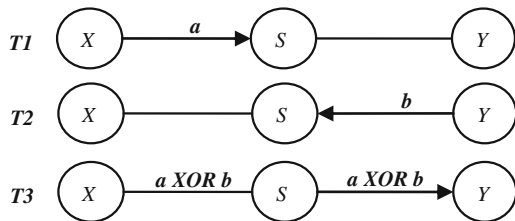


Fig. 3 Three transmissions are sufficient for exchanging data between *X* and *Y*



a from node *S* to node *Y* and follow by transmission *T4* which is for passing data *b* from node *S* to node *X*.

By using network coding on the other hand, both data *a* and *b* will be combined in node *S* through X-or operation to form $a \text{ XOR } b$. Then the combined data will be broadcasted and both node *X* and *Y* will receive the combined data as shown in Fig. 3. The received data will be processed and the particular data will be collected by each node for further processing while the unused data will be dropped. The combined data is transmitted during the third transmission. Here, only three transmissions are required.

Besides, we also know that the nodes in WSN are ubiquitous [1]. Therefore, we can exploit this characteristic to enhance the reliability of the system through cooperative data processing. In short, the sensed information data will be encoded and distributed to the selected adjacent nodes in the sensor field through certain mechanism. The node which received the data will do a simple processing before redistributing the received data to the selected nodes. The process will be repeated until the data reach the destination node. Due to the distribution of the data in the sensor field, it is expected that the data passes through several paths in the sensor field before reaching the destination. As a result, the system becomes more reliable in reducing the number of packet loss.

4 System Model

We use network model as shown in Fig. 4 which consists of three parts which are source, relay and destination. The source part consists of a source node that provides the information data through the sensor attached to it. The data from source node will be passed to the relay part. Relay part consists of several relay nodes forming a certain number of hops. Users can vary the number of nodes to ensure that the distance between two adjacent nodes are in the transmission range. The ubiquity of the system is increased as the number of node increases and we can exploit this characteristic to increase the data transmission reliability. In our case however, we consider various number of relay nodes. Up to fifteen nodes with maximum six-hops have been deployed to simplify the simulation. The destination part consists of a sink node.

The algorithm is aimed to be implemented for image sequence data transfer in multimedia data transmission. The following are the description of the data transmission process in our simulation. The information data will be fragmented into a few small packets. The size of the fragmented packets is obtained from the maximum size of packet that can be handled by the physical layer of the sensor node. For example, the users of TelG mote [15] need to ensure that the packet size must not be greater than 128 bytes since the sensor node use Xbee wireless module that can only handle the packet up to 128 bytes of size.

In every transmission, the source node will broadcast a packet to all relay nodes in the first hop. After receiving a packet, the relay node will buffer the packet and followed by network coding process which is done with probability P_{NC} . The process is done by combining the received packet with the packet that is previously stored in the buffer. The same activities occur in every relay nodes until the packet reach the destination node intact. Network coding must not be done in every transmission in order to ensure that the degree-one packets reach the destination. The reason is that the decoder used is based on LT-Codes decoder where a certain number of degree-one packets are required in order to reduce the ripple size or the degree of the received packets so that the original packets can be recovered [16].

5 Result and Discussion

In our simulation, we operate the system in erasure channel where the packet with error will be dropped. For performance evaluation, we have also introduced a few parameters such as the maximum hop, H and the number of nodes per hop, L . The maximum degree of the packet, D is obtained from the equation (1).

$$D = (1 + B)^{(H-1)} \quad (1)$$

The number of packets after fragmentation is K . The probability of network coding being done at a node is P_{NC} , while the probability of packet loss is P_E .

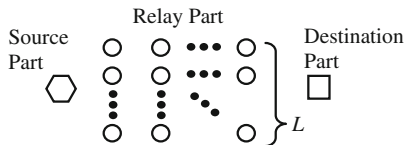


Fig. 4 System model for simulation

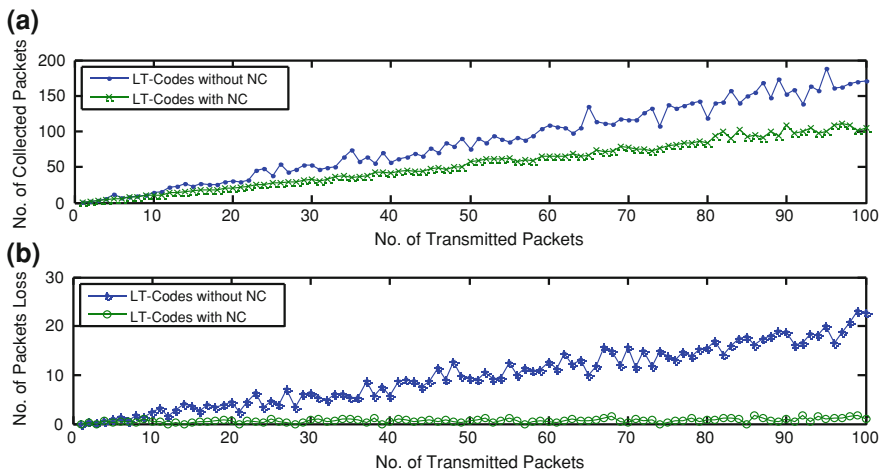


Fig. 5 a Number of packets collected to recover all data without any packet loss b Number of packets loss for equivalent number of collected packets and transmitted packets

Every node has a buffer of size B . The simulation is conducted using the network model in Fig. 4. We set the value of $H = 4$; $B = 1$ and hence the value of D will be 8. Since this algorithm is still immature, we will not compare it with other established algorithm. The data used for the simulation is computer generated data. We repeat the simulation with value of K from 1 to 100 in three different conditions.

In the first condition, we want to find out the effect of network coding on the number of packets required for data recovery. The receiver is allowed to collect the packets larger than K so that all data can be recovered without any losses. Secondly, we would like to observe the effect of network coding on the number of packets loss limited number of collected packets. In this condition, the receiver is allowed to collect the packets up to the value of K .

For the first and second condition, we are comparing the result between the network using network coding and the network without network coding. Figure 5 shows the result obtained from the simulation using $P_{NC} = 0.5$ and $P_E = 0.2$. We can see from Fig. 5a that the network with network coding has lower number of collected packets in order to recover overall data compare to the one without

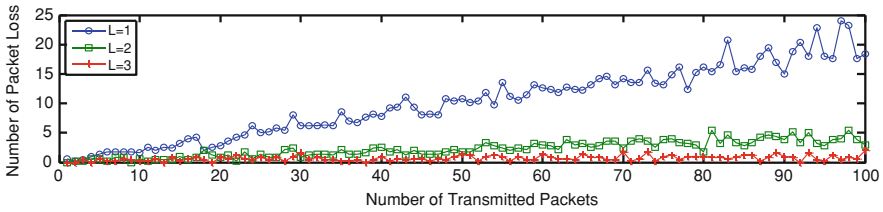


Fig. 6 Number of packet loss network with NC with various numbers of nodes per hop

network coding. Meanwhile in Fig. 5b, the number of packet loss for the network with network coding is always lower than the one without network coding. It is corresponded to the benefit of network coding itself which is to reduce the number of transmissions in the network and hence will reduce the transmission delay. In addition, the reducing of transmission delay is important for multimedia data transfer because of sensor nodes are battery powered. Based on traditional method, only one packet is transmitted at an instance of time. The transmitter will continuously transmit the packet without knowing whether the packet sent is received by the receiver or it is being dropped. Due to the merging of several packets into one packet in network coding, a lesser number of transmissions are sufficient to overcome packet loss.

In the last condition, the value of L is varied from one to three in order to study the effect of different number of nodes per hop on transmission reliability. We limit the number of the collected packets up to the value of K . The number of packets loss for every single value of L is obtained and illustrated in Fig. 6. As the number of nodes per hop increases, the number of packets loss decreases. We know that the increasing number of nodes will result in the improvement of the network ubiquity. Hence, we have exploited this characteristic to enhance the reliability of data transmission and reduce the number of packets loss.

6 Conclusion and Future Work

In recent years, network coding has received considerable attention due to its ability to increase the throughput and reduce energy consumption in wireless network system. In this paper, we proposed a network model that applied network coding in wireless multimedia sensor network. Through packet combination method, the system become more reliable and reduces the number of packet loss. The assumptions made for the simulation are described while the result obtained shows that the network coding is functioning according to the theory.

In order to improve this network model, there are lots of works to be done. Adaptive routing protocol can be implemented as a future work to enhance the multimedia transmission in this works. The adaptive routing protocol such as in [18]

that used biological inspired to find a route and in [19] that take care of parameters in physical layer to choose optimal forwarding decision suitable to be implementing in this work.

Acknowledgments The Author would like to thank to the Ministry of Science, Technology and Innovation (MOSTI) Malaysia for sponsorship, UTM-MIMOS Center of Excellent for their full support and good advice and for Research Management Center (RMC) Universiti Teknologi Malaysia. Thanks also to all anonymous reviewers for their invaluable comments and the guest editors who handle the review of this paper.

References

1. Ian FA, Tommaso M, Kaushik RC (2006) A survey on wireless multimedia sensor network. Broadband and Wireless Networking Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta
2. Wang-Yan (2008) Study on model and architecture of self-organization wireless sensor network. *Wireless Communications, Networking and Mobile Computing*
3. García-Hernández CF, Ibarguengoytia-González PH, Hernández JG, Pérez-Díaz JA (2007) Wireless sensor networks and applications: a survey. *Int J Comput Sci Network Secur* 7(3)
4. Sinem Coleri Ergen (2004) ZigBee/IEEE 802.15.4 Summary, 10 Sep
5. Karl H, Willig A (2007) Protocol architecture for wireless sensor networks. Wiley InterScience, New York, p 360
6. Yunus F, Ismail NN, Ariffin SHS, Shahidan AA, Norsheila F, Yusof SKS (2011) Proposed transport protocol for reliable data transfer in wireless sensor network. In: International conference on modelling, simulation and applied optimization (ICMSAO2011)
7. Fragouli C, Boudec JYL, Widmer J (2006) Network coding: an instant primer. *ACM SIGCOMM Comput Commun Rev* 36(1)
8. Thomos N, Frossard P (2007) Raptor network video coding. In: Proceedings of the first ACM international workshop on mobile video (in conjunction with ACM Multimedia 2007), Augsburg, Germany
9. Tan Chong, Zou Junni (2007) A multicast algorithm based on network coding for wireless sensor network. *IET Conf Pub* 2007:1043
10. Vukobratović D, Stefanović Č, Crnojević V (2008) On low-complexity network coding for data collection in wireless sensor networks. In: Sixteenth Telecommunications forum TELFOR
11. Goseling J, Weber JH (2008) Energy-benefit of network coding for multiple unicast in wireless networks. In: Proceedings of 29th symposium on information theory in the Benelux, pp 36–40
12. Baguda YS, Faisal N, Yusof SK, Syed SH, Rashid R (2009) Enhancing video quality over IEEE 802.11e WLAN using cross layer design. Springer-Verlag, Berlin
13. Rashid RA, Faisal N, Fikri AH, Halim A (2011) Wireless multimedia sensor network platform for low rate image/video streaming. *J Teknologi* 54(Sians & kej.) Keluaran Khas, Jan 2011
14. Ismail NN, Yunus F, Ariffin SHS, Shahidan AA, Rashid RA, Embong WMAEW, Faisal N, Yusof SKS (2011) MPEG-4 Video transmission using distributed TDMA MAC protocol over IEEE 802.15.4 wireless technology. In: International conference on modelling, simulation and applied optimization (ICMSAO2011)
15. Fikri AH, Hamid BA, Rashid RA, Faisal N (2009) Development of IEEE 802.15.4 based wireless sensor network platform for image transmission. *Int J Eng Sci* 9
16. Luby M (2002) LT-codes. In: Proceedings of ACM symposium foundations computer science, Vancouver, BC, pp 271–280

17. Saleem K, Faisal N, Baharudin MA, Ahmed AA, Hafizah S, Kamilah S (2011) BIOSARP— Bio-inspired self-optimized routing algorithm using ant colony optimization for wireless sensor network—Experimental performance evaluation. In: Mastorakis NE, Demiralp M, Mladenov VM (eds) *Computers and simulation in modern science (vol IV)*” Included in ISI/SCI web of science and web of knowledge, pp 165–175
18. Ahmed AA, Faisal N (2011) Secure real-time routing protocol with load distribution in wireless sensor networks, special issue paper in security and communication networks

Alternative Concept for Geometry Factor of Frequency Reuse in 3GPP LTE Networks

Modar Safir Shbat and Vyacheslav Tuzlukov

Abstract An intelligent radio resource management (RRM) is the core system of long term evolution (LTE) networks to provide the upcoming future applications with the broadband mobility, development of self organizing network (SON), and quality of service (QoS). Third generation partnership project (3GPP) standardizes that every 1 ms the physical radio resource blocks (PRBs) should be rescheduled during the transmission time interval (TTI). This proposal places a lot of processing load in the evolved node B's (eNodeBs). The way to speed up the scheduling process should be considered in addition to increasing the whole LTE network throughput caused by utilization of high order modulation. It requires more processing time on both ends of the transmission process. In this article, important radio resource scheduling aspects are discussed. An alternative concept for geometry factor based on a new feedback method of channel quality information (CQI) can be used to employ the determined frequency reuse policy among the LTE cells. The presented concept helps us to reduce the PRBs scheduling time, processing load, and complexity.

Keywords Long term evolution networks · Radio resource management · Geometry factor · Frequency reuse

M. S. Shbat (✉) · V. Tuzlukov
College of IT Engineering, Electronics Engineering Department,
Kyungpook National University, 1370 Sankyuk-dong,
Buk-gu, Daegu 702-701, South Korea
e-mail: modboss80@knu.ac.kr

V. Tuzlukov
e-mail: tuzlukov@ee.knu.ac.kr

1 Introduction

The present trend toward new generation of wireless broadband networks and the global mobility of user terminals fueled the need of developing the existed communication technologies. The networks and services are supported by a diversity of solutions to fulfill the exploding growth of mobile internet and related services that need more bandwidth and high capabilities in mobility and radio resources management. Recent increase in mobile data usage and demand of new applications have motivated the third generation partnership project (3GPP) to work on the long-term evolution (LTE) as the latest standard in the mobile network technology. Future network generations must be efficiently flexible to support scalability, as well as reconfigurable network elements in order to provide the best resource management solutions in hand under effective network employment with low cost. The ultimate target is to increase the valuable spectrum efficiency using more flexible and effective spectrum allocation and radio scheduling scheme to optimize the QoS, maximize system capacity, and satisfy the self-organizing network (SON) requirements.

Radio resources are scheduled every 1 ms in 3GPP LTE network and different frequency bandwidths and/or aggregated bandwidths can be assigned to an individual user based on the channel condition and availability. Owing to rapidly and instantaneously changing nature of radio channel quality there must be a sufficiently fast scheduling algorithm to compensate the changing channel conditions. Before assigning the modulation technique and coding rate to user equipment (UE) by eNodeB (the base station BS) in the LTE network based on the transmission channel condition, there must be defined the physical radio resource blocks (PRBs). Thus, the problem of scheduling and distribution of the PRBs in 3GPP LTE among users is a complicated process. Speeding up the scheduling process is an important point in the way to achieve the proposed standard of PRBs scheduling time. This article investigates the radio resource management considerations, the frequency reuse, and geometry factor main principles and presents an alternative concept for the geometry factor with feedback method from the UE to the eNodeB in order to reduce the complexity and the scheduling processing load. The rest of this article is organized as follows: the LTE networks RRM considerations are discussed in [Sect. 2](#). [Section 3](#) introduces the geometry factor concept for the frequency reuse employment. The introduced concept general analysis is presented in [Sect. 4](#). The conclusion remarks are discussed in [Sect. 5](#).

2 The LTE Networks RRM Considerations

In 3GPP LTE systems, OFDMA for downlink and SC-FDMA for uplink are accepted as multiple access techniques. Radio resource scheduling is a process in which the resource blocks are distributed among UEs. Before assigning the

modulation technique and coding rate for UE by eNodeB, the last performs scheduling on available physical radio blocks (PRB) and informs the UE about their allocated time/frequency resources and transmission formats to be used by the user. PRBs scheduling is based on UE capability, QoS, fairness, frequency reuse factor, inter cell interference (ICI), and measurement reports from the UE. Due to the rapidly and instantaneously changing nature of radio channel quality there must be a fast scheduling algorithm to compensate the changing channel conditions. Any RRM algorithm for LTE networks should efficiently use the expensive spectrum acquired by providers, and minimize the number of BSs, but maximize the number of users and also leads to interference limited systems. In general, we can summarize all these requirements based on the PRBs scheduling scheme using the following points or factors: the efficient frequency reuse, the optimum power allocation, the inter cell interference control (ICIC), the fairness, QoS, the SON requirements, and the vertical handover (VHO).

None PRBs scheduling algorithms can solve all existing problems associated with the maximum number of users with available transmission services and with the limited and imperfect channel information used by eNodeB, QoS, and fairness problems. The main idea to employ a frequency reuse is to assign the same frequency band in different cells that are usually far from each other to avoid high interference between neighboring cells. We can significantly improve the signal-to-interference-noise ratio (SINR) without using the same frequency band for neighboring cells [1].

Unfortunately, this improvement in SINR causes a reduction in the available spectrum per cell. The system capacity can be estimated using Shannon's formula [2]:

$$TP_k = \frac{BW}{K} \log_2(1 + SINR_k), \quad (1)$$

where k is the reuse factor meaning that only $1/k$ th part of the spectrum can be used by a single cell, BW is the LTE total bandwidth in Hz, $SINR_k$ is the SINR with reuse k . $SINR$ is given by [3]:

$$SINR = \frac{P_r}{P_{intracell} + P_{intercell} + N_0}, \quad (2)$$

where P_r is the received power density from the user, $P_{intracell}$ is the interference that comes from users inside the cell, $P_{intercell}$ is the interference from neighboring cells, and N_0 is the noise power.

3 The Alternative Geometry Factor Concept

In order to have a beneficial frequency reuse, an appropriate tradeoff between the bandwidth and SINR is important to utilize the spectrum by efficient way setting a frequency reuse factor to proper value and to maximize the cell/user throughput.

The frequency reuse factor should be chosen according to intercell interference power that depends on the cell size. Powerful interference favors a high reuse factor and vice versa. In this article, a soft frequency reuse (SFR) is used. This technique consists of splitting the bandwidth into two parts, namely, the full reuse (FR) and partial reuse (PR) parts. The FR part uses the reuse factor equal to 1 and the PR part is allocated to the cell edge-users. This structure allows us to have two level allocation scheme (TLA), where the first level is the cell-level resource allocation (CRA) and the second level is the user-level resource allocation (URA). It means that the cell users are divided into two categories, namely, the cell centre user (CCU) and the cell edge user (CEU). This classification can be done using the geometry factor G :

$$G = \frac{P_{serve}}{N + P_{nonserve}}, \quad (3)$$

where P_{serve} is the total power generated by the connected BS, $P_{nonserve}$ is the total power received from all BSs served as the interference sources, and N is the portion of the power from BSs that can be modeled as AWGN.

SFR is the applying frequency reuse factor (FRF) of 1 for CCUs and FRF of 3 to CEUs [4]. One third of the whole available bandwidth named the major segment can be used by CEUs where the packets should be sent with higher power. CCUs can access the entire physical radio resources with lower transmission power. To realize FRF of three for CEUs, the major segments among directly neighboring cells should be orthogonal (Fig. 1). The power allocation for each type of users can be determined as:

$$P_{CCU} = \frac{SP}{(\alpha - 1)T + S} = \frac{3P}{\alpha + 2}, \quad (4)$$

$$P_{CEU} = \alpha P_{CCU}, \quad (5)$$

where S is the total number of subchannels in LTE system, T is the number of available subchannels for the CEUs, α is the power ratio between the subchannel used by CEU and the subchannel used by CCU, and P is the reference power signifying the uniform transmit power used by each subchannel in a classical reuse-1 system. We can see that when α equals 1, P_{CCU} is equal to P_{CEU} , and the SFR is a reuse-1 system. As $\alpha \rightarrow \infty$, P_{CCU} and P_{CEU} will converge to 0 and $3P$, respectively, and the SFR becomes a reuse-3 system.

The introduced SFR scheme (also called reuse 1/3) has low complexity and good performance for CEUs. Additionally, it has two main drawbacks, namely, the signaling overhead and overall loss of throughput. In the next section, we try to overcome these drawbacks.

Since the channel quality information (CQI) has to be available at BS (eNodeB), the feedback information can be used for partitioning users. Another important topic here is the required number of feedback bits to cover and achieve optimal scenario for LTE system and, additionally, to reduce the signaling

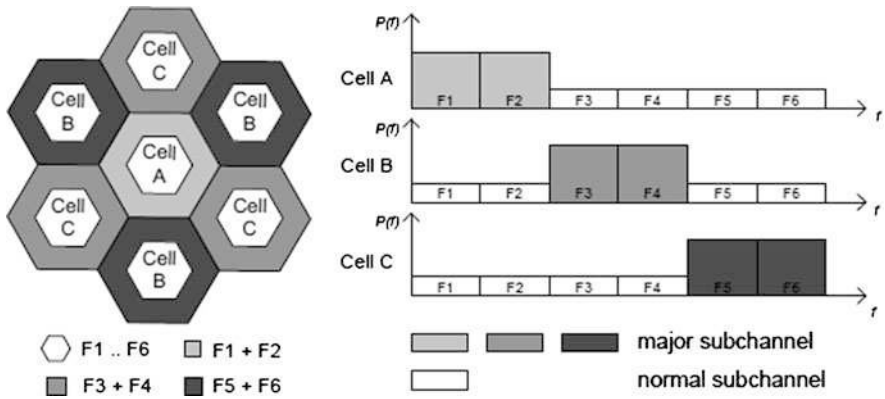


Fig. 1 Concept of the SFR scheme in LTE network

overhead problem. The number of feedback bits is the indicator of the feedback quality and is used by the BS (transmitter) to define the served users from the total number of users sending feedback to the BS. Based on the previous statement, we see that to apply any kind of scheduling scheme there is a need to evaluate the feedback quality and to decide the user should be served or not. In this case, the less the number of the feedback bits the less complexity and the best stability in the scheduling model.

In the proposed solution of this article, each eNodeB receives only one bit from each user instead of full information about SINR (in the case of MIMO system, UE sends information about the SINR for the best beam of every antenna element and this feedback consists of $N_{real} + N_{integer}$ numbers). This bit indicates either SINR of the receive antenna is over a given value (threshold) or not. Now, the transmission by M beams of the BS transmitter is carried out using a single bit of feedback from each user which measures the SINR and compares it with a pre-determined constant threshold δ . The threshold δ is considered as a network parameter known by the BS and all users. The only bit (“0” or “1”), as a feedback from the user, can inform the BS either SINR exceeds the threshold value or not. For pre-introduced LTE system, this threshold can be adjusted to indicate the CCUs at $SINR > \delta$ (“0”), otherwise CEUs. Another scenario is to use two thresholds, δ_1 in the case of “0” feedback bit and CCUs, and δ_2 in the case of “1” feedback bit and CEUs. After receiving the previous simple feedback from all users, the BS schedules a radio resource block or blocks for each user. The presented method is simple and ensures an effectiveness to decrease the signaling complexity of the network. By this way, we can see that the UE helps to replace the geometry factor by using simple feedback to indicate whether it is cell or edge user, and also the alternative concept reduces the processing load in the eNodeBs in the scale of the required time to define the G value [5]. The value of the suggested threshold δ can be the effective SINR that used to obtain transport radio

blocks. The effective SINR can be defined by performing nonlinear averaging of the several available physical radio resource blokes (PRBs) as follows:

$$\delta = SINR_{eff} = -\beta \ln \left(\frac{1}{N} \sum_{i=1}^N e^{-\frac{SINR_i}{\beta}} \right), \quad (6)$$

where N is the total number of sub-carriers to be averaged, and β is calibrated by means of link level simulation to fit the compression function to the additive white Gaussian noise (AWGN) block error ratio.

This model introduces the binary user/resource block assignment variable x that is “1” if the user m obtains resource block r and “0” otherwise. The expected throughput of the user m using the block r depends on the expected SINR. The expected SINR is derived from the latest SINR measurement. Thus, the expected throughput can be presented in the following form:

$$\hat{THR}_{m,r} == \Delta f \log_2(1 + \hat{SINR}_{m,r}), \quad (7)$$

where Δf is the resource block bandwidth. For the QoS criterion, we should take into account the guaranteed bit rate (GBR) as the only criterion under different services. Based on the user’s GBR and CSI, the required number of PRBs for each user can be determined as [6]:

$$N_m = \frac{GBR_m}{M BW_{PRB} S_m}, \quad (8)$$

$$S_m = \log_2(1 + \overline{SNIR}_m), \quad (9)$$

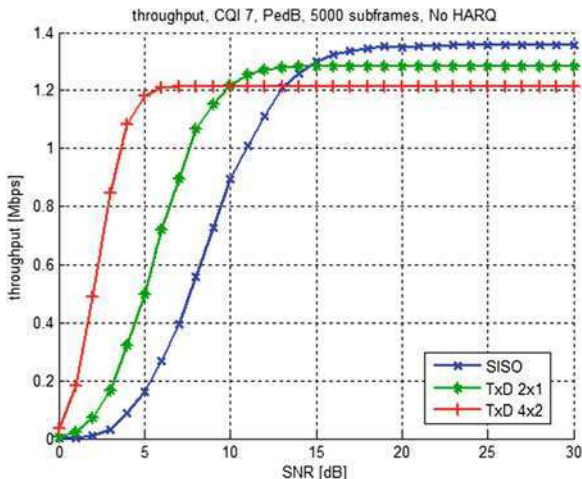
where \overline{SNIR}_m is the average SINR for user m over whole frequency band, S_m is the spectral efficiency of the user m , BW_{PRB} is the bandwidth of PRB, M is the number of OFDM symbols in PRB, and N_m is the required number of PRBs per TTI by the user m . The basic admission control criterion can be presented as the sum of PRBs per TTI required by new user requesting admission and the number of active users in the cell, and should be less than or equal to the total number of PRBs in the LTE system. This admission criterion can be presented in the following form:

$$\sum_{i=1}^k N_m + N_{new} \leq N_{total}. \quad (10)$$

4 General Analysis and Simulation Results

There are three major frequency reuse techniques that can be used in LTE networks to cancel ICI effects, namely, the hard frequency reuse (HFR) with the fixed frequency reuse factor (1 or 3 are popular); the partial frequency reuse (PFR); and

Fig. 2 SFR performance in terms of total cell throughput



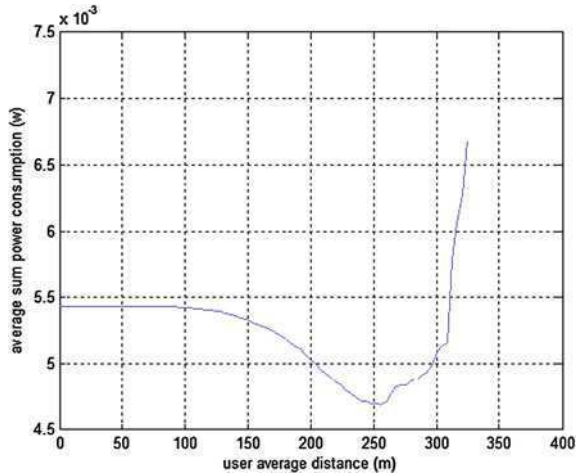
the soft frequency reuse (SFR) that is a part of the discussed technique in this article. A simple LTE system level simulation [7] shows us that the SFR has a good performance in terms of total throughput in the cell (Fig. 2).

It is proposed in the 3GPP LTE networks that every 1 ms the radio resources in the cell should be scheduled. Thus, the way to speed up the scheduling process is very essential and important. SFR processing load is acceptable with good performance, especially for the cell edge users. Another frequency reuse schemes are introduced and some of them may have better performance than SFR, but there are some disadvantages in complexity and high processing load in the eNodeBs. The simulation results for the cell average sum power consumption shows acceptable typical performance for the cell center and cell edge users (Fig. 3).

5 Conclusion

The introduced frequency reuse with the alternative concept for the geometry factor is flexible, and can be employed in different radio resource management, for example, the joint radio resource management (JRRM), the cognitive radio resource management (CRRM), and the dynamic fractional frequency reuse radio resource management scheme. SFR has a good performance, in both the average cell throughput and the cell edge user throughput. The proposed concept works to reduce the signaling overhead and speed up the scheduling process employing the simple feedback method instead of the geometry factor to distinguish the cell-center users (CCUs) and cell edge users (CEUs) between each other. Further improvement can be achieved by applying different PRBs assignment methods in the form of semistatic versions of SFR, which means that the frequency resource

Fig. 3 Average power consumption related to user distance



configuration is adjusted on a time scale corresponding to definite interval, for example, some seconds or longer, that makes the resource partition adaptive to the traffic load variety. This procedure leads to more complicated and higher signaling and proceeding load in the system. The effectiveness from applying the presented soft frequency reuse based on the geometry factor alternative concept has to be confirmed after deep analysis to be assured that we obtain a considerable improvement in the average scheduling delay (increase the scheduling speed) with same or better cell average throughput.

Acknowledgments This research was supported by the Kyungpook National University Research Grand, 2009, and Industry-Academic Cooperation Foundation, Kyungpook National University and SL Light Corporation Joint Research Grant (the Grant No. 201014590000).

References

1. Wang Y, Kumar S, Garcia L, Pedersen KI, Kovács IZ, Frattasi S, Marchetti N, Mogensen PE (2009) Fixed frequency reuse for LTE-advanced systems in local area scenarios. In: IEEE 69th vehicular technology conference, Barcelona, Spain
2. Shannon C (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423, 623–656
3. Krasniqi B, Wrulich M, Mecklenbräuker CF (2009) Network-load dependent partial frequency reuse for LTE. In: Ninth international symposium on communication and information technology (ISCIT 2009), pp 672–676
4. Xie Z, Walke B (2009) Enhanced fractional frequency reuse to increase capacity of OFDMA systems. In: NTMS'09 the third international conference on new technologies, mobility and security, NJ, USA

5. Shbat M, Khan Md. RR, Tuzlukov V (2011) Simple feedback with priority list radio resource scheduling scheme for 3GPP LTE networks. In: Lecture notes in electrical engineering, Springer, Berlin
6. Lu Z, Tian H, Sun Q, Huang B, Zheng S (2010) An admission control strategy for soft frequency reuse deployment of LTE systems. In: The seventh IEEE conference on consumer communications and networking conference CCNC'10
7. Ikuno JC, Wrulich M, Rupp M (2010) System level simulation of LTE network. In: IEEE 71st vehicular technology conference VTC2010, Taipei, Taiwan

Cognitive Radio Simplex Link Management for Dynamic Spectrum Access Using GNU Radio

M. Adib Sarijari, Rozeha A. Rashid, N. Fisal, A. C. C. Lo, S. K. S. Yusof, N. Hija Mahalin, K. M. Khairul Rashid Arief Marwanto

Abstract The explosion of new wireless communication technologies and services has led to the increase in spectrum demand. The fixed spectrum allocation approach has resulted in current day spectrum scarcity and poorly utilized licensed spectrum. In order to overcome these problems, a new concept of accessing the spectrum, defined as dynamic spectrum access (DSA), is proposed. DSA mechanism enables unlicensed or cognitive users (CUs) to temporarily utilize a spectrum hole for a period of time. In this work, DSA based on cognitive radio (CR) technology is chosen due to its features of able to sense, learn, adapt and react according to the environment. The proposed design of the system consists of four main functional blocks: spectrum sensing, spectrum management, spectrum

M. Adib Sarijari (✉) · R. A. Rashid · N. Fisal · S. K. S. Yusof · N. Hija Mahalin · K. M. Khairul Rashid · A. Marwanto
Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Johor, Malaysia
e-mail: adib_sairi@fke.utm.my

R. A. Rashid
e-mail: rozeha@fke.utm.my; sheila@fke.utm.my; kamilah@fke.utm.my

N. Hija Mahalin
e-mail: nhija.m@gmail.com

K. M. Khairul Rashid
e-mail: kaeroul@fkegraduate.utm.my

A. Marwanto
e-mail: ariefmarwanto@unissula.ac.id

A. C. C. Lo
Wireless and Mobile Communications Group,
Delft University of Technology, Delft, Netherlands
e-mail: A.C.C.Lo@tudelft.nl

decision and data transmission. Spectrum management is further divided into three parts: spectrum identification, synchronization and link management. This paper focuses on the establishment of link management module in simplex mode. The implementation is done using GNU Radio and USRP SDR platform. The GMSK based and IEEE 802.15.4 standard radios, equipped with DSA capability using CR technique, have been developed and tested. The results show that the link module has successfully maintained CU's seamless communication while the DSA mechanism offers significant improvement in terms of achieved packet rate ratio (PRR).

Keywords Dynamic spectrum access · Cognitive radio · Spectrum sensing · Spectrum management and spectrum decision

1 Introduction

There has been tremendous demand for radio spectrum recently due to emerging of new wireless communication technologies and services such as long term evolution (LTE) and LTE-advanced. Traditional static spectrum allocation policies as practiced in many countries including Malaysia has resulted in spectrum scarcity as most radio bands are already assigned to users by the regulators. However, a number of spectrum occupancy measurements [1–4] has shown that the licensed spectrum is poorly utilized where some bands are overcrowded while other portions are moderately or rarely utilized as shown in Fig. 1. Cognitive radio (CR) is a promising technology to provide highly reliable communication for all users in the network wherever and whenever needed and to facilitate efficient spectrum utilization. It promotes spectrum sharing approach where unlicensed or cognitive users (CUs) are allowed access to licensed channel as long as there is no interference to licensed or primary users' (PUs) transmission.

There are two basic approaches of spectrum sharing which are the underlay and overlay. Underlay approach allows CUs to simultaneously share the spectrum with licensed user but the transmission of information is strictly limited to be below the designated threshold [5, 6]. In contrast, the overlay spectrum sharing prohibits CUs to simultaneously use the same frequency which is in use by PUs. Hence, CUs have to robustly identify a spectrum opportunity (spectrum hole).

Dynamic spectrum access (DSA) is an enabling technology for overlay spectrum sharing. IEEE 1900.1 working group defined DSA as a technique which enable a radio to dynamically change its operating frequency in real-time based on the condition of the environment and the objectives of the system [7, 8]. With DSA, CUs can temporarily utilize unoccupied bands but need to be sufficiently agile to vacate the space (time, frequency or spatial) once PUs are detected as not to cause harmful interference [5–8].

SPECTRUM ALLOCATIONS IN MALAYSIA

INTERNATIONAL MOBILE	INTERNATIONAL FIXED
INTERNATIONAL SATELLITE	INTERNATIONAL AIR
AIRCRAFT	AIRCRAFT
UNIDENTIFIED	RADIO AMATEUR
INTERNATIONAL SATELLITE	INTERNATIONAL SATELLITE
FREE	RADIOLOGICAL
AIRSATELLITE SERVICE	RADIOLOGICAL
LAND MOBILE	SPACE OPERATIONS
MARINE MOBILE	SPACE RESEARCH
AIRSATELLITE SERVICE	SPACE RESEARCH
R WALKY	OR WALKY
F SATELLITE	F SATELLITE
L SPACE TELEGRAPHY	L SATELLITE
A DEPT. INTERNATIONAL MOBILE	G LOCAL USE BY GOVERNMENT
A DEPT SPACE	G DEPARTMENT OF DEFENSE

NOTE:
1) This chart shows only the bands for each service. For details of frequency allocations and frequency assignments, please refer to the Malaysian Radio Plan.
2) The quality grades in the spectrum management system is not proportional to the actual amount of spectrum available.
ISSUE BY: JUNE 2020

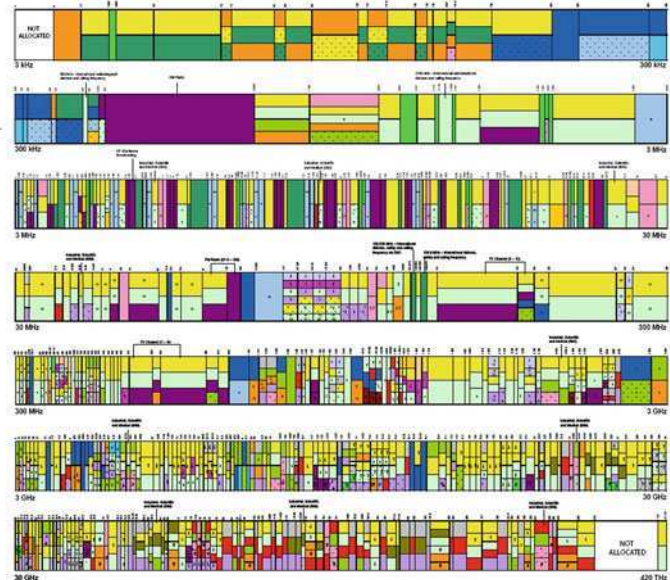


Fig. 1 Local spectrum utilization [1]

DSA can be realized through a number of available technologies, which include adaptive radio [9], cognitive radio (CR) [10], and reconfigurable radio [11, 12]. The most appropriate means to realize these radios are through the use of several existing software defined radio (SDR) platforms in the market [13–19]. A working prototype of DSA hardware implementation using GNU radio and USRP is presented in [20]. The work utilizes energy detector as its sensing method. Synchronization between two communicating CUs is established using simplex mode. A CU transmitter senses the channel to determine its status before the packet is transmitted. Each time it changes channel, CU transmitter will broadcast a number of synchronization packets on the found free channel. By sweeping channel and detecting the transmitted synchronization packet, the channel used by CU transmitter will be known by CU receiver.

In the work presented here, a similar DSA based CR system is developed using universal software radio peripheral (USRP) and GNU radio as hardware and software platforms, respectively. The proposed design consists of four main functional blocks which are spectrum sensing, spectrum management, spectrum decision and data transmission. However, the process of establishing communication between two CUs is further enhanced and made more effective by having a link management module, integrated with spectrum identification and synchronization as parts of spectrum management. This module is in simplex mode and is responsible for spectrum mobility where seamless communication requirements for CUs are maintained during the transition to better spectrum. An open platform

technology is also utilized to facilitate flexible modification according to users' requirement and specification.

The rest of the paper is organized as follows. [Section 2](#) presents the design concept of DSA based CR system. The details of Link Management module is explained in [Sect. 3](#). The CR network model is introduced in [Sect. 4](#) while [Sect. 5](#) discusses the implementation set-up and the results. The conclusion of this paper is outlined in [Sect. 6](#).

2 Design Concept of CR System for DSA

The proposed design of the CU system with DSA is illustrated by the block diagram in [Fig. 2](#). It consists of four main functioning blocks: spectrum sensing, spectrum management, spectrum decision and data transmission. Spectrum sensing is used to sense the spectrum and detects the presence of the PU on the scanned spectrum. Spectrum management is responsible for identifying the spectrum hole (spectrum hole identification), establishing synchronization and managing links with other CUs or secondary users (SUs).

Spectrum decision is responsible for summarizing the output from spectrum management which is the availability of the spectrum hole, the status of the synchronization and the status of the link and make decision on when and how to communicate, for instance the use and how long the channel can be utilized. Last but not least, the data transmission which functions is to forward data packets to the USRP (i.e. assemble and transmit the data). This paper highlights the work on link management in spectrum management module. More details on the operation of the CU can be found in [\[21\]](#).

3 Link Management

Link management resides in spectrum management module in the designed CR system. It is responsible for keeping seamless communication between CUs even when it changes its physical parameter such as the channel identification (id) used or spectrum band. In this work, the link management is in simplex mode. However, it can be easily extended to duplex mode as the system is built on open platform. The design of link management considers that the control packet has been successfully carried out.

The link management process begins after CU senses the spectrum hole. As shown in [Fig. 3](#), once spectrum hole is found, CU transmitter will initiate synchronization and connection id with the corresponding CU receiver by sending the control packet using the default channel (channel 20 as in [Table 1](#)). This is to ensure the corresponding CU has been informed that communication between both parties will begin or resume using the identified channel. Channel 20 is chosen as the default channel as it does not interfere with the neighboring access point.

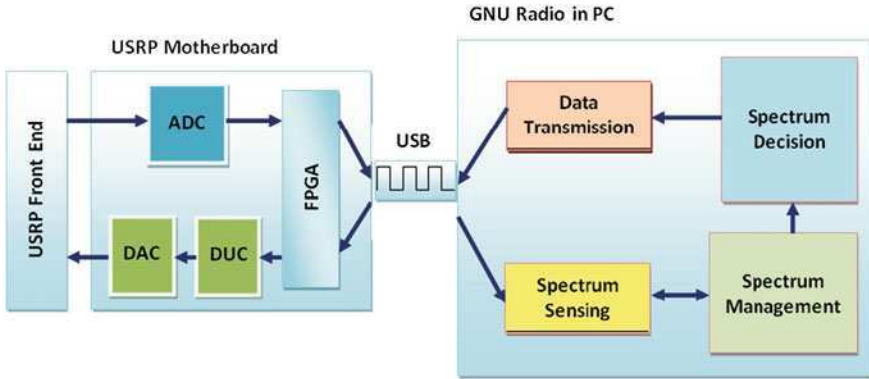


Fig. 2 Design concept of CR system for DSA

Fig. 3 Link management in CU system

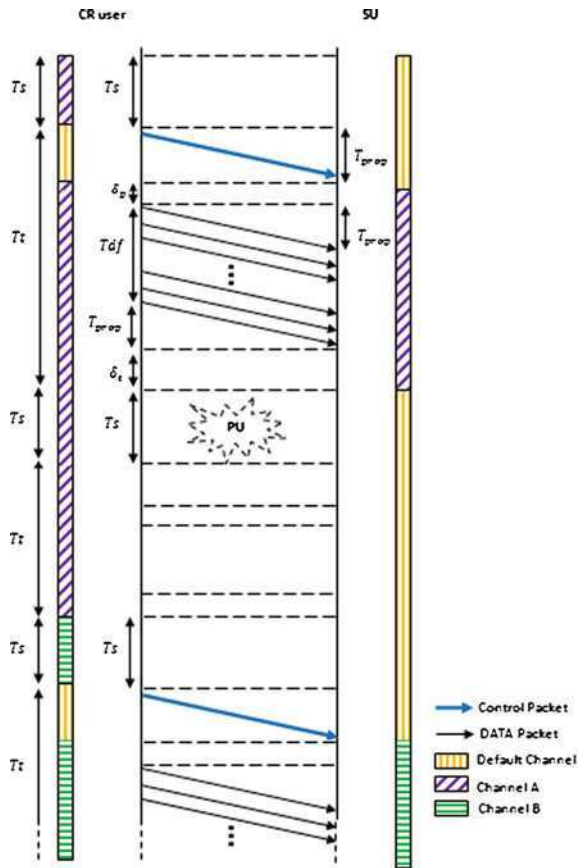


Table 1 Channel and frequency used in CR system [20]

Channel	Frequency (GHz)
1	2.404
2	2.408
3	2.412
4	2.416
5	2.420
6	2.424
7	2.428
8	2.432
9	2.436
10	2.440
11	2.444
12	2.448
13	2.452
14	2.456
15	2.460
16	2.464
17	2.468
18	2.472
19	2.476
20	2.480
21	2.484
22	2.488
23	2.492
24	2.496

The same process for sending data is repeated for the next time frame using the next identified spectrum hole as shown in Fig. 3.

In this figure, T_s is the sensing time equal to 31.59 ms. T_{prop} is the signal propagation delay as given in Eq. 1;

$$T_{prop} = \frac{d}{v} \quad (1)$$

where d is the distance between the transmitter and the receiver which in this case is equal to 0.66 m, and v is the speed of the speed of the light (300×10^6 m/s). Thus T_{prop} is equal to 2.2 ns. Tdf is the data packet transmission time and can be derived using Eq. 2,

$$Tdf = Tpx \text{ (No. of Packets)} \quad (2)$$

$$Tp = \frac{F_s}{L} \quad (3)$$

where F_s is the frame size, L is the link rate and Tp is the time needed to transmit a packet of data and can be calculated using Eq. 3. The frame size in GMSK based radio is 2 packet/frame and Tp is equal to 65.744 ms. Therefore Tdf is equal to

131.488 ms. δ_p is offset which is needed by the CU receiver to stabilize after changing to a new channel or spectrum band. Since USRP is used in this study and its response on the channel and spectrum band changes is very fast, therefore δ_p is set to zero. δ_t is the remaining time until the next frame and can be calculated using Eq. 4;

$$\delta_t = T_t - Tdf \quad (4)$$

In this work, T_t is equal to 192.25 ms. Therefore, δ_t is equal to 60.762 ms. The utilization of the link can be calculated using the following equation:

$$U = \frac{Tdf}{T_s + 2T_{prop} + \delta_p + Tdf + \delta_t} \quad (5)$$

Putting all the value in the equation gives U equal to 0.6839. The dominant parameters which degrade the utilization of the CR system is T_s and δ_t . The utilization can be improved by minimizing T_s and δ_t and it can be done as follow:

- Using faster and more powerful computer to execute the GNU Radio as this will speed up the processing of the signal processing block of the GNU Radio.
- Using joint sensing and data transmission as proposed in [22]. This technique is better if it is used with USRP2 which provides better bandwidth to exchange the data from SDR hardware to the GNU Radio.
- Choosing a lower sensing time for the CR system. However, this will degrade the Pd and Pfa.

4 CR Network Model

Figure 4 illustrates the CR network model used in this work. In this figure, PU and CU radios are operating in the same frequency band which in this work is at 2.4 GHz ISM band. The frequency started at 2.404 GHz which is channel 1 and end at 2.496 GHz which is channel 24. CU is assumed to operate in the PU coverage area.

4.1 CU Characteristics

Table 2 shows the parameters used by CU in this work. There are two types of CU radio system tested which are the GMSK based radio and IEEE 802.15.4 standard radio. Both radios operates in ISM 2.4 GHz band. The link rate of GMSK based radio is 500 kb/s and the bandwidth is 1 MHz. The maximum payload size of the radio is 4085 bytes which is limited by the maximum value of packet length in MAC header. The length can be increased by adding the size of the packet length



Fig. 4 CR network model

Table 2 Parameters used for SU radio

Parameter	GMSK based	IEEE 802.15.4 standard
Modulation	GMSK	O-QPSK
Bitrate	500 kb/s	256 kb/s
Band	2.4GHzISM Band	2.4GHz ISM Band
Bandwidth	1 MHz	3 MHz
Max. payload size	4085 bytes	128 bytes
MTU	4108 bytes	156 bytes

in the PHY header. The maximum transfer unit (MTU) for the GMSK based radio is 4108 bytes. For IEEE 802.15.4, the link rate is 256 kb/s with the bandwidth of 3 MHz. The modulation used in this radio is offset quadrature phase shift keying (O-QPSK) and the maximum packet length allowed is 128bytes and the MTU is 156 bytes.

4.2 PU Characteristics

The parameters used for the PU radio in this work are listed in Table 3. PU is operating in the same band as CU which is the ISM 2.4 GHz. This is to prove that PU and CU can share the same band. The bitrate of this radio is 500 kb/s with 1 MHz bandwidth. The modulation used for PU radio in this work is the differential phase shift keying (DQPSK) and the packet size transmitted is 1500 bytes. PU on and off time is based on work in [23] which is 352 and 650 ms, respectively. PU packet is generated every 30 ms within the on time.

Table 3 Parameters used for PU radio

Parameter	PU Radio
Modulation	DQPSK
Bitrate	500 kb/s
Band	2.4GHz ISM Band
Bandwidth	1 MHz
Pkt Size	1500 bytes

5 Implementation Setup

In the implementation of the CR system, four USRPs, one laptop and one personal computer (PC) were used as shown in Fig. 5. PC with USRP A acts as the CU receiver, PC with USRP B as the monitoring node (the spectrum analyzer), laptop and USRP C as the CU transmitter and laptop with USRP D acts as the PU transmitter. Daughter board used for this implementation is RFX2400 which covers frequencies from 2.3–2.9 GHz. In this implementation, only two channels are used as prove of concept purposes. The channels are channel 22 (2.488 GHz) as channel A and channel 23 (2.492 GHz) as channel B. These channels are considered since they do not overlap with the neighboring access points (WLAN) which can interfere with the USRP frequency operating at 2.4 GHz band.

5.1 Results of CR Network Implementation on GMSK Based Radio

The graphical results of the GMSK based CR network implementation are presented in Fig. 6a–d. In Fig. 6a, it is observed that PU and CR transmitters are transmitting at different channel frequencies in the same band; the signal with lower amplitude transmitting at 2.492 GHz is the CU signal with GMSK modulation while the higher one transmitting at 2.488 GHz is the PU signal with DQPSK modulation.

When PU appears on the channel where the CU is currently transmitting as shown in Fig. 6b, CU will stop its transmission as shown in Fig. 6c. CU will search for another free channel. Once the free channel is identified, CU will continue its communication by using this new free channel or spectrum hole as shown in the picture in Fig. 6d.

It is crucial for CU to change the channel frequency used not only to give the PU privilege of using the channel but also to protect it from interference. Figure 4 shows that the effect of interference on CU's packet reception rate for GMSK based radio. As shown in Fig. 7, with only 50% power transmission by PU, the packet reception rate (PRR) of CU drops to 0.084, a reduction of 91.6% from the original value.

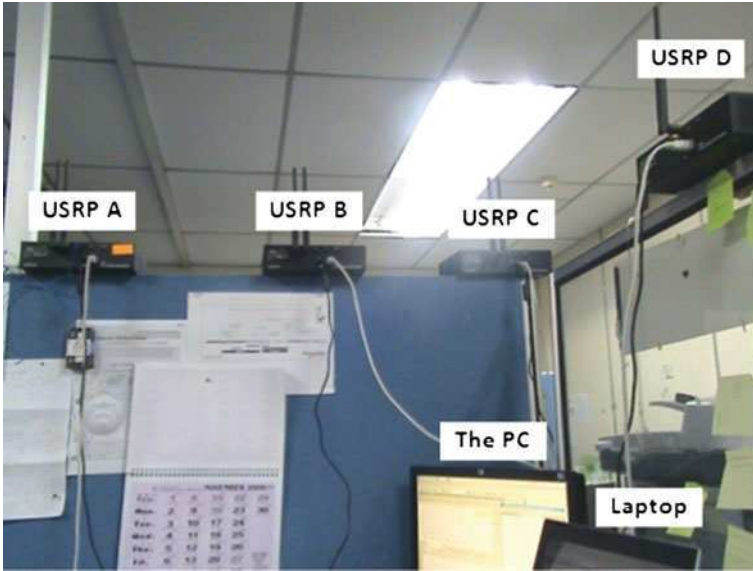


Fig. 5 Experimental setup of CR system implementation

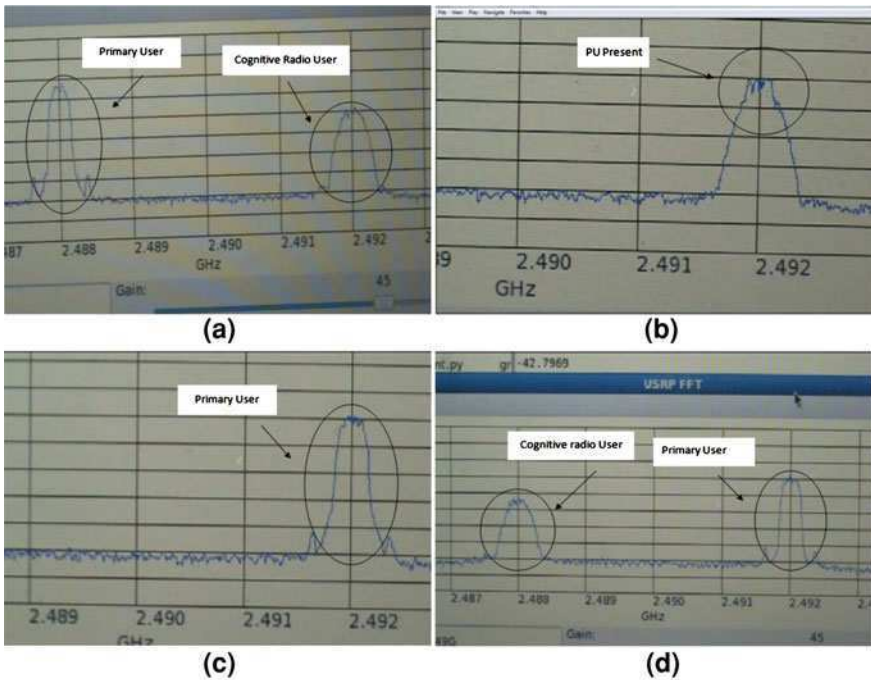


Fig. 6 a Overlay spectrum sharing between PU and CU b Detection of PU in the channel occupied previously by CU c CU stops transmission d CU finds another spectrum hole and resumes transmission

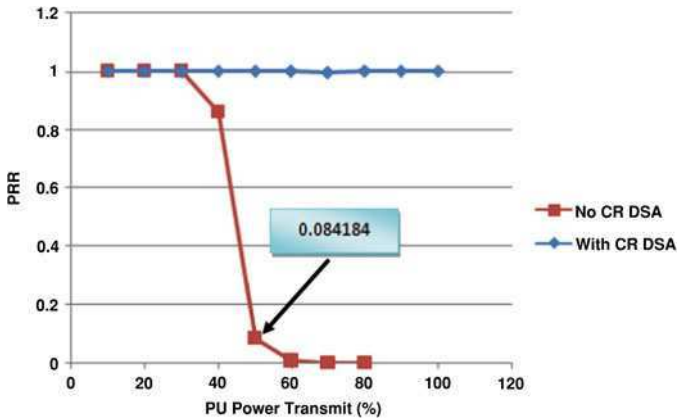


Fig. 7 PRR for GMSK based radio with and without CR system

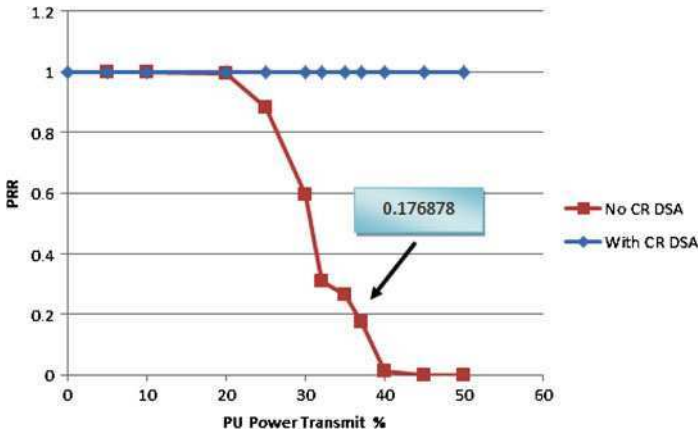


Fig. 8 PRR for IEEE 802.15.4 standard radio with and without DSA based CR system

5.2 Results of CR Network Implementation on IEEE 802.15.4 Standard Radio

Similarly for the IEEE 802.15.4 standard radio, the graph in Fig. 8 shows that the DSA based CR system improves the PRR of CU significantly. IEEE 802.15.4 standard radio without DSA mechanism suffers from the interference caused by the PU badly. As shown in this graph with only 37% power transmits used by the PU, CU’s PRR dropped to 0.176, which is a reduction by 82.6%. However, for the IEEE 802.15.4 radio which is equipped with DSA capability, the PRR value is maintained at around 1.0 even though PU transmits at 100% power.

6 Conclusion

The complete CR system for DSA which consists of spectrum sensing, spectrum management and spectrum decision has been implemented and tested. The working link management module which is part of the spectrum management is also successfully developed and implemented. The calculation shows that the minimum link utilization of the work is 68% and this can be improved by reducing the sensing time.

Furthermore, CU system deploying proposed DSA is shown to significantly improve the spectrum utilization in terms of packet reception rate for both GMSK based and IEEE 802.15.4 standard radios. Future works include the implementation of the link management in duplex mode and more measurements to evaluate the performance of the developed system.

References

1. Suruhanjaya Komunikasi dan Multimedia Malaysia (SKMM) (2010) Spectrum allocation in Malaysia. http://www.skmm.gov.my/link_file/what_we_do/spectrum/allocation/pdf/Malaysian_Spectrum_Allocations_Chart.pdf. Accessed on Oct 29
2. McHenry MA, Tenhala PA, McCloskey D, Roberson DA, Hood CS (2006) Chicago spectrum occupancy measurements & analysis and a long-term studies proposal. In: Proceedings of the first international workshop on technology and policy for accessing spectrum. Boston, MA, 5 Aug 2006
3. Cabric D, Mishra SM, Brodersen RW (2004) Implementation issues in spectrum sensing for cognitive radios. In: Asilomar conference on signals, systems, and computers. vol 1. pp 772–776, Nov 2004
4. Akyildiz IF, Lee W-Y, Vuran MC, Mohanty S (2006) Next generation dynamic spectrum access cognitive radio wireless networks: a survey. *Int J ComputTelecommun Network* 50:2127–2159
5. Berlemann L, Dimitrakopoulos G, Moessner K, Hoffmeyer J (2005) Cognitive radio, Management of spectrum and radio resources in reconfigurable networks. In: Wireless world research forum working group 6 White Paper
6. Rashid RA, Faisal N (2009) Issues of spectrum sensing in cognitive radio based system. In: Third south east asia technical universities consortium (SEATUC), Johor, Malaysia, 25–26 Feb 2009
7. International telecommunication union (ITU) (2002) Handbook frequency-adaptive communication systems and networks in the MF/HF bands
8. Mitola III J (2000) Cognitive radio: an integrated agent architecture for software defined radio. Ph.D. thesis, KTH- Royal institute of technology, Stockholm, Sweden
9. International telecommunication union (ITU) (2002) Handbook frequency-adaptive communication systems and networks in the MF/HF bands
10. Mitola III J (2000) Cognitive radio: an integrated agent architecture for software defined radio. Ph.D. thesis, KTH-Royal institute of technology, Stockholm, Sweden
11. Guenin J (2008) IEEE SCC41 standards for dynamic spectrum access networks. In: IEICE software and cognitive radio expo and technical conference. Tokyo, Japan, 1 Aug 2008
12. Hamid M (2008) Dynamic spectrum access in cognitive radio networks: Aspects of mac layer sensing. Master Thesis, Blekinge Institute of Technology, Ronneby, Sweden, Dec 2008

13. Baldiniet G et al (2008) Reconfigurable radio systems for public safety based on low-cost platforms. EuroISI 2008, LNCS vol 5376. pp 237–247
14. IEEE communications society (2008) 1900.1-2008 IEEE standard definitions and concepts for dynamic spectrum access: terminology relating to emerging wireless networks, System functionality, and spectrum management. 26 Sept 2008. pp c1–48. E-ISBN 978-0-7381-5776-4, Printed ISBN 978-0-7381-5777-1
15. Mitola III J (1992) Software radios-survey, critical evaluation and future directions. IEEE National telesystems conference. Washington, DC, May 1992, pp 19–20, 13/15–13/23
16. Mitola J (2002) The software radio architecture. IEEE Commun Mag 33(5):26–38
17. Flex radio systems (2010) <http://www.flex-radio.com/>. Accessed 2 Nov 2010
18. Berkeley emulation engine (2010) 2. <http://bee2.eecs.berkeley.edu/>. Accessed 2 Nov 2010
19. GNU radio (2010) <http://gnuradio.org>. Accessed 2 Nov 2010
20. Alice Crohas (2008) Practical implementation of a cognitive radio system for dynamic spectrum access. Master of science in electrical engineering thesis. Notre Dame, Indiana, July 2008
21. Adib Sarijari M, Rashid RA, Faisal N, Lo ACC, Yusof SKS, Mahalin NH (2011) Dynamic spectrum access using cognitive radio utilizing GNU radio and USRP. 26th wireless world research forum (WWRF26) Doha, Qatar, 11–14 Apr 2011
22. Do T, Mark BL (2009) Joint spatial-temporal spectrum sensing for cognitive radio networks. In: Proceedings of conference on information sciences and systems (CISS 09) Baltimore, MD, Mar 2009
23. Pei Y, Hoang AT, Liang Y-C (2007) Sensing-throughput tradeoff in cognitive radio networks: how frequently should spectrum sensing be carried out? IEEE 18th international symposium on personal, indoor and mobile radio communications. PIMRC 2007, Athens, pp 1–5, 3–7 Sept 2007

Do Children See Robots Differently? A Study Comparing Eye-Movements of Adults vs. Children When Looking at Robotic Faces

Eunil Park, Ki Joon Kim and Angel P. del Pobil

Abstract A 2 (face type: robot face vs. human face) \times 2 (participants' age: adults vs. children) between-subjects experiment with four conditions was conducted to explore whether adults and children view robotic faces differently. Participants were presented with a series of pictures including human or robotic faces while their eye movements were being recorded and analyzed by a Tobii x120 Eye Tracker. Results showed that adults had a longer eye fixation time on the eyes than children did for both human and robotic faces. However, children had a longer fixation time on the mouth and nose than adults for both human and robotic faces. Both implications and limitations of the present study as well as guidelines for future research are discussed.

Keywords Eye-tracking · Robot · Children · Eye-movements

E. Park (✉) · K. J. Kim · A. P. del Pobil
Department of Interaction Science, Sungkyunkwan University,
Seoul, South Korea
e-mail: pa1324@skku.edu

K. J. Kim
e-mail: veritate@skku.edu

A. P. del Pobil
e-mail: pobil@icc.uji.es

A. P. del Pobil
Robotic Intelligence Laboratory, University Jaume-I, Castello, Spain

1 Introduction

In social interaction between humans, we deliver information, emotion, intention, and desire by diverse ways of communication such as conversation, facial expression and gestures [1]. Although social characteristics of these functions were pointed out many times, we still lack empirical research in the real world. That is, the recent results of other robotic studies cannot fill up the high level of social interaction between human and robot [2–5]. Therefore, we need to suggest more sensitive and varied methods for interaction comparable to human-to-human communication. Many features (e.g. gender, age, and so on) affect social interactions and communication with robots [6, 7].

Concurrently with verbal communication between humans, they communicated by using non-verbal behavior and expression [8]. For example, facial expressions have been used to become aware conscious of the other man's situation or condition. Especially, in case of children, they do not have enough social cognition skills compared to adults. That is, the way of children to cognize robots' face may be different with adults' way. Social cognition skills of children improve gradually. Through this process, they read others' mind and understand that people have different feelings and thoughts. Then, they integrate various cues for social cognition from their process. For example, they combine many cues, which are not only verbal information but also non-verbal information such as facial expression. Specially, because facial expressions provide decisive way about expression of mind, facial recognition is the one of most essential thing to communicate and act feedback [8, 9]. However, few studies have focused on a method of face recognition in human-robot interaction.

That is, the aim of this paper is to find a sequential way of face recognition in communication between human and robot. It is a base-study for non-verbal communication of human-robot interaction. Additionally, we want to find a difference between adults and children with the overall aim.

2 Experiment

2.1 Design

The experiment was a between-subject design with four conditions: 2 (Age level of Participant: *Adults vs. Children*) * 2 (Face type: *Robot face vs. Human face*) (Table 1).

2.2 Participants

Forty participants were recruited from a large private university in Seoul. The age of children ranged from 7 to 12, with the mean 8.7 years (SD = 1.58). The age of

Table 1 Design of experiment

	Adults	Children
Robot face	10	10
Human face	10	10

adults ranged from 24 to 31, with the mean 27.6 years ($SD = 2.87$). Each condition had 10 participants (5 females).

2.3 Apparatus and Stimulus

Pictures of human and robot faces were prepared. For the human face, we hired eight actors (4 male and 4 female) who were instructed not to express any emotion during the photo shoot. Four actors' pictures were eliminated by facial-expression experts due to having facial expressions; 16 pictures from four actors were selected as stimulus materials to be used in the experiment.

For the robot face, images of 16 existing robots were collected. The facial-expression experts selected 16 pictures from four robots (i.e. Kaspar [10], Kobian [11], EveR-2 Muse [12], and Jules [13]) as stimulus materials.

Tobii x120 Eye Tracker (Fig. 1) was used to examine participants' eye movements. Fixation time on core features of the faces such as eyes, nose, and mouth was recorded.

2.4 Procedure

A 20-inch LCD monitor and Tobii x120 were prepared in a room with a chair. Participants were asked to look at the pictures of the faces while the eye-tracker was being calibrated. The experimenter then started the main session, and the eye-tracker recorded participants' eye movements. Sixteen images were randomly displayed on black background for 4 s. Between every image, black-screen was displayed for a second.

2.5 Measurement

Eye fixation time on mouth, eyes, and nose was recorded and calculated by eye-tracker (Tobii x120). For accurate comparison between conditions, we used percentage-results of fixation time in entire presented time.



Fig. 1 Tobii x120, an eye tracker used in our experiment

Fig. 2 A sample picture of eye tracking in adults group



2.6 Result

Effects of human vs. robot face. An analysis of variance (ANOVA) was conducted to analyze the effects of face type on the percentage of fixation time in each core feature (eyes, mouth, nose and sum of core feature). Although it was marginally significant, the results from the ANOVA indicated that participants who viewed the robot faces ($M = 23.47\%$, $SD = 22.00\%$) reported a longer fixation time on the mouth than those with the human faces ($M = 15.27\%$, $SD = 11.38\%$), $F(1, 36) = 3.91$, $p = 0.056$ (Fig. 2).

Effects of age. Participants' age also had significant effects on the percentage of fixation time in eyes, mouth, and nose. Results from the ANOVA showed that adults ($M = 51.85\%$, $SD = 16.37$) had a longer fixation time on eyes than children did ($M = 23.56\%$, $SD = 16.01$), $F(1, 36) = 28.89$, $p < 0.001$. The adults ($M = 9.12\%$, $SD = 6.72$), however, had a shorter fixation time on mouth than children did ($M = 29.62\%$, $SD = 19.56$), $F(1, 36) = 24.46$, $p < 0.001$. The adults

Fig. 3 Mean of percentage of fixation time on mouth

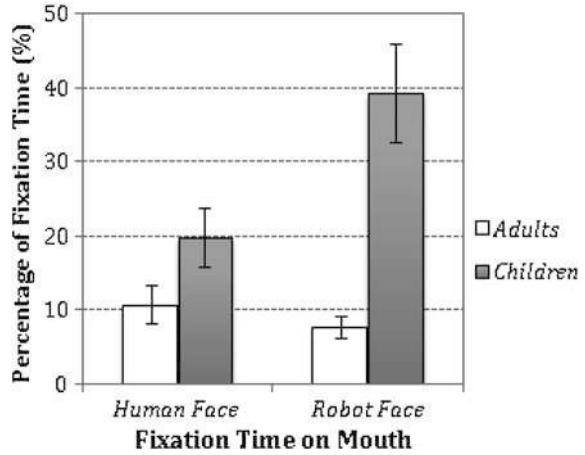


Fig. 4 Mean of percentage of fixation time on nose

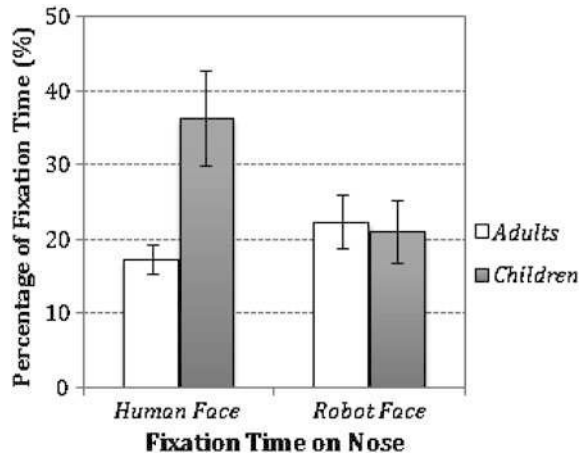


Table 2 Mean and standard deviation of fixation time in facial features

Facial features		Human face (%)	Robot face (%)
Adults	Eyes	52.14 (18.39)	51.56 (15.07)
	Nose	17.33 (6.19)	22.28 (11.42)
	Mouth	10.67 (8.24)	7.57 (4.71)
Children	Eyes	25.40 (20.26)	21.94 (11.12)
	Nose	36.32 (20.61)	20.91 (13.27)
	Mouth	19.88 (12.58)	39.36 (20.94)

($M = 19.81\%$, $SD = 9.29$) also had a shorter fixation time on nose than children did ($M = 28.61\%$, $SD = 18.63$), $F(1, 36) = 4.031$, $p = 0.052$.

Interaction. The interaction between the face type and age had effects on fixation time on mouth ($F(1, 36) = 7.42$, $p = .01$) and nose ($F(1, 36) = 5.34$, $p < 0.05$), such that children looked at the robot’s mouth longer than the nose while adults showed no difference due to human vs. robot face (Figs. 3 and 4, Table 2).

3 Discussion and Conclusion

The present study explored the face searching patterns for human and robotic faces. Participants tend to see the core features of human and robot faces such as eyes, nose and mouth. On average, participants in both human and robot face groups focused on core facial features in 81.34% of entire presenting time. However, we found a difference of concentration patterns in fixation time. Specially, the different concentrations of fixation were shown between adults and children.

Moreover, we were able to find different focusing results between adults and children. That is, adults tended to more focused on eyes compared to children's fixation. It is striking results that children usually gazed mouths of human and robot face compared to adults' gaze. It may be caused by two reasons. First, areas of robot's mouth were more spacious than areas of human mouth, averagely 8.4%. Second, eyes of robot faces did not have movements normally. However, eyes of human faces were able to shift normally.

Although our images of robot were designed based on humanoid robots, the differences of object were found in the fixation time. These patterns and focusing rates may be able to use design social interactive robot.

Our future study will include images of entire body of human and robot. Specially, if we designed robot images from human images, for example, robotic body and face, and Arnold Schwarzenegger [14] from The Terminator Series, we will find eye-movements in the whole body of two conditions.

Acknowledgments This study was supported by a grant from the World-Class University program (R31-2008-000-10062-0) of the Korean Ministry of Education, Science and Technology via the National Research Foundation.

References

1. Smalley G, Scott S (1982) For better or for best. Zondervan Publishing House, Grand Rapids
2. Park E, Kong H, Lim H, Lee J, You S, del Pobil AP (2011) The effect of Robot's behavior vs. Appearance on communication with humans. In: Sixth ACM/IEEE international conference on human-robot interaction, ACM, New York, pp 219–220
3. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. *Robot Auton Syst* 42(3):143–166
4. Seiter JS, Gass RH (2003) Perspectives on persuasion, social influence, and compliance gaining. Allyn & Bacon, Boston
5. Gockley R, Bruce A, Forlizzi J, Michalowski M, Mundell A, Rosenthal S, Sellner B, Simmons R, Snipes K, Schultz AC, Jue W (2005) Designing robots for long-term social interaction. In: IEEE/RSJ international conference on intelligent robots and systems, IEEE Press, New York, pp 1338–1343
6. Siegel M, Breazeal C, Norton MI (2009) Persuasive robotics: the influence of robot gender on human behavior. In: IEEE/RSJ international conference on intelligent robots and systems, IEEE Press, New York, pp 2563–2568

7. Schermerhorn P, Scheutz M, Crowell CR (2008) Robot social presence and gender: do females view robots differently than males. In: Third ACM/IEEE international conference on human-robot interaction, ACM, New York, pp 263–270
8. Gross AL, Ballif B (1991) Children’s understanding of emotion from facial expressions and situations. A review. *Develop Rev* 11(4):368–398
9. Felleman ES, Carlson CR, Barden RC, Rosenberg L, Marsters JC (1983) Childrens and adults recognition of spontaneous and posed emotional expressions in young children. *Dev Psychol* 19(3):405–413
10. Dautenhahn K, Nehaniv CL, Walters ML, Robins B, Kose-Bagci H, Mirza NA, Blow M (2009) KASPAR—a minimally expressive humanoid robot for human-robot interaction. *Appl Bionics Biomech* 6(3):369–397
11. Zecca M, Mizoguchi Y, Endo K, Lida F, Kawabata Y, Endo N, Itoh K, Takanishi A (2009) Whole body emotion expressions for KOBIAN humanoid robot—Preliminary experiments with different emotional patters. In: 18th IEEE international symposium on robot and human interactive communication IEEE Press, New York, pp 381–386
12. Lee D, Lee T, So B, Choi M, Shin E, Yang K, Baek M, Kim H, Lee H (2008) Development of an android for emotional expression and human interaction. In: Seventeenth world congress the international federation of automatic control, Seoul
13. Bristol Robotics Laboratory, <http://www.brl.ac.uk/projects/empathy/empathy.html>
14. Terminator Series, Wikipedia, http://en.wikipedia.org/wiki/The_Terminator

Relative Self-Localization Estimation for Indoor Mobile Robot

Xing Xiong and Byung-Jae Choi

Abstract It is important for an autonomous mobile robot to know where it is after movement. In this paper, we consider the problem of mobile robot indoor position estimation using only visual information from a single camera.

Keywords Ceiling key point extraction · Scale invariant feature transform (SIFT) · Radial distortion calibration

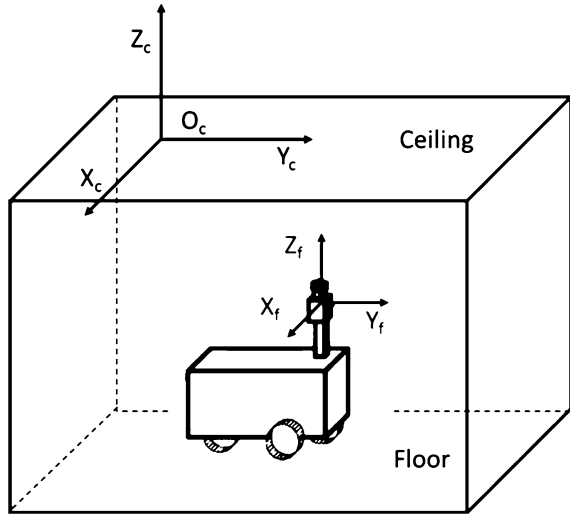
1 Introduction

Mobile robot self-localization is a mandatory task in accomplishing full autonomy during navigation. In an indoor environment, the floor is assumed to be planar. The ceiling consists of a series of blocks, which form a chessboard pattern parallel to the floor. A camera is mounted on the top of a mobile robot working on the floor. Its orientation is upright, directs to the block ceiling, as shown in Fig. 1.

X. Xiong (✉) · B.-J. Choi
School of Electronic Engineering, Daegu University Jillyang,
Gyeongsan, Gyeongbuk 712-714, Korea
e-mail: GaleWing@gmail.com

B.-J. Choi
e-mail: bjchoi@daegu.ac.kr

Fig. 1 Model of ceiling based visual positioning



2 Method

In this paper, the method is based on the use of a SIFT-based key point image-matching process. The SIFT algorithm was used to find the same feature point between two different images. The positioning procedure consists of following main parts: feature point extraction and matching, radial lens distortions of the key point, removal of non-ceiling key point, oval construction and determine the changed distance and angle.

3 Simulation Result and Conclusions

The experiment system consisted of a fish-eyes lens and a camera. The following image produced that robot, movement in short time, change in the position and orientation Fig. 2.

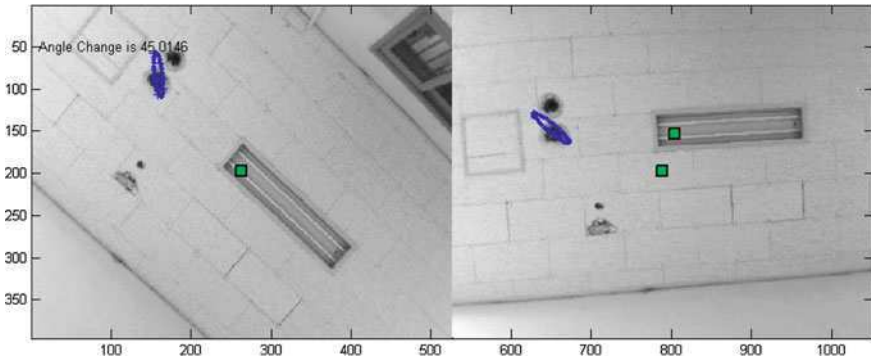


Fig. 2 Experimental result of the position and orientation

Our main contribution is that a new visual positioning method based on the features on ceiling is presented for an indoor mobile robot. The experimental results verify the effectiveness of the proposed method.

Acknowledgments This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology under Grant 2010-0006588.

References

- Yuen DCK, Bruce A (2005) MacDonald: vision-based localization algorithm based on landmark matching, triangulation, reconstruction, and comparison. *IEEE Trans Robot* 21(2):217–226
- Koenig A, Kessler J, Gross H-M (2008) A graph matching technique for an appearance-based, visual SLAM-approach using Rao-Blackwellized particle filters. In: *IEEE/RSJ international conference on intelligent robots and systems*, pp 1576–1581

Q(λ) Based Vector Direction for Path Planning Problem of Autonomous Mobile Robots

Hyun Ju Hwang, Hoang Huu Viet and TaeChoong Chung

Abstract This paper presents a novel algorithm to improve the efficiency of path planning for autonomous mobile robots. In an obstacle-free environment, the path planning of a robot is attained by following the vector direction from its current position to the goal position. In an obstacle environment, while following the vector direction, a robot has to avoid obstacles by rotating the moving direction. To accomplish the obstacle avoidance task for the mobile robot, the Q(λ) algorithm is employed to train the robot to learn suitable moving directions. Experimental results show that the proposed algorithm is soundness and completeness with a fast learning rate in the large environment of states and obstacles.

Keywords Reinforcement learning · Q-learning · Q(λ) algorithm · Path planning · Mobile robots

1 Introduction

In recent years, the path planning problem of autonomous mobile robots has been received a great interest from robotics researchers due to its important roles in applications of robots such as space exploration, ocean exploration,

H. J. Hwang · H. H. Viet (✉) · T. Chung
Artificial Intelligence Lab, Department of Computer Engineering,
School of Electronics and Information, Kyung Hee University, 1-Seocheon,
Giheung, Yongin, Gyeonggi 446-701, South Korea
e-mail: viethh@khu.ac.kr

H. J. Hwang
e-mail: hjoo@khu.ac.kr

T. Chung
e-mail: tcchung@khu.ac.kr

service and medical applications, and so on. The basic path planning of an autonomous mobile robot refers to determining a collision-free path for the robot from its position to the goal position through an obstacle environment without human intervention. Finding a solution to path planning as quickly as possible is one of the most fundamental problems to successful applications of autonomous mobile robots.

Reinforcement Learning (RL) is a branch of Artificial Intelligence in which an agent learns by interacting with its environment to gain experiences from the feedbacks or reward signals of the environment [1, 2]. By using a trial-and-error process, a RL agent is able to learn a policy that maximizes the cumulative reward intake of the agent over time. In literature, there have been several RL algorithms suggested to solve the path planning problem of autonomous mobile robots. Among those algorithms of RL, the Q-learning algorithm [3] has been frequently used to solve this problem. Smart et al. [4] introduce a framework based on the Q-learning algorithm for the path planning problem of mobile robots. In an environment with sparse reward functions, the chances of finding a reward by chance are very small indeed. Therefore, a learning robot is not effective in the early stages of the learning process in such environment. A proposed solution to this problem is to split the learning process into two phases. In the first phase, the robot is controlled by a supplied control policy which implemented by either control codes, or a human directly controlling the robot with a joystick. In the second phase, the robot learns policies by applying the Q-learning algorithm. Their method easily incorporates human knowledge about how to support a task in a learning system. Zamstein et al. [5] present a solution to the path planning problem for a real robot. The Q-learning is chosen to learn an optimal policy of actions by using only the current sensor inputs in their Koolio system. Although the Q-learning is not the most efficient of RL methods but it is chosen for the Koolio system because it can be easily programmed. Indrani et al. [6] propose an extension of the Q-learning algorithm for the path planning problem of a mobile robot. By using a flag variable to keep a track of the necessary for updating in the entries of the Q-table, their algorithm avoids unnecessary computations to reduce both space and time-complexity for updating the Q-table. Another approach proposed by Vien et al. [7] is that it combines an ant colony optimization algorithm with the Q-learning algorithm to solve the path planning problem for a mobile robot. Experiments show that their Ant-Q algorithm finds a very good path with a high convergence rate.

With the observations mentioned above, it can be seen that there is a tendency among researchers solving the path planning problem of mobile robots by using the Q-learning algorithm. The Q-learning is a model-free method. This means that it does not require an explicit model of an environment, thus it can be popularly employed to solve the path planning problem. When the Q-learning algorithm is augmented with eligibility traces, it is known as Watkins's $Q(\lambda)$ algorithm or $Q(\lambda)$ algorithm [2]. Eligibility traces are a recent memory to store traces. If an eligibility trace stores a trace of the state-action pairs taken, it is possible to back more than one step at a time. This can thus

increase the learning speed of the Q(λ) algorithm in order to converge to a near-optimality path comparing with the Q-learning algorithm. Therefore, we have reasons to believe that the Q(λ) algorithm that is going to be applied to our approach instead of the Q-learning algorithm will be more effective for the path planning problem of autonomous mobile robots.

It is assumed that the robot knows the goal position at each instant time through sensory inputs. In an obstacle-free environment, the path planning of the robot is achieved by following the vector direction from its position to the goal position through states of the environment. However, this is not true in an obstacle environment. To solve this problem, the Q(λ) based Vector Direction, called QVD(λ) algorithm, for the path planning problem of autonomous mobile robots is proposed. In our approach, the objective is to make the learning robot mimic the human reasoning. That is at each instant time, the robot moves one step to the next position by following the vector direction of its position and the goal position if the next position is not an obstacle. Otherwise, it has to rotate the moving direction to avoid an obstacle and move to the next free position.

The rest of this paper is organized as follows: Sect. 2 presents the proposed method. The experimental results are described in Sect. 3. Finally, we conclude our work in Sect. 4.

2 Proposed Method

In this section, the assumptions for the path planning problem of an autonomous mobile robot are going to be described and the QVD(λ) algorithm for solving this problem is also presented.

2.1 Assumptions

Assumption 1. The environment of the robot consists of the goal position and obstacles. The position and shape of obstacles are totally unknown by the robot.

Assumption 2. The robot is equipped with all necessary sensors to know its position, the goal position, and to detect obstacles if collisions occur during navigating time.

Assumption 3. The robot initially has no knowledge of the effect of its actions on what position it will occupy next and the environment provides rewards to the robot that this reward structure is also initially unknown to the robot.

Assumption 4. From its current position, the robot can move to an adjacent position in one of the eight directions, *East*, *North-East*, *North*, *North-West*, *West*, *South-West*, *South*, and *South-East*, except that any direction that takes the robot into obstacles or outside its environment, in which case the robot keeps its current position.

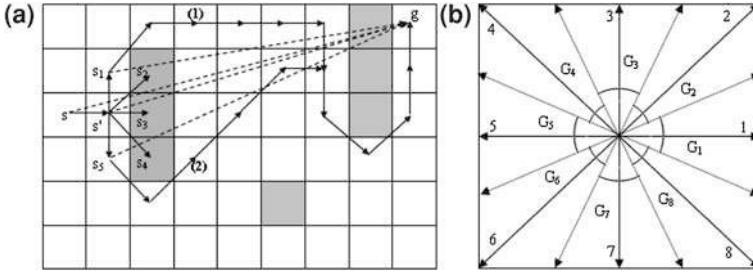


Fig. 1 **a** The illustrative environment, **b** Eight regions and eight actions at the state s

The robot’s objective is to discover a collision-free path as quickly as possible from its position to the goal position through its environment.

2.2 The QVD(λ) Algorithm

Figure 1a shows an illustrative environment to analyze our approach, where shaded cells represent obstacles and the robot can occupy any other cells, called states, of this environment.

It is assumed that the robot position is at the state $s = (x_r, y_r)$, and the goal position is at the state $g = (x_g, y_g)$. The task of the robot is to find suitable moving directions at states to reach to the goal position. The vector direction of the vector \vec{sg} is defined as equation (1).

$$\theta = \arctan \frac{y_g - y_r}{x_g - x_r}, \theta \in [-\pi, \pi] \tag{1}$$

To reduce the number of states of the environment, the interval $(-\pi, \pi)$ is divided into the eight angular regions as in equation (2). Based on the defined regions, the set of eight actions, $A(s)$, at the state s , along with the action selection rule, is defined as in equation (3), where actions 1, 2, 3, 4, 5, 6, 7, and 8 correspond to the eight moving directions of the robot including *East*, *North-East*, *North*, *North-West*, *West*, *South-West*, *South*, and *South-East* as shown in Fig. 1b.

$$\left\{ \begin{array}{l} G_1 = (-\pi/8, \pi/8] \\ G_2 = (\pi/8, 3\pi/8] \\ G_3 = (3\pi/8, 5\pi/8] \\ G_4 = (5\pi/8, 7\pi/8] \\ G_5 = (-7\pi/8, -\pi] \cup [7\pi/8, \pi] \\ G_6 = (-7\pi/8, -5\pi/8] \\ G_7 = (-5\pi/8, -3\pi/8] \\ G_8 = (-3\pi/8, -\pi/8] \end{array} \right. \tag{2}$$

$$A(s) = \{1, 2, 3, 4, 5, 6, 7, 8\}, a = i \in A(s), \text{ if } \theta \in G_i \quad (3)$$

If the robot is navigated in an obstacle-free environment, using the equations (1, 2), and (3), it is sure that the robot will reach to the goal position with the shortest path. If the robot is navigated in an obstacle environment, then some actions at some states cannot be implemented because of obstacles. Therefore, an angular deviation of the angle θ is proposed as in equation (4) to an avoid obstacle.

$$\theta' = \theta + n \frac{\pi}{4}, n = -3, \dots, 4 \quad (4)$$

An illustrative example is shown in Fig. 1a, where the dotted lines represent vector directions at the states $s, s', s_I,$ and s_5 . At the beginning, the robot is at the state s and the angle θ of the vector $\vec{s}g$ is determined by equation (1). Based on equations (2) and (3), it can be seen that the angle $\theta \in G_I$, so the selected action in $A(s)$ is $a = 1$ (*East* direction) and the robot moves to s' . Since s' is a state, so the robot occupies the state s' . In this case, the angle θ' is equal to the angle θ ($n = 0$). Next, the angle θ of the vector $\vec{s'}g$ is determined by equation (1), and the selected action in $A(s')$ is $a' = 1$ (*East* direction). However the robot cannot move to s_3 because s_3 is an obstacle. Therefore, the robot has to rotate an angle $\theta' = \theta + \pi/4$ ($n = 1$), or $\theta' = \theta - \pi/4$ ($n = -1$). It is assumed that the selected angle is $\theta' = \theta - \pi/4$, then the selected action in $A(s')$ with respect to the angle θ' is $a' = 8$ (*South-East* direction). But the robot cannot move to s_4 because s_4 is an obstacle. So, robot has to rotate an angle $\theta' = \theta - 2*\pi/4$ ($n = -2$) and the selected action in $A(s')$ with respect to the angle θ' is $a' = 7$ (*South* direction) and the robot moves to the state s_5 . By the similar way, the robot moves one step to the next position by following the vector direction of its position and the goal position if the next position is a state. Otherwise, it has to rotate the moving direction to avoid an obstacle and move to the next state.

To determine coefficients $n \in [-3, 4]$ in equation (4) for all states of the environment, the QVD(λ) algorithm is proposed to train the robot to learn an optimal action $a \in A(s)$ of the state s for avoiding an obstacle, and then the value of n is determined from the action a as in equation (5).

$$n = a - 4, a \in A(s) \quad (5)$$

The reward function given to the robot is defined as in equation (6), where s is the current state, s' is the next position after taking action $a \in A(s)$.

$$r(s, a, s') = \begin{cases} 1, & \text{if } s' \text{ is the goal state} \\ 0, & \text{if } s' \text{ is a state} \\ -1, & \text{if } s' \text{ is an obstacle.} \end{cases} \quad (6)$$

Algorithm 1. The QVD(λ) algorithm

```

1:   Initialize  $Q(s,a) = 0$  and  $e(s,a) = 0$ , for all  $s \in S$ ,  $a \in A(s)$ 
2:   Repeat (for each episode):
3:     Initialize  $s$ ,  $a = 4$  (i.e., default value of  $n = 0$ )
4:     Repeat (for each step of episode):
5:       (a) Compute the angle  $\theta$  based on equation (1)
6:       (b) Compute coefficient  $n$  based on equation (5)
7:       (c) Compute  $\theta'$  based on equation (4)
8:       (d) Take action  $a$  based on the angle  $\theta'$  [equations (2, 3)], observe  $r$ ,  $s'$ 
9:       (e) If ( $s'$  is a state) then
10:         $a' = 4$  (i.e., default value of  $n = 0$ )
11:         $a^* = 4$  (i.e., default value of  $n = 0$ )
12:      Else
13:         $a' \leftarrow \varepsilon$ -greedy( $s', Q$ )
14:         $a^* \leftarrow \operatorname{argmax}_b Q(s', b)$  (if  $a'$  ties for the max, then  $a^* \leftarrow a'$ )
15:      (f)  $\delta \leftarrow r + \gamma Q(s', a^*) - Q(s, a)$ 
16:      (g)  $e(s, a) \leftarrow I$ 
17:      (h) For all  $s$ ,  $a$ :
18:         $Q(s, a) \leftarrow Q(s, a) + \alpha \delta e(s, a)$ 
19:        If  $a' = a^*$  then  $e(s, a) \leftarrow \gamma \lambda e(s, a)$ 
20:        Else  $e(s, a) \leftarrow 0$ 
21:      (i)  $s' = s$ ;  $a' = a$ 
22:    Until  $s$  is the goal state.

```

The complete QVD(λ) algorithm is shown in Algorithm 1. The $Q(s, a)$ is the action value, the $e(s, a)$ represents the eligibility trace of the state-action pair (s, a) , the state s' is the next state after taking action a , r is the reward value, α ($0 < \alpha < 1$) is the learning rate, γ ($0 < \gamma < 1$) is the discount rate, and λ ($0 \leq \lambda \leq 1$) is trace-decay parameter. An ε -greedy strategy is used to select actions in the QVD(λ) algorithm. According to this strategy, the robot chooses the action having the highest value of the Q-value at the state s with probability of $1-\varepsilon$, and chooses a random action (*non-greedy* action) with a small probability of ε . At each episode, the robot begins at the start position s and it chooses the default action corresponding to the vector direction of the state s and the goal position. At each step of the current episode, the vector direction θ , the coefficient n , and the rotating angle θ' are determined. Based on the angle θ' , the robot moves to the next position s' and it is received a reward r . If s' is a state then the next default action is chosen again. Otherwise, an ε -greedy strategy is used to select the next action. The eligibility traces are updated in two steps. First, if a *non-greedy* action is taken, they are set to zero for all state-action pairs. Otherwise, the eligibility traces for all state-action pairs are decayed by $\gamma\lambda$. Second, the eligibility trace value of the current state-action pair is assigned to I . After reaching the goal position, the robot returns to its start position to begin a new episode. The algorithm terminates after a predefined number of episodes.

3 Experiments

In this section, computer simulations using the Matlab software are implemented to estimate the efficiency of the QVD(λ) algorithm. Besides, the Q-learning algorithm is also employed to implement our approach, called Q-learning based Vector Direction or QVD algorithm. The environments of these simulations are represented by the cells of a uniform grid. Each cell with a zero value is considered as a state of the environment. Otherwise, it is considered as an obstacle. The basic parameters of the all simulations are set as follows: $\alpha = 0.1$, $\gamma = 0.95$, $\lambda = 0.95$, $\varepsilon = 0.05$. After each episode, the value of ε is set again by $\varepsilon = 0.99\varepsilon$.

The first simulation environment consists of 47 states as shown in Fig. 2b. The task of the robot is to travel from the start position (S) to the goal position (G) as quickly as possible. The Fig. 2 depicts the results of this simulation. Figure 2a shows that the QVD(λ) algorithm converges after about 4 episodes and the QVD algorithm converges after about 10 episodes. In addition, the path found by the QVD(λ) algorithm is shorter than the path found by the QVD algorithm. Figure 2b and c show two paths found by the algorithms QVD(λ) and QVD after 100 episodes, respectively.

The second simulation environment is a maze as shown in Fig. 3b. The maze consists of $50 \times 50 = 2,500$ cells in which 20% cells make obstacles, so the number of states of the environment is 2,000 states. The task of the robot is to travel from the start position (S) in the bottom left corner to the goal position (G) in the top right corner of the maze. The simulation results are shown in Fig. 3. Figure 3a shows that the QVD(λ) algorithm converges after about 40 episodes and the QVD algorithm converges after about 140 episodes. At each episode the paths found by the QVD(λ) algorithm are shorter than the paths found by the QVD algorithm. Figure 3b and c depict two paths found by the algorithms QVD(λ) and QVD after 200 episodes, respectively.

Finally, to evaluate the QVD(λ) algorithm in a larger environment of states and obstacles, we design a maze consisting of $100 \times 100 = 10,000$ cells in which 20% cells make obstacles, so the number of states of the environment is 8,000 states. The task of the agent is to travel from the start position (S) in the bottom left corner to the goal position (G) in the top right corner of the maze environment. The Fig. 4 depicts the results of this simulation. Figure 4a shows that the QVD(λ) algorithm converges after about 160 episodes, but the QVD algorithm converges after about 260 episodes. At each episode, the paths found by the QVD(λ) algorithm are much shorter than the paths found by the QVD algorithm. It is clear that the QVD(λ) algorithm converges far faster than the QVD algorithm. Figure 4b and c depict two paths of the algorithms QVD(λ) and QVD after 300 episodes, respectively.

With the simulations implemented above, it can be concluded that the QVD(λ) algorithm guarantees to find a collision-free path and the path obtained is a near-optimality path. Besides, the QVD(λ) converges much faster than the QVD algorithm and the path found by the QVD(λ) algorithm is shorter than the path

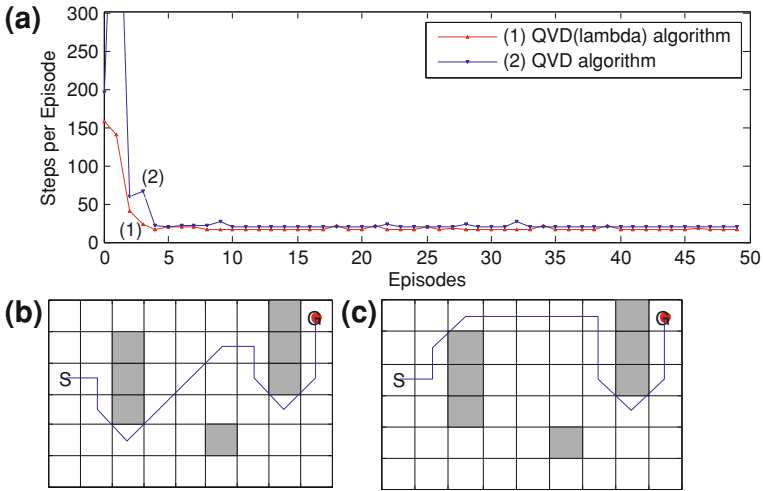


Fig. 2 **a** Comparison steps per episodes of algorithms QVD(λ) and QVD, **b** The path is found by the QVD(λ) algorithm after 50 episodes, **c** The path is found by the QVD algorithm after 50 episodes

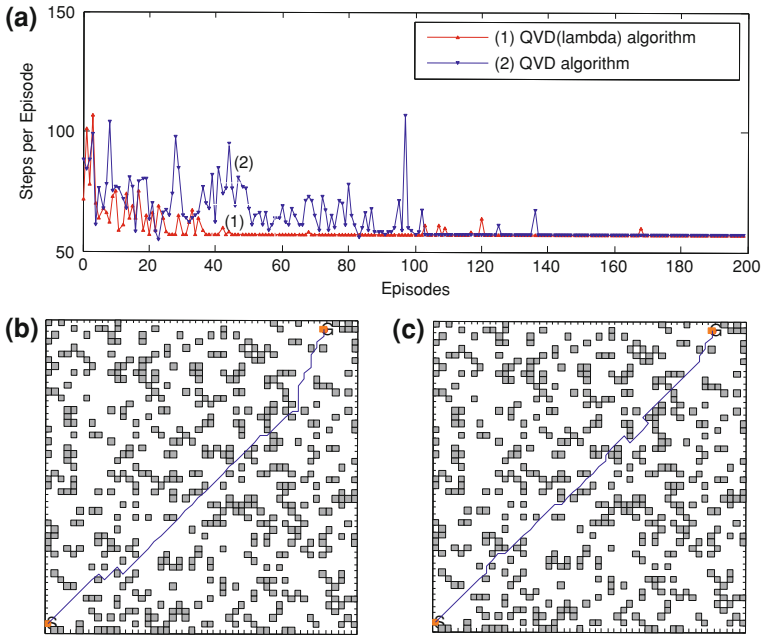


Fig. 3 **a** Comparison steps per episodes of algorithms QVD(λ) and QVD, **b** The path is found by the QVD(λ) algorithm after 200 episodes, **c** The path is found by the QVD algorithm after 200 episodes

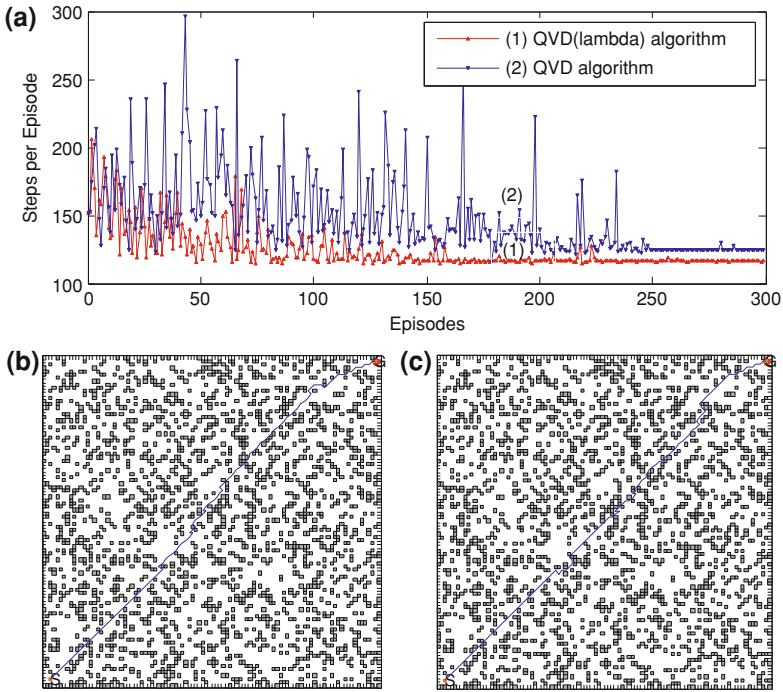


Fig. 4 **a** Comparison steps per episodes of algorithms QVD(λ) and QVD, **b** The path is found by the QVD(λ) algorithm after 300 episodes, **c** The path is found by the QVD algorithm after 300 episodes

found by the QVD algorithm. Therefore, the Q(λ) algorithm applied to our approach is efficient for the path planning problem of autonomous mobile robots.

4 Conclusions

In this paper, we propose a novel learning algorithm for the path planning of autonomous mobile robots, called the QVD(λ) algorithm. The proposed algorithm trains the robot to learn the vector directions of the robot’s positions and the goal position to find suitable moving directions for avoiding obstacles of the environment. Experimental results show that our approach is efficient for solving the path planning problem of autonomous mobile robots in an environment that the number of states and obstacles are so large. In addition, the learning speed of the algorithm QVD(λ) is much faster than the QVD algorithm. Simulation results demonstrate that our approach guarantees to find a collision-free path in a short time because the robot does not need to learn the vector direction if the next position of the current state is not an obstacle, but in some case it cannot find the shortest

collision-free path. However, we have just emphasized our study on the simulations of the maze world problem. We plan to apply the proposed approach to the real robot in the real environment. Besides, since the proposed approach provides a fast algorithm to solve the path planning problem. Therefore, it can be extended to the path planning problem of autonomous mobile robots in a dynamic environment [8].

Acknowledgments This research 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 (2010-0012609).

References

1. Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: a survey. *J Artif Intell Res* 4:237–285
2. Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. The MIT Press, Cambridge
3. Watkins C (1989) Learning from delayed rewards. Dissertation, Ph.D., King's College
4. Smart WD, Kaelbling LP (2002) Effective reinforcement learning for mobile robots. In: IEEE international conference on robotics and automation (ICRA'02), vol 4. IEEE Press, Washington, pp 3404–3410
5. Zamstein L, Arroyo A, Schwartz E, Keen S, Sutton B, Gandhi G (2006) Koolio: path planning using reinforcement learning on a real robot platform. In: 19th Florida conference on recent advances in robotics, Miami, May 2006
6. Chakraborty IG, Das PK, Konar A, Janarthanan R (2010) Extended Q-learning algorithm for path-planning of a mobile robot. LNCS, vol 6457. Springer, Heidelberg, pp 379–383
7. Vien NA, Viet NH, Lee SG, Chung TC (2007) Obstacle avoidance path planning for mobile robot based on ant-q reinforcement learning algorithm. LNCS, vol 4491. Springer, Heidelberg, pp 704–713
8. Mohammad AKJ, Mohammad AR, Lara Q (2011) Reinforcement based mobile robot navigation in dynamic environment. *Robotics Comput-Integr Manuf* 27:135–149

Registered Object Trajectory Generation for Following by a Mobile Robot

Md Hasanuzzaman and Tetsunari Inamura

Abstract This article presents an algorithm to generate trajectory of a visually localized object by linking centre points from successive image frames. In this method object is registered using a mouse and the system dynamically creates several templates of that object with different resolutions and slides those templates over the whole image and measures the matching score at every position. Based on predefined threshold of minimum distance object is localized and the centre position of that object is preserved. Using the coordinates of two consecutive centres of localized object, the system calculates the object movement in terms of number of pixels and direction in radian. Finally, the system maps the visual information with floor spaces where the robot will be moved. The algorithm is tested using a mobile robot where the robot follows the trajectory of a registered object.

Keywords Object registration · Object localization · Object trajectory generation · Trajectory following by robot

1 Introduction

To control robots or intelligence machines is one of the important research topics in recent year because robots are playing important roles in today's society, from factory automation to service applications to medical care and entertainment. With

M. Hasanuzzaman (✉) · T. Inamura
National Institute of Informatics (NII), 2-1-2, Hitotsubashi, Chiyoda-ku,
Tokyo 101-8430, Japan
e-mail: hasan@nii.ac.jp

T. Inamura
e-mail: inamura@nii.ac.jp

the advance in AI, Computer Vision and Object Recognition algorithm, the research is focusing not only the safest physical interaction, but also on a socially correct interaction. There are three basic types of robots: Manipulators, Mobile robots and hybrid robots. Mobile robots are especially necessary for tasks that are difficult, hazardous or dangerous for human. There are significant amount of researches on human-robot or human-intelligent machine interaction system [1–3] in recent years. Many researchers have already proposed and implemented successful mobile robot navigation systems so that the mobile robot can reach to the target point successfully. However if the target object moves freely it is very difficult to track that object because the speed and direction of the object varies over time. To adapt speed and direction of the target is one of the major problems for following object trajectory. Several algorithms are presented using active camera system to able to automatically calibrate itself to keep track of target object [4]. Tracking object is a complex task due to several reasons, such as noise in image, complex object motion, non-rigid nature of object, partial and full object occlusions, complex object shapes and scene illumination changes, etc. [4]. There are three broad classes for object tracking: point tracking (Kalman Filter [5]), Kernel tracking (Mean-shift [6]), and Silhouette Tracking (Hough Transform [7]).

The first step of following an object is to localize the object, then track that object by following its direction and displacement over time. Thus, the robot needs to localize the object using its vision system or another client PC should do that and transfer to the robot. There are several approaches for object localization. One of the widely used method is template matching approach-where small part of the image is matched with a template image [8]. This approach is subdivided into two approaches: feature-based approach and template-based matching. The feature-based approach uses the feature of the search and template image, such as edge and corners, as the primary match measuring metrics to find the best location of the template in the source image. To reduce the number of features, the SIFT (scale invariant feature transform) descriptor is widely used method for matching image features [9]. In this method, SIFT keypoints of objects are first extracted from a set of reference images and stored in a database. An object is recognized in a new image by individually comparing each from the new image to database and finding candidate matching features based on Euclidean distance of their matching vectors. However, perfect scale invariance cannot be achieved in practice because of sampling artifacts, noise in the image data and computational cost [10]. For template without strong features a normal template-based approach is effective but this matching may potentially require sampling of large number of points, it is possible to reduce the number of sampling points by reducing the resolution of the search and template images by the same factor. Multiresolution templates or multiresolution image pyramids is one of the effective tools to detect object if the size of the object varies due to variation of distances between object and camera.

Many researchers used object tracking algorithm in robotic applications. Sang-joo Kim et al. proposed and implemented a tracking and capturing a moving object using a mobile robot [11]. In this work they estimated the position of the target based on the kinematic relationship of consecutive image frames and estimated the

movement of the target object using Kalman filter for tracking. Based on estimated trajectory they used motion planning of a mobile robot to capture the target object. Min-Soo Kim et al. also used kalman filter for trajectory estimation of a moving object that are used for robot visual servo control system [12]. Bhuiyan et al. used template-based eye detection method and measured optical flow to track the eye ball and utilized it to control robot action [13].

However, all of the above studies did not consider dynamic selection of object as well as did not use angular movement and linear displacement of an object to control a mobile robot path. The goal of this research is to follow the selected object trajectory using a mobile robot. In this system the selected object is localized or detected using multi-resolution template-based template matching approach from a real-time capture images by a client PC camera. Using the coordinates of the consecutive centers of a localized object the system calculates the object angular movement and object displacement. Then the system generates the trajectory by linking center points of the localized object. The system proposes visual space to floor space mapping algorithm so that a robot can follow visual trajectory.

2 Proposes System Description

Figure 1 shows the proposed system architecture of an object trajectory generation system for following by a mobile robot. The system is capable of registering new object and making trajectory of the registered object by linking successive center points of the localized object. The system is also able to control robot to follow the object trajectory by mapping visual space to floor space where the robot will be moved. This system uses object angular movement and displacement in visual space to control robot. Following subsections describe each module briefly.

2.1 Object Registration

The system uses standard CCD camera to capture the real-time image. It captures 30 image frames per second with RGB color and the resolution of the image is (640×480) . For localization the system should know the object or register the object. The algorithm for new object registration is described below;

Step 1: Capture the image using a single camera and show RGB image on the display. The source image is defined by $S(M \times N)$, where M and N represents image width and height respectively.

Step 2: Select any object using left button of mouse (click four points that bounded that object as rectangular).

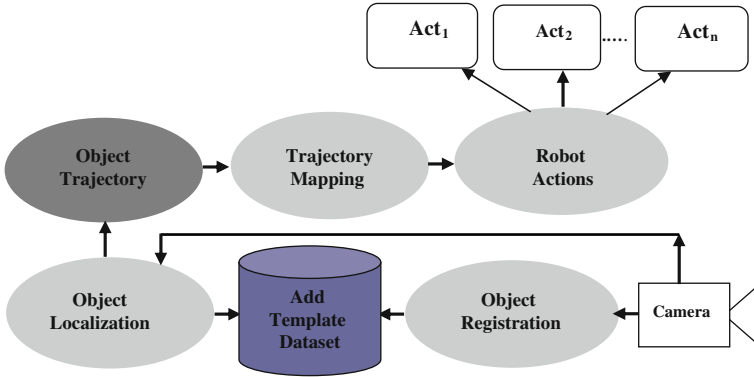


Fig. 1 Proposed system architecture

Step 3: The system crops the marked area and saves as template image. The template image is defined by $T(M' \times N')$ where M' and N' represents the template image width and height respectively ($M' < M$, $N' < N$).

Step 4: If the user types a name of that object the system registers it as a known object and saves the template image and object name in the knowledge base otherwise ignore it.

Step 5: After registering an object the system considers that object as known object.

The system uses on-line interactively registering approach, so that user can register new object in interactive manner selecting new object. For single object tracking only the last registered object is used but for multiple objects tracking all the registered objects will be useful.

2.2 Object Localization

After registering the object the system detects that object using multi-resolution template matching approach. The detail algorithm is as follows;

Step 1: Read the source image, $S(M \times N)$.

Step 2: Read the last registered template image $T(M' \times N')$ and create multiple templates of that object with different resolutions. In this work we have used 11 templates of an object with different resolutions.

Step 3: Slide each template over the image and calculate minimum Euclidian distance based on template matching approach.

Step 4: Calculate normalized minimum Euclidian Distance for each template size. Here a template image is compared with source image to find the area most similar to the template. The OpenCV function `cvMatchTemplate` is used to do the matching [14]. This function provides minimum Euclidian distance (d_i for i -th template)

and location (x, and y coordinates of the minimum distance area) of the probable match area. The normalized minimum Euclidian distance is defined by,

$$\hat{d}_i = \frac{d_i}{M'_i N'_i}$$

where, \hat{d}_i is normalized minimum Euclidian distance of i -th template, M'_i and N'_i represents the i -th template image width and height respectively.

Step 5: Find the minimal among the all minimum normalized Euclidian distance and if it is less than predefined threshold then the object is detected and bounded by the size of the mask that best match.

Step 6: Preserve the center position of the rectangular box as the location of the object.

2.3 Object Trajectory Generation

A trajectory can be described mathematically either by the geometry of the path or the position of object over time. The first step is to locate the object and calculate its center position (c_x, c_y). Once the object is localized that means its position is determined in each frame, the tracking algorithm traces the object from frame to frame. This article describes object trajectory generation method based on object position over time. The system calculates the displacement of the object between two successive positions as well as the direction of movement. To calculate displacement we consider that the object trajectory between two successive frames is straight and to measure the direction of movement we calculate slope of a straight line. To measure distance we did not consider the acceleration rate [15], we simply calculate the distance between two points of straight line.

Suppose, the detected object in i -th image frame is surrounded by a rectangle PQRS. Where, $P(x^l, y^l), Q(x^h, y^l), R(x^h, y^h)$ and $S(x^l, y^h)$ represents the 4-vertex points of the surrounded rectangle. In the $(i + 1)$ -th image frame the object may move any position from its original position, or even move out of tracking domain. Supposed the detected object in the $(i + 1)$ -th frame surrounded by another rectangle P, Q, R, S. Where, $P(x^{l'}, y^{l'}), Q(x^{h'}, y^{l'}), R(x^{h'}, y^{h'})$ and $S(x^{l'}, y^{h'})$ represents the 4-vertex points of this surrounded rectangle. The trajectory is determined using following steps.

Step 1: Calculate center point of the object (center of the rectangle C') in i -th image frame using equation,

$$C = \left(\frac{x^l + x^h}{2}, \frac{y^l + y^h}{2} \right)$$

Step 2: Calculate center of the object (center of the rectangle C') in the $(i + 1)$ -th image frame using equation.

$$C' = \left(\frac{x' + x^{h'}}{2}, \frac{y' + y^{h'}}{2} \right)$$

Step 3: Calculate the horizontal (δx) and vertical distance (δy) changes between two center points C and C'

$$\delta x = \left(x' + x^{h'} - x^l - x^h \right) / 2$$

$$\delta y = \left(y' + y^{h'} - y^l - y^h \right) / 2$$

Step 4: Calculate the direction of movement by finding slope m , and direction θ .

$$m = \frac{\delta y}{\delta x} \quad (\delta x \neq 0)$$

$$\theta = \tan^{-1}(m) * 180 / 3.14$$

If $\delta x = 0$, then m is infinite that means object is moving vertically, in that situation m is define as 90. If m is zero that means the object is moving horizontally and if both δx and δy are zero that means the object is static.

Step 5: Calculate the moving distance or object displacement over time,

$$L = \sqrt{\delta x^2 + \delta y^2}$$

Step 6: Draw a line to connect two consecutive center points and display the output image with trajectory (with all the connecting points).

2.4 Mapping Visual Data to Floor Space for Robot Movement

In this system we show path using selected object trajectory and robot will follow that path. To implement this work we need to map visual data (direction and distance) to floor space where the robot will be moved. This system uses mobile robot named "PeopleBot". The robot can freely move in the floor space and we can control using client-server architecture. The system uses ARIA (Activmedia Robotics Interface for Applications) open source API to tele-operate the robot. We consider a robot is placed in such a position where at least 128 cm floor space in all the surroundings. Object direction is considered as the robot direction of movement and assume that robot will move 1 cm in the floor space if the object moves 5 pixels. The detail algorithm is given bellow.

Step 1: Run the robot server program and connect with client PC.

Step 2: The evaluated direction of the object movement and distance of movement is estimated by visual analysis in a client PC and send it robot server. Robot server receives it and activates movement function that we have been designed based on (θ, L) . Here θ represents direction in terms of radian and L represents distance in terms of pixels.

Step 3: Robot will rotate based on direction information and wait 100 ms.

Step 4: Robot will progress 1 cm for 5 pixels movement of the visual object and wait 100 ms.

Step 5: Go to step 2 or stop when the object is stopped or out of scene.

3 Experiments and Results

The experimental results are summarized in two parts. In part one we have presented the result of object trajectory generation method and in the second part we have discussed a human-robot interaction scenario where a mobile robot follows an object trajectory.

3.1 Result of Trajectory Generation

For performance analysis the system uses three video clips of three objects. A single camera captures those videos with a capture rate of 30 fps and resolution of the capture image is 640×480 pixels. The client PC is used Intel(R) Core(TM) 2 Duo CPU with 3.06 GHz clock speed and 3GB of RAM. The system uses OpenCV functions for object localization and tracking [14]. Figure 2, shows the several sample visual outputs of object localization and trajectory generation method with image sequence number. Table 1 presents the accuracy of object localization method for three objects.

Each video is 2 s long and has 60 frames where some frames do not have required object and in some frames object is presented but due to pose variation the system could not located the required object. In case of video clip 1, among 60 frames there are 36 frames with 'Pink-ball' and 36 are localized because the object is circular and color is uniform and unique. In case of video clip 2, among 60 frames there are 38 frames with 'Duster' which is rectangular and uniform color. Among them 36 are properly localized and the system could not localized 2 of them due to partial presence (Example Frame#8 in Fig. 2b). In case of video clip 3, there are 52 frames with 'Hand-palm' and among them 36 are properly localized and the system could not localized 16 of them due to partial presence and changes of pose (Example Frame#59 in Fig. 2c). The accuracy of the object localization method is presented in Table 1. In Table 1, "# Presence" represents the total number of frames with required object, "# Localize" represents the total number frames where object is localized and "Localization Accuracy" represents the ratio



Fig. 2 Example outputs of trajectory generation method. **a** Trajectory of a “Pink Ball” (clip #1). **b** Trajectory of a “Duster” (clip #2). **c** Trajectory of a “Hand Palm” (clip #3)

Table 1 Evaluation of object localization method

Video clip #	# Presence	# Localize	Localization accuracy (%)
1 (Pink-ball)	36	36	100
2 (Duster)	38	36	94.73
3 (Hand palm)	52	36	69.23

of total number of frames where the object is localized to total number of frames with required object in a video. From the table (3rd row) we can see the limitation of single view template based object localization method for hand palm. This happen due to uses of single viewed template. This limitation will be overcome if we take multi view poses for an object to create template. This method is suitable for tracking objects whose pose do not vary considerably during the course of tracking. We can handle pose variation if we use adaptive mean-shift based tracking algorithm or others algorithm that used histogram matching-based object localization method or not rely on object shape.



Fig. 3 Object trajectory and robot initial position. **a** Object trajectory. **b** Robot initial position

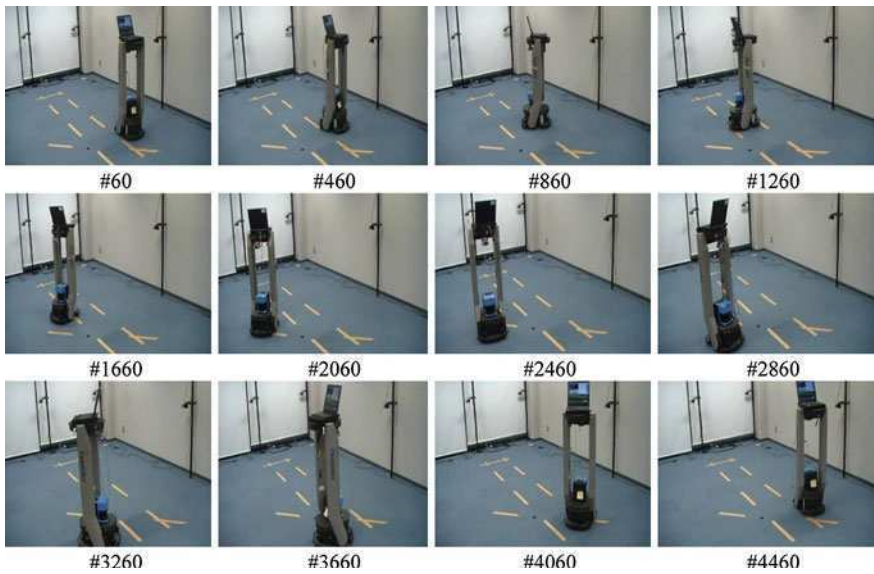


Fig. 4 Sample robot movement sequences related to object trajectory

3.2 Robot Following Object Trajectory

This system uses a mobile robot named ‘PeopleBot’ to implement the human-robot interaction scenario. A Client PC captures and processes the image and sends direction and displacement information of the moving object based on visual analysis to a robot server. Robot server executes the command for rotating a robot and progress an estimated distance based on mapped data. Figure 3 shows the

example output frame of an object trajectory (Fig. 3a) and initial position of a mobile robot (Fig. 3b). After executing each command either rotation or progress the robot sleep 100 ms that means 200 ms delay is required for each frame.

To compare robot trajectory with object trajectory we took a video image of robot movement during following the trajectory and its duration is about 168 s. Figure 4 shows several sequences of image frames of robot movement to follow the trajectory of an object. The video image is converted to sequence of image frames (30 frames per second) and it produces 5040 image frames. In Figure 4, we have shown 12 image frames to show the robot movement regarding corresponding trajectory. In the floor there are several marked lines and by relating robot position with those lines the reader can guess the robot trajectory. In this experiment we consider that there is no obstacle in this space. If obstacle is present the robot will stop. In our near future work we will consider obstacle in robot space and modify our algorithm to adapt robot path. If robot will find obstacle the system will calculate the new slope value and distance with the next points (skip the current points) and this process will be continued until the robot can avoid the obstacle.

4 Conclusion

In this article, we discuss an algorithm to generate trajectory of a registered object for following by a mobile robot. In this algorithm, we use dynamically created multi-resolution templates to locate a moving object. By linking the center points of the registered object over time the system generates object trajectory and a robot follows that trajectory. The system can dynamically include new object in the template database if user select the object by mouse and type a name using keyboard. The algorithm is tested by implementing a human-robot interaction scenario with a mobile robot where robot follows the trajectory of an object. The major advantage of the system is that it uses angular movement and linear displacement of an object to control a mobile robot path so robot path is identical to object trajectory. In this system object localization is faster since it uses only few templates of the last registered object and multi-resolution templates reduces the effect of object to camera distance. Robot trajectory is precise since the system provides direction and distance precisely. As we discussed in the experiment and result section, the system could not locate the non-circular object precisely if its pose was changed due to movement or partial occlusion. This algorithm did not consider the presence of obstacle in front of robot path. The remaining issue of this work is to develop more robust object tracking method as well as handle obstacle when robot observe it for following object trajectory.

References

1. Bartneck C, Okada M (2001) Robotic user interface. In: Hc'01 Human and computer conference, Aizu, pp 130–140
2. Fong T, Nourbakhsh I, Dautenhahn K (2003) A survey of socially interactive robots. *Robotics Auton Syst* 42(3–4):143–166
3. Hasanuzzaman M, Zhang T, Ampornaramveth V, Ueno H (2006) Gesture-based human–robot interaction using a knowledge-based software platform. *Int J Ind Robot* 33(1):37–49
4. Yilmaz A, Javed O, Shah M (2006) Object tracking: a survey. *ACM Comput Surv* 38(4): article 13
5. Broida T, Chellappa R (1986) Estimation of object motion parameters from noisy images. *IEEE Trans Pattern Anal Mach Intell* 8(1):90–99
6. Comaniciu D, Ramesh V, Meer P (2003) Kernel-based object tracking. *IEEE Trans Pattern Anal Mach Intell* 25:564–575
7. Sato K, Aggarwal J (2004) Temporal spatio-velocity transform and its application to tracking and interaction. *Comput Vis Image Underst* 96(2):100–128
8. Brunelli R (2009) *Template matching techniques in computer vision: theory and practice*. Wiley, New York. ISBN 978-0-470-51706-2
9. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. *Int J Comput Vis* 60:91–110
10. Yan Cui, Nils Hasler, Thorsten Thormahlen, Hans-Peter Seidel (2009) In: *ICIAR'09*. Springer, Berlin, pp 258–267
11. Kim SJ, Park JW, Lee JM (2005) Implementation of tracking and capturing a moving object using a mobile robot. *Int J Control Autom Syst* 3(3):444–452
12. Kim M-S, Koh J-H, Nguyen HQP, Kang H-J (2009) Robot visual servo through trajectory estimation of a moving object using Kalman filter. In: Huang D-S et al. (eds.) *ICIC 2009*, LNCS, vol 5754. Springer, Berlin, pp 1122–1130
13. Bhuiyan MA, Ampornaramveth V, Muto S, Ueno H (2004) On tracking of eye for human–robot interface. *Int J Robotics Autom* 19(1):42–54
14. Bradski G, Kaehler A (2008) *Learning OpenCV, computer vision with the OpenCV library*. O'Reilly Media, ISBN: 978-0-596-51613-0
15. Wang Y, Doherty JF, Van Dyck RE (2000) Moving object tracking in video. In: *IEEE applied imagery and pattern recognition workshop*, pp 95–101, Washington

An Improved Algorithm for Constrained Multirobot Task Allocation in Cooperative Robot Tasks

Thareswari Nagarajan and Asokan Thondiyath

Abstract This paper presents an improved algorithm for solving the complex task allocation problem in constrained multiple robot cooperative tasks. The existing multiple robot task allocation mechanisms do not discuss much about complex tasks, instead they treat tasks as simple, indivisible entities. Complex tasks are tasks that can be decomposed into a set of subtasks and so can be executed by several possible ways. The goal of cooperative task allocation algorithm for multiple mobile robots is to find which robot should execute which task in order to maximize the global efficiency and minimize the cost. Some factors such as benefit, cost, resources, and time should be considered during the course of task allocation. The meta-heuristic algorithm proposed here solves the task allocation problems with the characteristics like each task requires a certain amount of resources and each robot has a finite capacity of resource to be shared between the tasks it is assigned. The cost of solution which includes static costs when using robots, assignment cost, and communication cost between the tasks if they are assigned to different robots are also taken into account in developing the solution. A peer search scheme algorithm for solving the constrained task allocation problem is presented. Computational experiments using this algorithm have shown that the proposed method is superior in terms of computation time and solution quality.

Keywords Mobile robots · Multiple robot cooperative tasks · Task allocation · Heuristic algorithm · Peer structure

T. Nagarajan · A. Thondiyath (✉)
Robotics Laboratory, Department of Engineering Design, Indian Institute
of Technology Madras, Chennai 600036, Tamilnadu, India
e-mail: asok@iitm.ac.in

T. Nagarajan
e-mail: Nagarajan.thareswari@gmail.com

1 Introduction

Due to its outstanding flexibility, robustness, and autonomy, multiple robot systems that can carry out complex tasks which cannot be performed by a single robot is finding applications in many industrial fields in recent years [1]. Task assignment is one of the key modules in a multi-robot system and it addresses the issue of finding a task-to-robot assignment that achieve some system objectives such as improving efficiency, saving cost, optimizing the global utility, or rationalizing the distribution of resources. The task allocation problem is to assign a set of tasks to a set of robots so that the overall cost is minimized. This cost may include a fixed cost for using a robot, a task assignment cost, which may depend on the task and robot, and a communication cost between tasks that are assigned to different robots. The task assignment problem can be constrained or unconstrained; depending on whether or not the robots have capacity to be shared between the tasks they are assigned. This problem arises in multiple mobile robot systems, where a number of tasks are to be assigned to a set of robots to guarantee that all tasks are executed within a certain cycle time. The aim is to minimize the cost of using the robots as well as the inter-robot communication bandwidth. In this paper we propose an algorithm, based on a peer structure group search scheme, for solving the constrained task allocation problem. The results of the computational experiment using the algorithm show that the proposed scheme outperforms the other methods.

2 Task Allocation Problem

Based on the nature of task availability, multiple robot task allocation problems are classified into static and dynamic allocation problems. If the tasks are known to the robot before execution, then it is referred as static [2]. If it is made known to the robot during execution then it is termed as dynamic task allocation [3]. In threshold based task allocation, each robot has a threshold for each task, and pheromone is used to reflect the urgency or importance of tasks. Typical multi-robot system which uses this method is ALLIANCE [4] and its corresponding system with parameter leaning capacity is named as L-ALLIANCE [5]. Threshold based method [6] is used in multi-robot system with many simply functioned robots and this needs very little communication but has low efficiency. The basic idea of market-based approaches is to facilitate task allocation through contract negotiations; single task or combinatorial tasks are auctioned in such an approach.

Market based task allocation [7] which needs very little computation fits for systems with a large number of unknown quantities of selfish subsystems. It is convenient for increasing or decreasing subsystems dynamically [1], however, it requires much communication and cannot promise an optimal solution [8]. Some distinctive task allocation approaches using artificial intelligent techniques and intelligent computation algorithms to deal with some characterized tasks in different

fields can be seen in [9]. When there is more than one robot it resembles a Multiple Travelling Salesman problem wherein all the robots would have to be used in a cost effective manner to serve the tasks by making proper utilization of the resources (available robots). It is emphasized that in multiple travelling salesman problems the less addressed criterion is the balancing of workloads amongst salespersons [10]. It is evident from the literature that there are many multiple robot task allocation methodologies available. However, there are not much literature which discusses the allocation problems in complex tasks, dynamic and constrained task allocation, and balanced utilization of the tasks among the group of robots. Moreover, most of the researchers concentrated on minimizing the total distance traveled by the robots rather than looking at the overall utilization and efficiency of allocation. Our work is more concerned with the utilization of the robots and thus we consider the problem as constrained task allocation problem. The objective is to assign a set of tasks to a set of robots so that the overall cost is minimized.

3 Problem Formulation

The objective of this constrained task allocation problem is to minimize the total allocation cost for optimality which in turn provides the proper utilization of the robots present in the system. The idea behind developing the formulation is to provide a task allocation algorithm that not only produces the optimal solution but also flexible enough to practical problems. The following assumptions are made while developing the allocation strategy:

- Tasks are separable and can be sorted with priority.
- The task should be executed by any robots present in the system with the needed capability to execute the respective task.
- The execution cost of task which has restriction for any particular robot has to be infinite.
- Task pre-emption is not allowed, i.e. once a robot begins to execute a task, it must continue to completion without interruption.
- All relevant parameters to the task allocation problem are known in advance.
- Prior information about all available robots are available

Consider there are N tasks to be allocated between M robots present in the system. The allocation cost is calculated by summing the communication cost (C), static cost (S), and execution cost (E). The communication cost, denoted by C_{ij} , is the cost incurred if the task i and j are assigned to two different robots and the tasks have some information which has to be exchanged between them. It is assumed to be independent of robots. If there is no dependency between tasks, then it assumes the value zero. Execution cost, denoted by E_{ij} is the cost for executing the i -th task on j -th robot and this becomes zero if the task cannot be executed by a particular robot. The static cost, S_k , is defined as the cost for using the k -th robot for any task, where $k = 1, \dots, M$. The size or capability of the robots are represented by b_k where $k = 1, \dots, M$. The task

requirement is given by a_i where $i = 1, \dots, N$. Since the objective here is to reduce the total allocation cost, the objective function is defined as

$$[Min]Z = \sum_{k=1}^M S_k.y_k + \sum_{i=1}^{N-1} \sum_{j=1}^N C_{ij} \cdot \left(1 - \sum_{k=1}^M X_{ik}.X_{jk} \right) + \sum_{i=1}^N \sum_{k=1}^M E_{ik}.X_{ik} \quad (1)$$

Subjected to:

$$\sum_{k=1}^M X_{ik} = 1 \quad i = 1, \dots, N \quad (2)$$

where $X_{ik} \in \{0,1\}$ indicates whether task i is assigned to robot k ($i = 1, \dots, N$; $k = 1, \dots, M$). The following constraints describe the conditions needed to compute an optimal task allocation schedule.

$$\sum a_i.X_{ik} \leq \sum b_k.y_k \quad k = 1, \dots, M \quad (3)$$

where, y_k indicates whether any task is assigned to robot k ($k=1, \dots, M$) and

$$X_{ik}, y_k \in \{0, 1\}, \text{ for all } i, k$$

The first equation is to minimize the allocation cost. The second equation imposes that each task is to be assigned to one and only one robot and the third equation imposes the capacity constraint. The multiple robot task allocation problems becomes strongly Non-Deterministic Polynomial-time hard (NP-hard) if the number of robot is greater than three [11]. Hence, some kind of meta-heuristic procedure needs to be developed for dealing with the problem and finding near-optimal solutions.

4 Algorithm Development

The approach considered here for the constrained task allocation is shown in Fig. 1. Here the mission or application is divided into subtasks based upon their computational criteria, and then the communication needed between them is designated, paving way for the parallel computations and proper utilization of robots. This task model, along with the robot model, is given to the task allocation algorithm for finding out the optimal allocation scheme.

The task allocation algorithm starts with an initial solution created by sorting the static costs of robots in an ascending order and then assigning the tasks to the robots after satisfying the constraints. Based on this initial solution, a set of peer structure search moves are explored to find the optimal solution. The cost function described in the previous section is used at every search to identify an optimal solution. The peer search moves keep reallocating or exchanging the tasks to other robots. The flow diagram of the peer structures is shown in Fig. 2.

Fig. 1 Approach for task allocation strategy

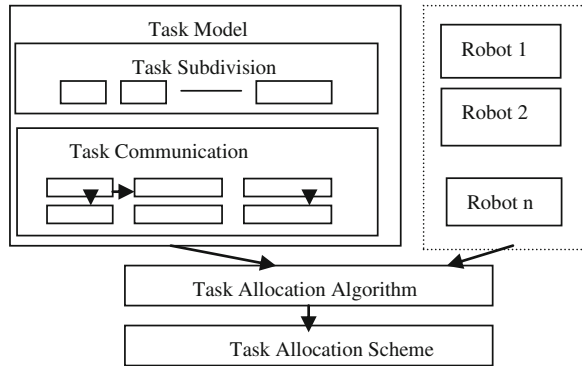
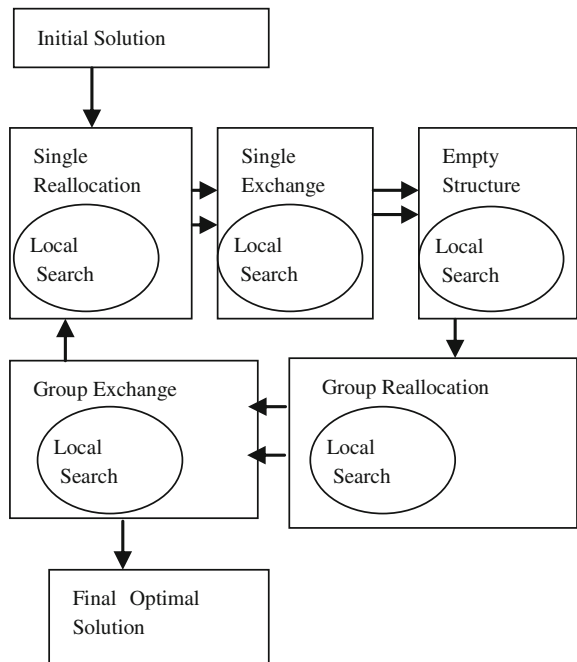


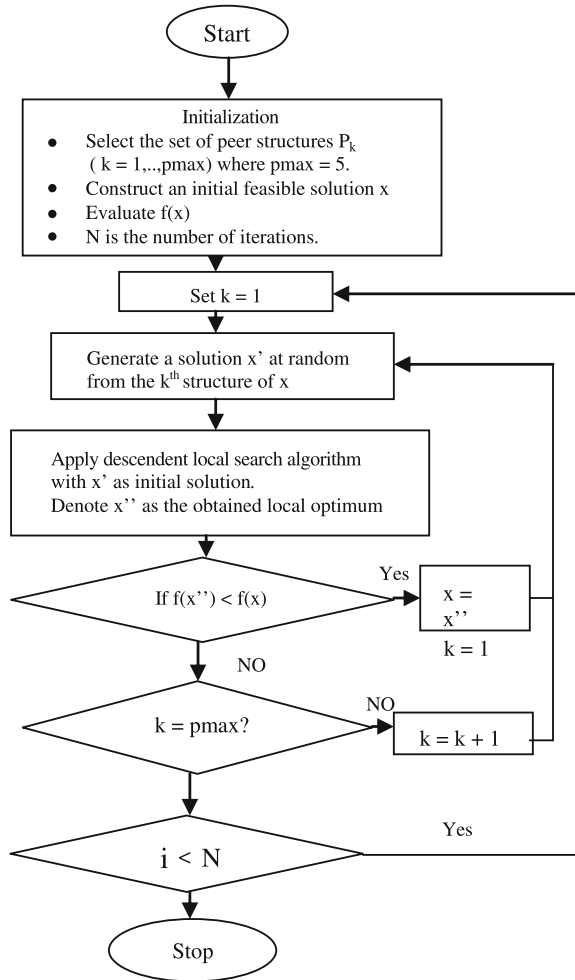
Fig. 2 Peer structure based search strategy



A meta-heuristic algorithm, based on search method is proposed here for peer search move. A solution is referred as x , and $f(x)$ is the cost of the solution x , where x belongs to the search space. The algorithm used is given in Fig. 2.

Peer Structure: As the peer structure significantly affects the solution quality, it is necessary to clarify how the peer structure is defined. Let ‘S’ be the set of all defined moves and x is the initial allocation solution. We use $P(x)$ to denote the peer structure of x , i.e., the subset of moves in S applicable to x . For any move s belongs to $P(x)$, the new solution obtained by applying move s to x is called a peer structure of x . None of the peer structure moves allow non-feasible solution as

Fig. 3 Flow chart for peer structure algorithm



the solution x is the feasible solution taken from the feasible permutations of the solution space. Hence it is guaranteed that the algorithm will always yield a feasible solution. Local search in such a peer structure is defined as performing allocation for all feasible (peer structure) moves of the given initial solution x . The following types of peer structure moves are considered: (1) reallocating a task from one robot to another robot, (2) exchanging a task from one robot to another robot, (3) emptying a robot by reallocating its assigned tasks to other robots, (4) reallocating a group of tasks from one robot to another, and (5) reallocating a group of tasks from different robots to one robot. The first two moving schemes are simple moves between the robots, but the rest are complex moving schemes available in the peer structure.

As shown in Fig. 3, the algorithm will start the search using the peer structures, P_k . In this case, we have five structures as shown in Fig. 2. They are ordered in this

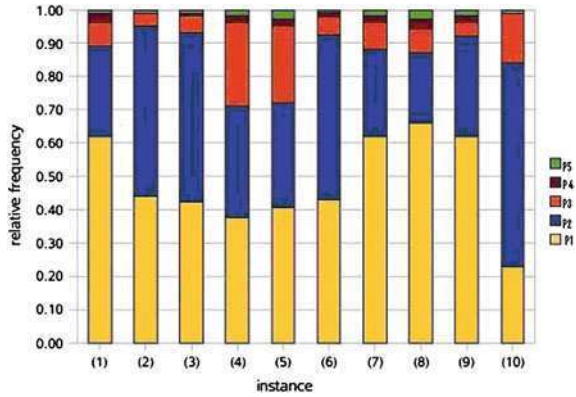
fashion (1) $P_1(x)$ -reallocate a task i from robot k to robot l . (2) $P_2(x)$ - exchange two tasks (task i from robot k to robot l and task j from robot l to robot k). (3) $P_3(x)$ -reallocate a group of tasks from robot k to robot l . (4) $P_4(x)$ -reallocate a group of tasks from different robots to robot l . (5) $P_5(x)$ -empty robot k .

The greedy algorithm for task allocation is used for finding the initial solution. The greedy algorithm sorts the robots by increasing the static cost of the robot and then tasks are divided into subtasks. The algorithm finds the cost for allocating robots in the ordered way by keeping the capacity of the robots and the requirement of the tasks as the constraint criteria. The task that minimizes the cost function for the given robot will be allocated with the respective task. Then the algorithm does this with the other unassigned tasks. The greedy algorithm is given as algorithm 1. A descendent local search algorithm is used to find the optimal solution for each peer structure. The descendent local search algorithm finishes when no improvement is obtained, which yields a solution that is a local optimum for the peer group moves. The algorithm used for the local search is described as algorithm 2.

Algorithm 1	Algorithm 2
<p>Beta = random number $\in [0-1]$ Initially, $R_k = \{\emptyset\}$, $k=1, \dots, m$ Sort robots by increasing the static Cost, k is the first robot while (there are non-assigned tasks) do P is the set of non-assigned tasks p that $a_p = b'_k$ where b'_k is the remaining capacity of the robot and a_p is the requirement of the task While ($P \neq \{\emptyset\}$) do Compute C_j t is the task that minimize C_j; add t to robot k; $P_k = P_k + t$ and $b'_k = b'_k - a_t$ Determine J (set of non assigned tasks j that $a_j = b'_k$) End While Go to next robot, k End While</p>	<p>Generate an initial solution, x, Evaluate ($f(x)$) While (no final condition) do While ($k \leq pmax$) do Randomly choose a solution for $P_k(x)$, as x' x'' is the result of applying local search to x' If $f(x'') < f(x)$ then $x = x''$ and $k = k + 1$ else $k = k + 1$ End While End While Return best found solution as x''</p>

Efficiency of Peer structure: When applying peer structure search algorithm, the peer structure are mostly ordered by increasing complexity and afterwards this static peer structure order is applied during the whole search process. As the number of tasks associated with a move increases, the efficiency of this moving scheme decreases. However, the quality of the solution obtained by using complicated moving scheme may be better. There is usually a tradeoff between efficiency and the quality of the solution. The relative frequency of the peer structure is tested for the various instances and the results shows that the first two peer

Fig. 4 Relative frequencies of peer structure



structure reallocating a task from one robot to another robot (P1) and exchanging a task from one robot to another robot (P2) having the highest occurrence, which accords with the static order of the peer structure shown in the Fig. 4. The relative frequency of a peer structure is the overall number of contribution to a solution improvement of a peer structure divided by the total number of solution improvements of all peer structure.

5 Simulation Results

The algorithm presented above has been verified through computer simulations using open source software and optimizer. The objective of the computational experiment was to evaluate the effective working of the algorithm. Experiment was conducted for various instances to check the feasibility of the algorithm. The results of the simulation for the experiment are explained below.

Sample Problem: A task allocation problem which consists of five subtasks and three robots is considered here with two instances. The task requirements range from a few up to 40 units and robot capabilities range from 50 to 90 units respectively. The static cost ranges from 1 to 3 units and the communication cost matrices are very dense with C_{ij} ranging from a few to 100 units. Task 3, 4 and 5 are dependent on each other, i.e. these tasks are cooperative and needs information exchange between them. The task 1 and task 2 are independent i.e. there is no information exchange happens between these two tasks. The communication cost (in terms of bandwidth) between the tasks is listed in Table 1. The robot capabilities and task requirements are listed in Tables 2 and 3. The execution cost for the robots on the tasks are given in Table 4 and the assigning cost is shown in Table 5.

The initial solution is found by ordering the static cost of the robots and then the allocation scheme is found using the greedy algorithm as shown in Fig. 5.

This initial solution is used by the algorithm to explore the peer structures by applying the local search. Here the algorithm allocates/exchanges tasks as per the

Table 1 Communication cost for the tasks

C	T1	T2	T3	T4	T5
T1	0	0	0	0	0
T2	0	0	0	0	0
T3	0	0	0	100	30
T4	0	0	100	0	40
T5	0	0	30	40	0

Table 2 Task requirement

Task Requirement	T1	T2	T3	T4	T5
	37	36	35	25	22

Table 3 Robot capabilities

Robot Capability	R1	R2	R3
	88	83	56

Table 4 Execution cost of the robots

E	R1	R2	R3
T1	1	3.2	2.8
T2	2	4	3
T3	3.5	3.5	0.5
T4	1.6	2	1.1
T5	2.1	1.6	0.7

Table 5 Assigning cost for the robots

R1	R2	R3
1	3	4

peer structure and evaluates the cost at every stage and finds the optimum solution with all the constraints satisfied.

Table 6 shows the final allocation of tasks for the given problem. As shown here, tasks 3, 4, and 5 are assigned to Robot 2 (R2). This minimizes the communication cost as all the dependent tasks are executed by the same robot. Since R2 has a capability of 83, all the three tasks can be executed on it as the total requirement for these tasks are only 82. Figure 6 shows the screen shot of the results obtained using the algorithm. In order to verify the algorithm effectiveness, the previous problem was modified slightly in term of robot capabilities as shown in Table 7, and run again.

The results of the task assignment are shown in Figs. 7 and 8. The initial assignment shows that tasks 1, 2, and 3 are assigned to R1 and 4 and 5 are assigned to R2. The final assignments shown in Table 8 clearly indicates the capability of R1 to carry out tasks 2–5, thus reducing the communication cost and maximizing the robot utility.

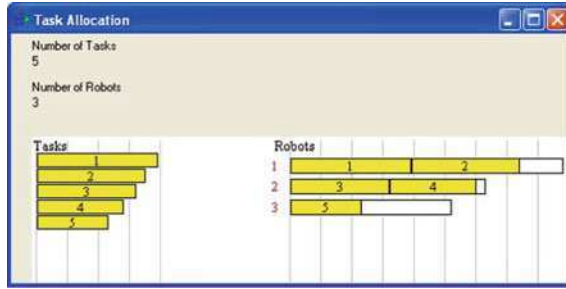


Fig. 5 Initial solution for the problem

Table 6 Final assignment

	R1	R2	R3
T1	1	0	0
T2	1	0	0
T3	0	1	0
T4	0	1	0
T5	0	1	0

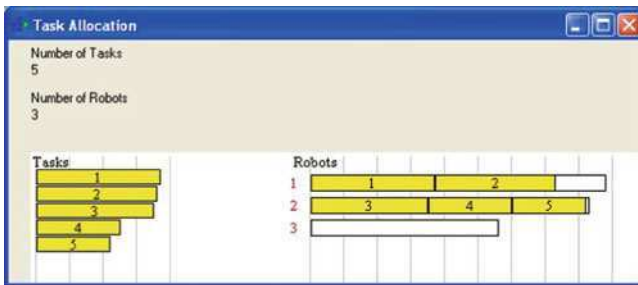


Fig. 6 Solution for problem using the peer structure scheme

Table 7 Robot capabilities

Robot	R1	R2	R3
Capability	118	60	73

Fig. 7 Initial solution using greedy method

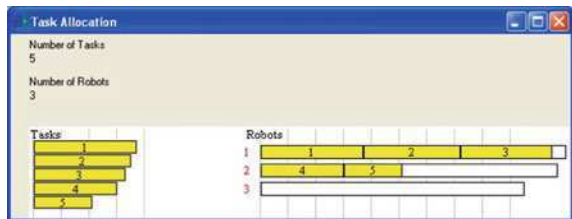


Fig. 8 Solution using peer structure scheme

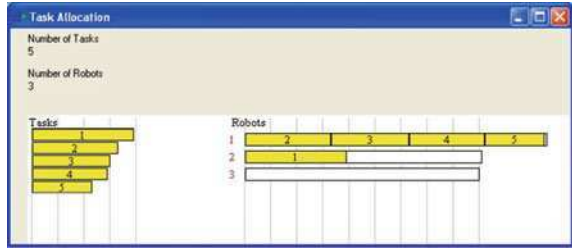


Table 8 Final assignment

	R1	R2	R3
T1	0	1	0
T2	1	0	0
T3	1	0	0
T4	1	0	0
T5	1	0	0

The improvement offered by the proposed algorithm is greater in situations where the number of required robots (on average) is greater than the number of available robots. This is not surprising, as these are exactly the cases in which it is possible to take greater advantage of the complicated moving schemes in the peer structures. The remaining capacity of the robots may be low, and it may be very difficult to reallocate a group of tasks to a robot or to empty a robot, which is exactly what is done in peer structure 3, 4 and 5.

6 Conclusions

This paper has presented an algorithm for solving task allocation in multiple robots, which are using cooperative and complex tasks based on peer search scheme using search algorithms. The algorithm has been designed for both complex and cooperative task in constrained and unconstrained environments. Simulations verify the effectiveness of the proposed scheme. It is shown that the problem belongs to the class of NP-hard problems because it has more than two robots. Under the assumption that $NP \neq P$, no polynomial time algorithm exists. Therefore, in the worst case exponential time is needed to search through the whole search space X . Thus, exact approaches such as Integer Linear Programming and Constraint Programming are not capable of solving real-life instances. An indicator of the complexity of an instance is the size of the search space size (X), which can be calculated with specifying the set of all possible combinations for the given tasks and the number of robots needed for each combination in the configuration. We also plan to conduct further experiments and simulation in varying environments, with tasks of varying complexity, requiring different

numbers of robots and also do comparative studies with other methods. The system would have to assign not only a task, but also combine robots in a group if a task requires participation of several robots.

References

1. Zlot R, Stentz A (2005) Complex task allocation for multiple robots. In: Proceedings of the 2005 IEEE international conference on robotics and automation, Spain, pp 1515–1522
2. Berhaut M, Huang H, Keskinocaki P, Koenig S, Elmaghraby W, Griffin P (2003) Robot exploration with combinatorial auctions. In: International conference on intelligent robot and systems, Las Vegas, 27–31 Oct 2003
3. Jiang YC, Jiang JC (2005) A multi-agent coordination model for the variation of underlying network topology. *Expert Syst Appl* 29(2):372–382
4. Parker LE (1998) ALLIANCE: an architecture for fault tolerant multirobot cooperation. *IEEE Trans Robotics Autom* 14(2):220–240
5. Parker LE (1997) L-ALLIANCE: task-oriented multi-robot learning in behavior-based systems. *Adv Robotics* 11(4):305–322
6. Gerkey BP (2003) On multi-robot task allocation [D]. University of South California
7. Zlot R, Stentz A (2006) Market-based multirobot coordination for complex tasks. *Int J Robotics Res* 25(1):73–101
8. Ren-Ji C (2000) Coordination theory and implementation study of multiple behavior-based robotic systems. JiaoTong University, Shanghai
9. Watkins CJCH, Dayan P (1992) Q-learning. *Mach Learn* 8:279–292
10. Chandran T, Narendran T, Ganesh K (2006) A clustering approach to solve the multiple travelling salesman problem. *Int J Ind Syst Eng* 1:372–387
11. Cheng-Heng F, Ge SS (2005) Cooperative backoff adaptive scheme (COBOS) for multirobot task allocation. In: Proceedings of the IEEE Transactions on Robotics, vol 21(6)

Simulation-Based Evaluations of Reinforcement Learning Algorithms for Autonomous Mobile Robot Path Planning

Hoang Huu Viet, Phyo Htet Kyaw and TaeChoong Chung

Abstract This work aims to evaluate the efficiency of the five fundamental reinforcement learning algorithms including Q-learning, Sarsa, Watkins's $Q(\lambda)$, Sarsa(λ), and Dyna-Q, and indicate which one is the most efficient of the five algorithms for the path planning problem of autonomous mobile robots. In the sense of the reinforcement learning algorithms, the Q-learning algorithm is the most popular and seems to be the most effective model-free algorithm for a learning robot. However, our experimental results show that the Dyna-Q algorithm, a method learns from the past model-learning and direct reinforcement learning is particularly efficient for this problem in a large environment of states.

Keywords Reinforcement learning · Autonomous mobile robots · Path planning

1 Introduction

Mobile robotics is a research area that deals with autonomous and semi-autonomous navigation. Path planning problem is recognized as one of the most fundamental problems to applications of autonomous mobile robots. The path

H. H. Viet (✉) · P. H. Kyaw · T. Chung
Artificial Intelligence Lab, Department of Computer Engineering,
School of Electronics and Information, Kyung Hee University,
1-Seocheon, Giheung, Yongin, Gyeonggi 446-701, South Korea
e-mail: viethh@khu.ac.kr

P. H. Kyaw
e-mail: phyo@khu.ac.kr

T. Chung
e-mail: tcchung@khu.ac.kr

planning or trajectory planning problem of autonomous mobile robots refers to determining a collision-free path from its position to a goal position through an obstacle environment without human intervention [1].

Reinforcement learning (RL) is an approach to artificial intelligence that emphasizes learning by an agent from its interaction with the environment [2, 3]. The goal of the agent is to learn what actions to select in situations by learning a value function of situations or “states”. The learning agent is not conducted which actions to take, but it has to discover an optimal action of each state which yields the high rewards in a long-term objective. In literature, there have been several RL algorithms suggested to solve the path planning problem of autonomous mobile robots. Among those algorithms of RL, the Q-learning algorithm [4] has been frequently employed to solve the path planning problem [5–8]. The strength of RL methods is that it does not require an explicit model of an environment, thus it can be popularly employed to solve the mobile robot navigation problem. However, one primary difficulty faced by RL applications is that the most RL algorithms learn very slowly. As such, this work aims to evaluate five popular algorithms of RL including Q-learning, Sarsa, Watkins’s $Q(\lambda)$, Sarsa(λ), Dyna-Q based on computer simulations for the path planning problem of autonomous mobile robots, and to indicate which one is the most efficient for this problem. The rest of this article is organized as follows: Sect. 2 shows a short review of the algorithms that are going to be evaluated in this article. The evaluations are discussed in Sect. 3. Finally, we conclude our work in Sect. 4.

2 Background

2.1 Basic Concepts

Reinforcement learning emphasizes the learning process of an agent through trial-and-error interactions with an environment. In the standard RL model, an agent connects to its environment via perceptions and actions. On each step of interaction the agent receives a *state*, s , of the environment as an input and then the agent takes an *action*, a . The action changes the state of the environment, and a scalar value of the state-action pair is sent to the agent, called a *reward function* r of the state-action (s, a) pair. The set of all states makes the *state space*, S , of the environment, and the set of actions of the state s makes the *action space*, $A(s)$. The *value function* of a state (or state-action pair) is the total amount of rewards that an agent can expect to accumulate over the future starting from that state. A reward function indicates what is good in an immediate sense, whereas a value function specifies what is good in the long-run. A *policy* is a mapping from perceived states of the environment to actions taken in those states. A *model* of the environment is something that mimics the behavior of the environment. Given a state and an action, the model might predict the resultant of the next state and the reward

function. The objective of the agent is to learn actions that tend to maximize the long-run sum of the value of the rewards.

An *on-line* learning method learns while gaining experience from the environment. An *off-line* learning method waits until it is finished gaining experience to learn. An *on-policy* learning method learns about the policy it is currently following. An *off-policy* learning method learns about a policy while following another.

A *greedy strategy* refers to a strategy that agent always chooses the action with the highest value of the value function. The selected action refers to a *greedy action* and it is said that the agent is *exploiting* the environment. An ε -*greedy strategy* refers to a strategy that agent chooses the *greedy action* with probability of $1-\varepsilon$, and chooses the *random action* with a small probability of ε . The random action refers to a *non-greedy* action and it is said that the agent is *exploring* the environment.

If the agent-environment interaction process is broken into subsequences, each subsequence refers to an *episode* and the end state of each subsequence is called the *terminal state*. The learning task broken into episodes is called *episodic tasks*. In episodic tasks, the state space S denotes the set of all *non-terminal states* and the state space S^+ denotes the set S plus the *terminal state*.

In the RL algorithms, the parameter $\alpha \in (0, 1)$ denotes the learning rate, the parameter $\gamma \in (0, 1)$ denotes the discount rate, the parameter δ denotes the temporal-difference error, the parameter $\lambda \in (0, 1)$ denotes the decay-rate parameter for eligibility traces, the parameter ε denotes probability of random action in ε -*greedy* strategy, and $Q(s, a)$ denotes the action-value function of taking action a in state s .

2.2 Temporal Difference Learning Algorithms

In the temporal difference (TD) learning approach, two algorithms that can be identified as the main idea of TD method would certainly be Sarsa and Q-learning. The Sarsa algorithm (short for *state, action, reward, state, action*) is an *on-policy* TD learning algorithm, whereas the Q-learning algorithm is an *off-policy* TD learning algorithm. These two algorithms consider transitions from a state-action pair to a state-action pair and learn the action-value function of state-action pairs. While the Sarsa algorithm backups up the Q-value corresponding to the next selected action, the Q-learning algorithm backups up the Q-value corresponding to the action of the best next Q-value. Since these algorithms need to wait only one time step to backup the Q-value. So, they are *on-line* learning methods. The algorithms Sarsa [3] and Q-learning [4] are shown in Algorithm 1 and Algorithm 2, respectively.

<p>Algorithm 1: Sarsa algorithm</p> <p>Initialize $Q(s,a)$ arbitrarily</p> <p>Repeat (for each episode):</p> <p>Initialize s</p> <p>$a \leftarrow \varepsilon\text{-greedy}(s,Q)$</p> <p>Repeat (for each step of episode):</p> <p>Take action a, observe r, s'</p> <p>$a' \leftarrow \varepsilon\text{-greedy}(s',Q)$</p> <p>$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma Q(s',a') - Q(s,a)]$</p> <p>$s \leftarrow s'; a \leftarrow a'$</p> <p>Until s is terminal</p>	<p>Algorithm 2: Q-learning algorithm</p> <p>Initialize $Q(s,a)$ arbitrarily</p> <p>Repeat (for each episode):</p> <p>Initialize s</p> <p>Repeat (for each step of episode):</p> <p>$a \leftarrow \varepsilon\text{-greedy}(s,Q)$</p> <p>Take action a, observe r, s'</p> <p>$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$</p> <p>$s \leftarrow s'$</p> <p>Until s is terminal</p>
--	--

2.3 Eligibility Traces

<p>Algorithm 3: Sarsa(λ) algorithm</p> <p>Initialize $Q(s,a)$ and $e(s,a) = 0$, for all s, a</p> <p>Repeat (for each episode):</p> <p>Initialize s, a</p> <p>Repeat (for each step of episode):</p> <p>Take action a, observe r, s'</p> <p>$a' \leftarrow \varepsilon\text{-greedy}(s',Q)$</p> <p>$\delta \leftarrow r + \gamma Q(s',a') - Q(s,a)$</p> <p>$e(s,a) \leftarrow 1$</p> <p>For all s, a:</p> <p>$Q(s,a) \leftarrow Q(s,a) + \alpha \delta e(s,a)$</p> <p>$e(s,a) \leftarrow \gamma \lambda e(s,a)$</p> <p>$s \leftarrow s'; a \leftarrow a'$</p> <p>Until s is terminal</p>	<p>Algorithm 4: Q(λ) algorithm</p> <p>Initialize $Q(s,a)$ and $e(s,a) = 0$, for all s, a</p> <p>Repeat (for each episode):</p> <p>Initialize s, a</p> <p>Repeat (for each step of episode):</p> <p>Take action a, observe r, s'</p> <p>$a' \leftarrow \varepsilon\text{-greedy}(s',Q)$</p> <p>$a^* \leftarrow \operatorname{argmax}_b Q(s',b)$</p> <p>$\delta \leftarrow r + \gamma Q(s',a^*) - Q(s,a)$</p> <p>$e(s,a) \leftarrow 1$</p> <p>For all s, a:</p> <p>$Q(s,a) \leftarrow Q(s,a) + \alpha \delta e(s,a)$</p> <p>If $a' = a^*$, then $e(s,a) \leftarrow \gamma \lambda e(s,a)$</p> <p>Else $e(s,a) \leftarrow 0$</p> <p>$s \leftarrow s'; a \leftarrow a'$</p> <p>Until s is terminal</p>
---	---

Eligibility traces are one of the basic mechanisms of RL. Almost any TD method can be combined with eligibility traces to obtain a more general method that may learn more efficiently. An eligibility trace is a temporary record storing a trace of the state-action pairs taken over time. When eligibility traces are augmented with the Sarsa algorithm, it is known as the Sarsa(λ) algorithm. The basic algorithm is similar to the Sarsa algorithm, except that backups which are carried out over n steps later instead of one step later. The Watkins's Q(λ) [hereinafter called Q(λ)] algorithm is similar to the Q-learning algorithm, except that it is supplemented eligibility traces. The eligibility traces are updated in two steps. First, if a *non-greedy* action is taken, they are set to zero for all state-action pairs. Otherwise, they are decayed by $\gamma\lambda$. Second, the eligibility trace corresponding to the current state-action pair is reset to 1. The algorithms Sarsa(λ) and Q(λ), referred from [3], using replacing traces are shown in Algorithm 3 and Algorithm 4, respectively.

2.4 Dyna-Q Algorithm

The Dyna-Q algorithm is the integration of planning and direct RL methods. Planning is the process that takes a model as an input and produces a policy by using simulated experience generated uniformly at random, whereas direct RL method uses a real experience generated by the environment to improve the value function and policy. The Dyna-Q algorithm is shown in Algorithm 5 [3]. The $Model(s, a)$ represents the next predicted state and reward of the model for the state-action pair (s, a) and N is the number of planning steps. Step (d) is the direct RL, steps (e) and (f) are model-learning and planning, respectively. If steps (e) and (f) are omitted, the planning step $N = 0$, the remaining algorithm is the Q-learning algorithm.

Algorithm 5: Dyna-Q algorithm

<p>Initialize $Q(s,a)$ and $Model(s,a)$, for all s, a Do forever: (a) $s \leftarrow$ current (<i>non-terminal</i>) state (b) $a \leftarrow \epsilon$-greedy(s, Q) (c) Take action a, observe r, s' (d) $Q(s,a) \leftarrow Q(s,a) + \alpha [r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$ (e) $Model(s,a) \leftarrow s', r$ (assuming deterministic environment) (f) Repeat N times: $s \leftarrow$ random previously observed state $a \leftarrow$ random action previously taken in s $s', r \leftarrow Model(s,a)$ $Q(s,a) \leftarrow Q(s,a) + \alpha [r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$</p>

3 Evaluations

In this section, assumptions of the path planning problem are defined. Evaluations based on simulations of the algorithms are implemented to determine which one is the most efficient for the autonomous mobile robot path planning.

3.1 Assumptions

Assumption 1 The environment of the robot consists of a goal position and obstacles. The position of the goal, the position and shape of obstacles are unknown by the robot.

Assumption 2 The robot is equipped with all necessary sensors such that the robot knows its position, detects obstacles if collisions occur, and determines the goal if it reaches to the goal position during navigating time.

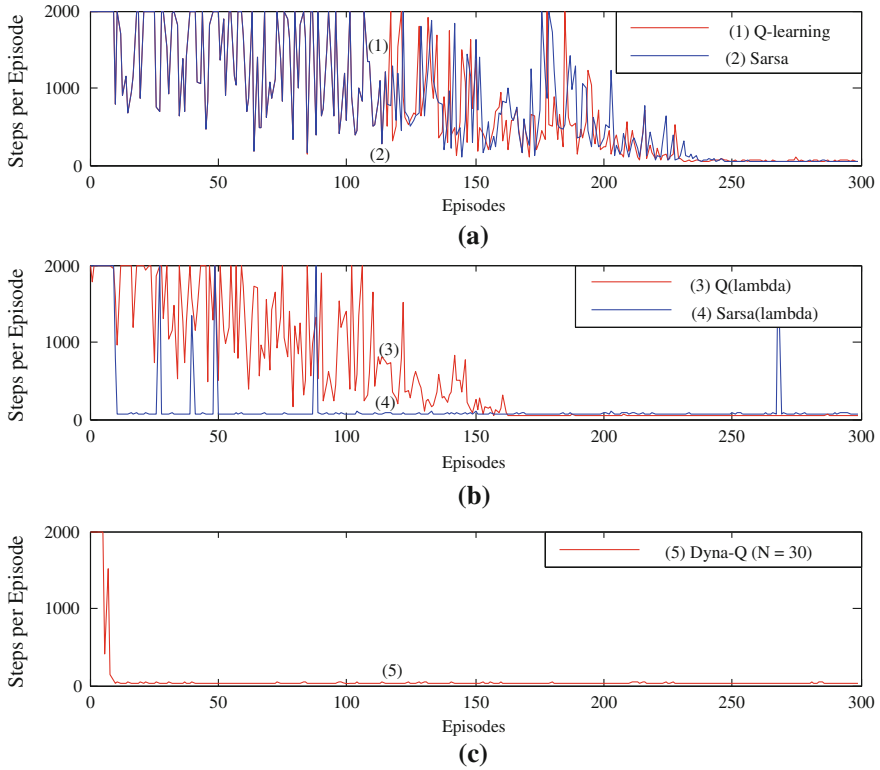


Fig. 1 Comparison steps per episodes of algorithms Q-learning, Sarsa, $Q(\lambda)$, Sarsa(λ), and Dyna-Q

Table 1 The performance of the algorithms Q-learning, Sarsa, $Q(\lambda)$, Sarsa(λ), Dyna-Q described in the first simulation

Criterion	Q-learning	Sarsa	$Q(\lambda)$	Sarsa(λ)	Dyna-Q
Episodes	230	250	170	270	10
Steps	227, 469	237, 401	170, 234	48, 653	14, 138
Path length	53	49	41	62	34

Assumption 3 The robot initially has no knowledge of the effect of its actions on what position it will occupy next and the environment provides rewards to the robot and that this reward structure is also initially unknown to the robot.

Assumption 4 From its current position, the robot can move to an adjacent position in one of the eight directions, East, North-East, North, North-West, West, South-West, South, and South-East, except that any direction that takes the robot into obstacles or outside of environment, in which case the robot keeps its current position.

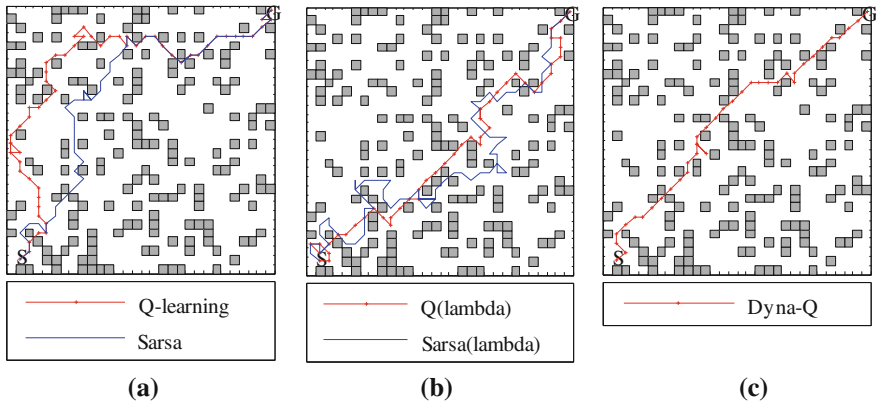


Fig. 2 Paths are found by the algorithms **a** Q-learning and Sarsa, **b** $Q(\lambda)$ and Sarsa(λ), **c** Dyna-Q after 300 episodes

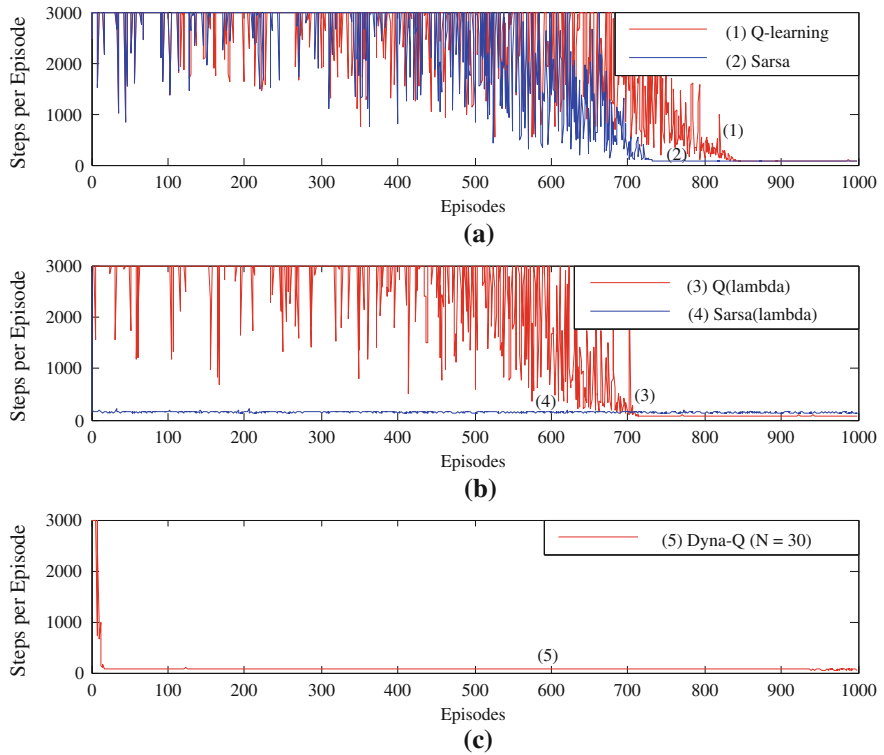


Fig. 3 Comparison steps per episodes of algorithms Q-learning, Sarsa, $Q(\lambda)$, Sarsa(λ), and Dyna-Q

Table 2 The performance of the algorithms Q-learning, Sarsa, $Q(\lambda)$, Sarsa(λ), Dyna-Q described in the second simulation

Criterion	Q-learning	Sarsa	$Q(\lambda)$	Sarsa(λ)	Dyna-Q
Episodes	850	720	710	10	20
Steps	1, 963, 980	1, 731, 725	1, 728, 585	6, 757	30, 556
Path length	74	73	74	128	58

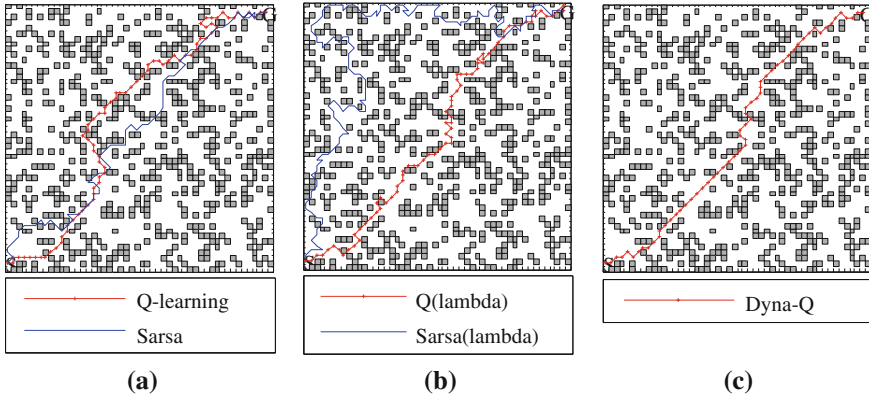


Fig. 4 Paths are found by the algorithms **a** Q-learning and Sarsa, **b** $Q(\lambda)$ and Sarsa(λ), **c** Dyna-Q after 1,000 episodes

Assumption 5 If the robot reaches to the goal position, a reward of 1 is given for the robot. Otherwise, a reward of zero is given for it. After reaching the goal position, the robot returns to the start position to begin a new episode.

The task of the robot is to discover a collision-free path from the start position (S) to the goal position (G) through its environment. Evaluations of algorithms for the path planning problem are based on the speed of convergence of the algorithms to a near-optimality path and length of the path obtained.

3.2 Simulations and Evaluations

In this section, two simulations using the Matlab software are implemented to evaluate the efficiency of the algorithms. The environments of these simulations are represented by the cells of a uniform grid. Each cell with a zero value is considered as a state of the environment. Otherwise, it is considered as an obstacle. The basic parameters for the all simulations are set as follows: $\alpha = 0.1$, $\gamma = 0.95$, $\lambda = 0.95$, $\epsilon = 0.05$. After each episode, the value of ϵ is set again by $\epsilon = 0.99\epsilon$.

The environment of the first simulation is a maze as shown in Fig. 2. The maze consists of $30 \times 30 = 900$ cells in which 20% cells make obstacles, so the number of states of the environment is 720 states. The maximum step of each episode is

Table 3 The evaluations of the algorithms Q-learning, Sarsa, $Q(\lambda)$, Sarsa(λ), and Dyna-Q

Criterion	Q-learning	Sarsa	$Q(\lambda)$	Sarsa(λ)	Dyna-Q
Soundness	Yes	Yes	Yes	Yes	Yes
Completeness	Yes	Yes	Yes	Yes	Yes
Optimality	Bad	Bad	Medium	Bad	Good
Speed of convergence	Slow	Slow	Medium	Rapid	Rapid

2,000 steps. Figure 1 shows the steps per episodes of algorithms Q-learning, Sarsa, $Q(\lambda)$, Sarsa(λ), and Dyna-Q. Table 1 depicts the performance of these algorithms, where episodes refer to the number of episodes taken to converge to a near-optimality path, steps refer to the sum of steps taken to converge to a near-optimality path, and path length refers to the length of the path found by the algorithms after 300 episodes. Figure 2 depicts the paths found by the algorithms Q-learning, Sarsa, $Q(\lambda)$, Sarsa(λ), and Dyna-Q after 300 episodes. It can be seen from Table 1 and Fig. 2 that the Dyna-Q algorithm obtains a near-optimality path with the shortest length in the smallest number of steps among five algorithms.

The next simulation is to evaluate the efficiency of the five algorithms in a larger environment of states and obstacles. In this simulation, the environment is a maze as shown in Fig. 4. The maze consists of $50 \times 50 = 2,500$ cells in which 25% cells make obstacles, so the number of states of the environment is 1,875 states. The maximum step of each episode is 3,000 steps. Figure 3 shows the steps per episodes of algorithms Q-learning, Sarsa, $Q(\lambda)$, Sarsa(λ), and Dyna-Q. Table 2 depicts the performance of these algorithms, where parameters are the same as in Table 1, except path length refers to paths found by the algorithms after 1,000 episodes. Figure 4 depicts the paths found by the algorithms Q-learning, Sarsa, $Q(\lambda)$, Sarsa(λ), and Dyna-Q after 1,000 episodes. In this simulation, the Sarsa(λ) algorithm converges to a near-optimality path quickly, but the path found by this algorithm is much longer than the path found by the Dyna-Q algorithm. The Dyna-Q algorithm is really effective in this simulation.

Based on simulation results shown above, some evaluation criteria [1] of these algorithms are summarized in Table 3, where the soundness means that the planned path is guaranteed to be collision-free, the completeness means that the algorithm is guaranteed to find a collision-free path if one exists, the optimality means that the length of the actual path obtained versus the optimal path, and speed of convergence means that the computer time taken to find a near-optimality path. Here, the criteria of optimality and speed of convergence to a near-optimality path are only compared among the algorithms.

4 Conclusions

In this work, we reviewed and evaluated some popular RL algorithms for the path planning problem of autonomous mobile robots. In the first sense of RL algorithms, the Q-learning algorithm is the most popular and seems to be the most

effective model-free algorithm for a learning robot. However the simulation results show that the Q-learning is not really effective for finding a collision-free path in an environment that the number of states and obstacles are so large. Both the Sarsa algorithm and the Q-learning algorithm converge quite slowly and the paths found by these two algorithms are not good paths. The algorithms $Q(\lambda)$ and $Sarsa(\lambda)$ improve quite well the speed of convergence to a near-optimality path. But, the Dyna-Q algorithm is particularly efficient in solving the path planning problem of autonomous mobile robots. With the experimental results shown above, we believe that the Dyna-Q algorithm is the best choice among algorithms Q-learning, Sarsa, $Q(\lambda)$, $Sarsa(\lambda)$, and Dyna-Q to solve the path planning problem. However, we have just emphasized our work on the simulations of the maze domain. We plan to extend the Dyna-Q algorithm to the real robot in the real environment.

Acknowledgments This research 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 (2010-0012609).

References

1. Dudek G, Jenkin M (2010) Computational principles of mobile robotics. Cambridge University Press, New York
2. Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement learning: a survey. *J Artif Intell Res* 4:237–285
3. Sutton RS, Barto AG (1998) Reinforcement learning: an introduction. The MIT Press, Cambridge
4. Watkins C (1989) Learning from delayed rewards. Ph.D. Dissertation, King's College
5. Smart WD, Kaelbling LP (2002) Effective reinforcement learning for mobile robots. In: IEEE international conference on robotics and automation (ICRA'02), vol 4. IEEE Press, Washington, pp 3404–3410
6. Zamstein L, Arroyo A, Schwartz E, Keen S, Sutton B, Gandhi G (2006) Koolio: path planning using reinforcement learning on a real robot platform. In: 19th Florida conference on recent advances in robotics, Florida
7. Chakraborty IG, Das PK, Konar A, Janarthanan R (2010) Extended Q-learning algorithm for path-planning of a mobile robot. In: LNCS, vol 6457. Springer, Heidelberg, pp 379–383
8. Mohammad AKJ, Mohammad AR, Lara Q (2011) Reinforcement based mobile robot navigation in dynamic environment. *Robotics Comput-Integr Manuf* 27:135–149

Control Mechanism for Low Power Embedded TLB

Jung-hoon Lee

Abstract This research proposes a new embedded translation look-aside buffer (TLB) structure that can reduce the power consumption effectively by using simple hardware control logics. The proposed TLB structure is constructed as two fully associative TLBs and one of the two TLBs is selectively accessed by the dynamic selection method. It is shown that on-chip power consumption of the proposed TLB can be reduced by around 42% comparing with the conventional fully associative TLBs with the same number of entries.

Keywords Translation look-aside buffer (TLB) · Low power design · Temporal locality · Memory hierarchy

1 Introduction

Low-power techniques for memory systems can be used for all the design levels from the high levels including algorithm selection, system integration, and architecture design, and up to the low levels including gate/circuit design and process. However, applying the low-power design technology to the higher levels of architecture, algorithm, and system levels may cause a larger effect with relatively less design effort and cost than changing process technology or optimizing gate and circuit design [1]. Because most of TLB structures are constructed as a fully-associative TLB with CAM cells, power consumption of the TLB varies

J. Lee (✉)

ERI, Electrical and Electronic Engineering, GyeongSang National University,
900 Ga-jwa, Jinju, South Korea
e-mail: leejh@gsnu.ac.kr

linearly depending upon the number of its entries [2]. In case of the Strong ARM [3] and ARM920T [4], the amount of power dissipated at the TLB corresponds to around 17 and 10% of the overall power consumption respectively. Although the physical size of a TLB is small, compared with a cache memory, it accounts for a significant fraction of the total power consumption.

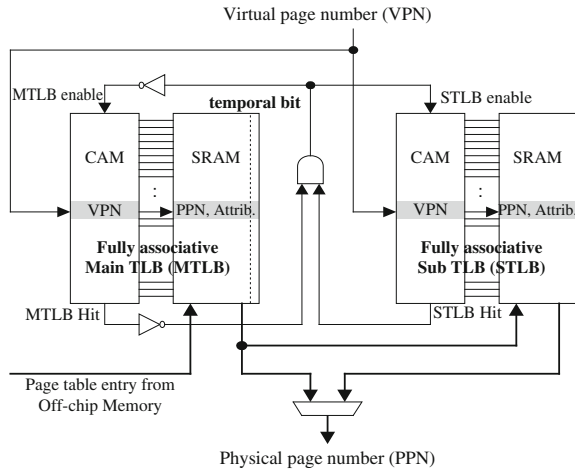
In many applications, such as portable devices, energy efficiency is more important than performance. In order to reduce power consumption, maintaining a micro-TLB above the conventional TLB level turns out to be an effective approach for instruction TLB with low miss ratio. But in case of data TLB, performance degradation tends to be more significant than the gain by power reduction. The other TLB studies are memory cell redesign, such as modified CAM cell, voltage reduction, and optimized TLB structures. Our focus is to optimize the basic TLB structure with the aid of a simple mechanism. Thus simple control and construction can be achieved by this method. Conclusively, the proposed TLB system is designed as a low power/high performance TLB structure for embedded processors.

2 Control Mechanism for Dual TLB

2.1 Proposed TLB Structure

The proposed selective TLB structure is shown in Fig. 1. The selective TLB is constructed as two fully associative TLBs, i.e., main-TLB and sub-TLB, and one of the two TLBs is selectively accessed. When the CPU accesses memory, the main-TLB (MTLB) is searched first for a match. If a miss occurs at the MTLB, then the sub-TLB (STLB) is searched during the next cycle. If a hit at the STLB occurs, when the next virtual address is generated, the STLB is searched first. Two consecutive misses at the both places cause a miss handling service to the operating system. This scheme of dynamic search ordering has been made possible by page characteristics, that is, if one page is loaded, that page has high probability of consecutive hits because one page has many hundreds or thousands of information. Therefore this scheme improves the average access time of conventional dual TLB system, which was a major weak point. Each entry of MTLB holds a new control bit, called a temporal bit. This single bit is used to select pages with temporal locality. Generally if one page is loaded, that page has high probability of consecutive hits and thus it cannot be used as a sign of temporal locality, and therefore temporal bit is kept as 0. The temporal bit is set to 1 only if other TLB entry is accessed and the TLB reference returns to the original page. This mechanism is accomplished to compare a virtual page number (VPN) accessed just before with a newly accessed VPN. Replacement policy of the two TLBs is chosen as FIFO algorithm. If the MTLB is full, the oldest entry is replaced. And then if temporal bit of the entry is set, the entry is moved into the STLB.

Fig. 1 Proposed dual TLB structure



2.2 Proposed TLB Control Algorithm

Algorithms for the proposed dual TLB management are described in detail. When the first virtual address is generated, MTLB is searched. If a hit occurs at the MTLB, address translation is performed. If a miss at the MTLB occurs, then the STLB is searched during next cycle. And also if a miss at the STLB occurs consecutively, a new page table entry is placed on MTLB. Continuously when the next virtual address is generated, the MTLB is searched during one cycle until the MTLB is filled up. When the MTLB is full and a MTLB miss occurs, one entry within the MTLB is selected and replaced with a new page entry. This page entry replaced from the MTLB is placed at the STLB if it shows high possibility to be accessed in the future. If a page table entry is to be placed in STLB, possible cases are:

- Hit in main-TLB: if a page is found in the MTLB, the actions are not different at all from any conventional TLB hit. The requested physical address is sent to the cache and compared with tag bits of the cache. Also next TLB search is performed in the order of the MTLB and STLB in each of next two cycles. If the currently generated tag value does not equal to the preceding tag value, its corresponding temporal bit of the entry is set to 1.
- Hit in sub-TLB: when the CPU generates a virtual address and if a page is found in the STLB, the actions are not different at all from MTLB hit. But next TLB search is done in the order of the STLB and MTLB in each of next two cycles. When consecutive hits at the STLB occur, address translations are performed at the STLB until a miss occurs.
- Miss in both places: a miss occurs at both TLBs and while the O/S is handling the miss, controller will check whether the MTLB is filled up or not. If MTLB is full, the oldest entry is replaced. And then if its corresponding temporal bit is

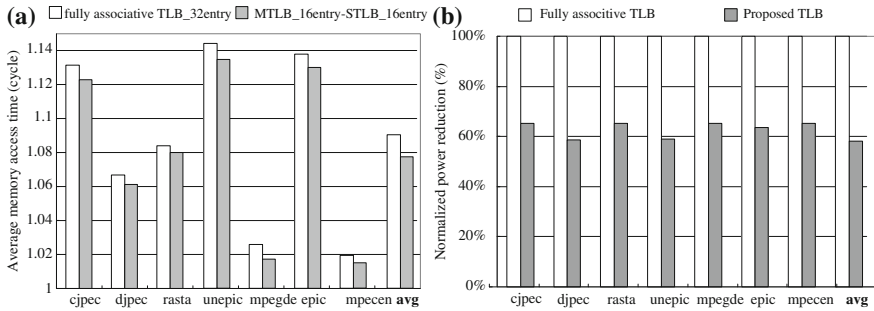


Fig. 2 Comparison of the performance and the power consumption

set, the entry is moved into the STLB. These actions are to exploit temporal locality selectively and also the lifetime of each entry with temporal locality can increase. If MTLB is not full, incoming new value is placed in MTLB and sent to the cache at the same time.

3 Performance and Evaluation

Two major performance metrics, i.e., the average memory access time and power consumption are used to evaluate and compare the proposed TLB system with conventional fully associative TLB with the same number of entries, e.g., 32 entries. It is assumed that CPU clock is 200 MHz, memory latency is 15 cpu cycles, and memory bandwidth is 1.6 Gbytes/sec. These parameters are based on the values used for common 32-bit embedded processors (e.g., Hitachi SH4 or ARM920T).

It is shown that average memory access time of the proposed TLB can be reduced by about 10% and power consumption can be reduced by around 42% comparing with the conventional fully associative TLB (Fig. 2).

4 Conclusion

High performance and low power consumption are two important factors to consider in designing many embedded systems. This research proposes a new TLB system that can reduce the power consumption effectively by using simple hardware control. The proposed TLB system consists of two fully associative TLBs and either of the two TLB is selectively accessed from access pattern. Also in order to obtain high performance, using of a new control bit can select pages with temporal locality effectively.

According to the results of simulation and analysis, performance improvement can be achieved reasonably and this is shown by comparing with the conventional fully associative TLB with the same number of entries. And it is shown that on-chip power consumption of the proposed TLB can be reduced by around 42% comparing with the conventional fully associative TLB with the same number of entries.

References

1. Jeyapaul R, Marathe S, Shrivastava A (2009) Code transformations for TLB power reduction. In: 22th IEEE international conference on VLSI. IEEE Press, New Delhi, pp 413–418
2. Kadayif I, Sivasubramaniam A, Kandemir M, Kandiraju G, Chen G (2005) Optimizing instruction TLB energy using software and hardware techniques. In: ACM transactions on design automation of electronic systems, vol 10, ACM, pp 229–257
3. StrongARM, <http://en.wikipedia.org/wiki/StrongARM>
4. Segars S (2001) Low power design techniques for microprocessor. In: Proceedings of international solid-state circuit conference. IEEE Press, San Francisco

Part V
IT Multimedia for Ubiquitous
Environments

A Noise Reduction Method for Range Images Using Local Gaussian Observation Model Constrained to Unit Tangent Vector Equality

Jeong Heon Kim and Kwang Nam Choi

Abstract We present a method for smoothing heavy noisy surfaces acquired by on-the-fly 3D imaging devices to obtain the stable curvature. The smoothing is performed in a way that finds centers of probability distributions which maximizes the likelihood of observed points with smooth constraints. The smooth constraints are derived from the unit tangent vector equality. This provides a way of obtaining smooth surfaces and stable curvatures. We achieve the smoothing by solving the regularized linear system. The unit tangent vector equality involves consideration of geometric symmetry and it minimizes the variation of differential values that are a factor of curvatures. The proposed algorithm has two apparent advantages. The first thing is that the surfaces in a scene with various signals to noise ratio are smoothed and then they can earn suitable curvatures. The second is that the proposed method works on heavy noisy surfaces, e.g., a stereo camera image. Experiments on range images demonstrate that the method yields the smooth surfaces from the input with various signals to noise ratio and the stable curvatures obtained from the smooth surfaces.

Keywords Range image · Noise · Local gaussian observation model · Linear system

J. H. Kim (✉) · K. N. Choi
Department of Computer Science and Engineering, Chung-Ang University,
Heukseok-dong Dongjak-gu, Seoul, Korea
e-mail: jhkim@vim.cau.ac.kr

K. N. Choi
e-mail: knchoi@cau.ac.kr

1 Introduction

Image processing is a fundamental field in computer vision. The advance of imaging technology gives good opportunities to various image acquisitions. The processing of 3D information from these images has become an important issue in visualization and vision. The development of 3D sensing technologies makes production of high resolution range image possible, and it follows that they continuously suffer from noise. The surfaces of interest need to be extracted from the noisy data. Consequently, the need for noise reduction methods of 3D image processing has been increased recently.

Denoising or smoothing images is one of the most prevalent works of image processing. Traditionally the methods for denoising 2D images focus on local values of quantity field. Furthermore, we consider local geometric relationship to form smooth surfaces of 3D objects. The 3D object has a stable distribution of a feature based on the surface shape such as surface curvatures in differential geometry because of the consideration of local geometric relationship.

Curvature is one of the fine features with transformation invariant to describe the surface. The invariant is a good characteristic for computer vision—object recognition, pose estimation, motion estimation and image matching [1–3]. Curvature represents the sharpness of a surface or a curve and computed from 1st and 2nd partial derivatives in 3D Euclidean space. However, curvature is very sensitive to noise because of the characteristics based on derivative feature. The noisy surface from 3D imaging sensors produces uneven curvature distribution over all observed objects. They indicate that most locations have sharp bends. In other words, the noisy surfaces yield unsteady curvature even on flat surfaces and it does not correspond with our expectation. Thus a sharp point is not discriminated from other flat points. Useful curvature for discrimination is obtained from smooth surfaces; therefore, we should improve methods to smooth noisy surfaces.

The development of the vision technology could bring into existence the outstanding mobile device for range image acquisition. We can rapidly and easily obtain the range images in anywhere at any time, while there is a trade-off: We have more and biased noises. Magnitudes and directions of the noise are biased by the device, with the consequence that the noise has different distribution on each direction. For such reasons as mentioned, observed surfaces are irregular and the noise of those has anisotropic distribution. The smoothing methods are required especially for the applications that use these on-the-fly devices.

Surface reconstruction by fitting a Radial Basis Function (RBF) is one of the popular techniques. Carr et al. [4] smooth the scattered range data by convolving with a smoothing kernel (low pass filtering) during the evaluation of the RBF. They also show the discrete approximation to the smoothing kernel. This allows arbitrary filter kernels, including anisotropic and spatially varying filters. Smooth interpolation by moving least squares (MLS) approximation is also one of the powerful approaches. The mesh-independent MLS-based projection strategy for general surface interpolation is proposed [5]. This is applicable to smoothing a

$(d - 1)$ -dimensional manifold in \mathbb{R}^d , $d \geq 2$; and the resulting surface is C^∞ smooth. Image-smoothing technique by diffusion is general. Anisotropic diffusion for images is proposed by Perona and Malik [6]. Tasdizen et al. [7] develops the surface smoothing via level set surface models and anisotropic diffusion of normals. Anisotropic diffusion is a much better method for denoising than isotropic diffusion which behaves like a low pass filter. The targets of these methods are not only the range images from on-the-fly devices. On-the-fly range images have various noise levels in a scene. One smoothing level is not satisfied of the on-the-fly range images.

Our goals are obtaining denoised smooth surfaces and the smooth stable curvature in range images from on-the-fly 3D imaging devices. The range images are illustrated in scattered points and the smooth surface can be obtained by fitting the scatter points. In this paper we focus on the non-iterative approximation to noisy surfaces with satisfaction of constraint that is the differential geometry representation of the surface with stable curvature. We formulate this in the context of regularization with two linear systems. The one is the maximum log-likelihood estimate (MLE) of the likelihood in which points are observed. The other is the smooth constraints that are the equality of neighbor unit tangent vectors. The proposed method offers two apparent advantages. The first thing is that it yields the stable curvature. They are computed from the smooth variation of surface normal vectors. The unit tangent vectors of a point on surface are the factor of surface normals. The minimization of unit tangent vector variation makes the variation of curvature minimal. The unit tangent vector equality involves consideration of geometric symmetry. The consideration improves the results better than those of traditional noise reduction methods in curvatures. The second advantage is that it is non-iterative approximation. On-the-fly 3D imaging devices produce range images. The applications using the device work sequentially with the range images. The direct linear system solvers for proposed method are powerful and well researched [8, 9].

2 Observation Model and Constraint

On-the-fly 3D imaging devices such as stereo camera, Flash RADAR and structured light 3D scanner produce range images rapidly. Our approach is non-iterative so as to process continuous input range images periodically. The range image is the format of lattice and is formed of discrete data in general. The range images are similar to 2D color images and they can be applied to 2D image filters. However, the Gaussian or median filters that are useful for noise reduction are not suitable to make the stable curvature; they do not consider geometric symmetry except values only. One of our goals is on the stable curvature. It demands the constraint that a surface has inherent stable curvature beyond the simple smoothing. One of the methods for the approximation by constraints is regularization in linear algebra.

We derive linear system to describe the range image from a probability point of view for defining the regularization problem. The constraint is represented in linear system through the neighborhood relation.

2.1 MLE of the Point Observation Likelihood

Suppose that each observation point in the range image is the random variable observed from the true coordinate with a Gaussian distribution. A set \mathcal{N} contains a point \mathbf{x}_p and its neighbor. To simplify this problem, we shall assume that points in \mathcal{N} have independent probability distribution; probability distribution of each axis is independent at a point. Suppose that \mathcal{N} contains k points, $\mathbf{x}_1, \dots, \mathbf{x}_k$. Then, the likelihood of \mathcal{N} on one axis is

$$\begin{aligned} p(\mathcal{N}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) &= \prod_{i=1}^k p(x_i|\mu_i, \sigma_i^2) \\ &= \prod_{i=1}^k \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right] \end{aligned} \quad (1)$$

where mean $\boldsymbol{\mu}$ denote the vector $\boldsymbol{\mu} = (\mu_1, \dots, \mu_k)^T$ and variance $\boldsymbol{\sigma}^2$ denote the vector $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_k^2)^T$.

The maximum log-likelihood estimate of (1) for $\boldsymbol{\mu}$ must satisfy

$$\sum_{i=1}^k -\frac{x_i - \mu_i}{\sigma_i^2} = 0 \quad (2)$$

and

$$\sum_{i=1}^k \frac{\mu_i}{\sigma_i^2} = \sum_{i=1}^k \frac{x_i}{\sigma_i^2}. \quad (3)$$

An observation point x_i is known value. A variance σ_i^2 is drawn from the accuracy of 3D imaging devices on distance. The true coordinate that is center of Gaussian distribution μ_i is unknown, furthermore the accuracy from the true coordinate is unknown. We assume that the accuracy from x_i similar to the accuracy from μ_i . While the observation point is observed within the accuracy, we define the accuracy as a 3σ . Now, remained unknown variable is only μ_i .

2.2 Reproduction to Linear System

The derived MLE represent in linear system for the regularization. The linear algebra form of (3) is

$$\mathbf{w}^T \boldsymbol{\mu} = \mathbf{w}^T \mathbf{x}, \quad (4)$$

where vector $\mathbf{w} = (1/\sigma_1^2, \dots, 1/\sigma_k^2)^T$.

Suppose that the size of range image is $m \times n$. Then, we can rewrite (4) as

$$\mathbf{a}^T \boldsymbol{\mu} = \mathbf{a}^T \mathbf{x}, \quad (5)$$

where vector $\mathbf{a} = (a_1, \dots, a_l)^T$. The element a_i of \mathbf{a} is

$$a_i = \begin{cases} \frac{1}{\sigma_i^2} & \text{if neighborhood of } \mathbf{x}_p \\ 0 & \text{otherwise} \end{cases}. \quad (6)$$

While the range image contains s samples, (5) can be rewritten as

$$\mathbf{A}\boldsymbol{\mu} = \mathbf{A}\mathbf{x}, \quad (7)$$

where matrix $\mathbf{A} = (\mathbf{a}_1, \dots, \mathbf{a}_s)^T$.

2.3 Unit Tangent Vector Equality

Surface is expressed as a mapping of an open set D of 2D Euclidean space \mathbb{R}^2 into 3D Euclidean space \mathbb{R}^3 by a coordinate patch $\boldsymbol{\mu} : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$ in differential geometry [10]. Expressing the coordinate patch $\boldsymbol{\mu}$ as a function on D yields the formula

$$\boldsymbol{\mu}(u, v) = (f_1(u, v), f_2(u, v), f_3(u, v)) \quad (8)$$

where f_1, f_2, f_3 are arbitrary functions.

For each point (u_0, v_0) in D , the curve $\boldsymbol{\mu}(u, v_0)$ is called the u -parameter curve on $v = v_0$ of $\boldsymbol{\mu}$; and the curve $\boldsymbol{\mu}(u_0, v)$ is the v -parameter curve on $u = u_0$ of $\boldsymbol{\mu}$. We now calculate the tangent vectors $\boldsymbol{\mu}_u, \boldsymbol{\mu}_v$ at u_0, v_0 of the u -parameter curve and the v -parameter curve by partial differential on each direction. The partial differentials of range image are

$$\boldsymbol{\mu}_u = \boldsymbol{\mu}(u_0 + 1, v_0) - \boldsymbol{\mu}(u_0, v_0), \quad (9)$$

$$\boldsymbol{\mu}_v = \boldsymbol{\mu}(u_0, v_0 + 1) - \boldsymbol{\mu}(u_0, v_0). \quad (10)$$

The equality of neighbor u -direction unit tangent vectors, the constraint for smoothing are

$$\frac{\boldsymbol{\mu}(u_0 + 1, v_0) - \boldsymbol{\mu}(u_0, v_0)}{\|\boldsymbol{\mu}_u\|} = \frac{\boldsymbol{\mu}(u_0 + 2, v_0) - \boldsymbol{\mu}(u_0 + 1, v_0)}{\|\boldsymbol{\mu}_{u+1}\|} \quad (11)$$

and

$$0 = \left(\frac{1}{\|\boldsymbol{\mu}_u\|} \right) \boldsymbol{\mu}(u_0, v_0) - \left(\frac{1}{\|\boldsymbol{\mu}_u\|} + \frac{1}{\|\boldsymbol{\mu}_{u+1}\|} \right) \boldsymbol{\mu}(u_0 + 1, v_0) \\ + \left(\frac{1}{\|\boldsymbol{\mu}_{u+1}\|} \right) \boldsymbol{\mu}(u_0 + 2, v_0).$$

If $\|\boldsymbol{\mu}_u\|$ and $\|\boldsymbol{\mu}_{u+1}\|$ are given, then we can formulate in linear algebra of $\boldsymbol{\mu}$ and rewrite the linear system of full range image in the same manner as MLE of the point observation likelihood. However, because $\|\boldsymbol{\mu}_u\|$ and $\|\boldsymbol{\mu}_{u+1}\|$ are unknown, we draw from $\|\mathbf{x}_u\|$. The partial differential \mathbf{x}_u of noisy data is very sensitive. The estimation of $\boldsymbol{\mu}_u$ from smooth surface is complicate because of the sensitive. Thus, we separate $\|\mathbf{x}_u\|$ of full image into two categories; neighborhood and non-neighborhood. We define $\|\boldsymbol{\mu}_u\|$ as the representative of $\|\mathbf{x}_u\|$ category.

We write the linear equation of the equality of neighbor v -direction unit tangent vectors in the same manner as mentioned above. We can combine two linear equations of constraint and we have the unit tangent vector equality constraint

$$\Gamma \boldsymbol{\mu} = \mathbf{0}. \quad (13)$$

2.4 Tikhonov Regularization

Regularization is a general technique to prevent over-fitting. Consequently, the regularization smooths out the noisy surface with the constraint. The most common and well known form of regularization is the one known as Tikhonov regularization. The solution of linear system by Tikhonov regularization $\boldsymbol{\mu}_\lambda$ is defined the minimizer of weighted sum of residual norm and the side constraint

$$\boldsymbol{\mu}_\lambda = \arg \min \left\{ \|\mathbf{A}\boldsymbol{\mu} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda^2 \|\Gamma \boldsymbol{\mu}\|_2^2 \right\} \quad (14)$$

where the regularization parameter λ controls the balance of minimization between residual norm and side constraint. We solve the three linear systems that involve (14) of other two axes.

3 Experiments

The experiments use preprocessed range images that are removed the impulse noise. Accuracy of range images is based on the device specification. Variance σ^2 of Gaussian distribution, the observation model, is draw from $accuracy = 3\sigma$.

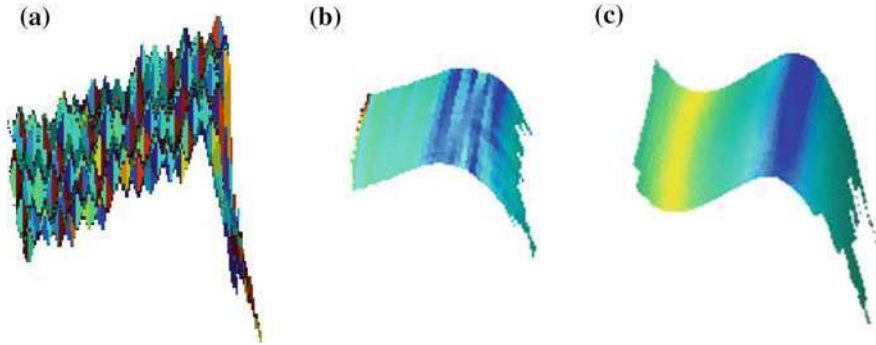


Fig. 1 a Magnified surface of the partial range image from the stereo camera, color shows mean curvatures. The Gaussian filtered surface **b** shows unstable curvatures and data loss. The smoothed surface by the proposed method **c** has stable curvatures

We solve the linear system with regularizer using the method of normal equation in the least-squares sense.

3.1 Compare with Gaussian Filter

The most common and well known method of noise reduction is the one known as Gaussian filter. We compared the proposed method by Gaussian filter. Figure 1 shows the result. Figure 1a is the magnified surface of the partial range image from the stereo camera used for the experiment; color shows mean curvatures. The Gaussian filter makes limited smooth surfaces. The result at edge is irregular in particular. The data loss occurs because of the region without observation. The Gaussian filtered surface Fig. 1b shows unstable curvatures and data loss. The smoothed surface by the proposed method Fig. 1c has stable curvatures at all around involved edge. The plane is sufficiently flat and the edge is natural. The data loss does not occur due to the approximation at the region without observation.

3.2 Experiments with Various Signals to Noise Ratio

The experiment with various signals to noise ratio is shown in Fig. 2. Figure 2a is the original happy Buddha range image from the Stanford 3D scanning repository. We add the Gaussian noise with different variance on Y-axis value that is represented various signal to noise ratio in Fig. 2b. The Gaussian filtering was applied with 3 by 3 window and 0.5 standard deviation over a number of iteration.

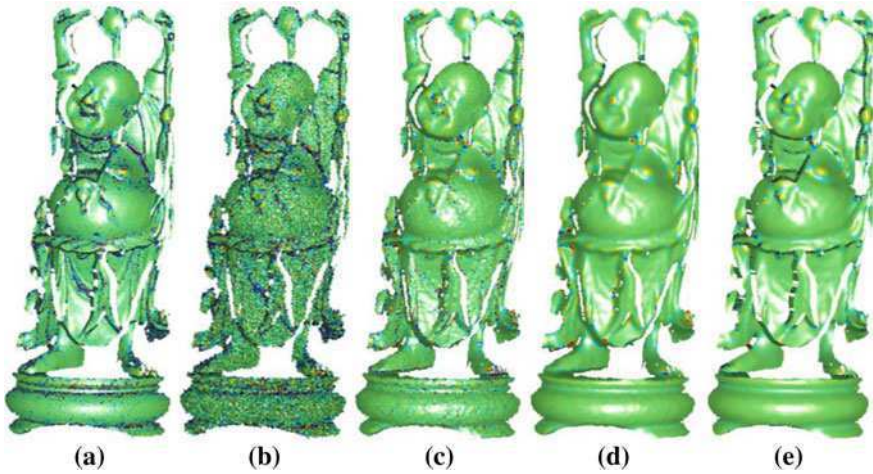


Fig. 2 **a** Original happy Buddha range image, color shows mean curvature, **b** noisy image with Gaussian, **c** Gaussian filtered image with ten iterations, **d** Gaussian filtered image with 30 iterations, and **e** smoothed image with the proposed method

Figure 2c is the result of with 10 iterations and Fig. 2d is the result of with 30 iterations. The upper part of Fig. 2c is nearly smoothed while the bottom is still rough because of the different noise level. The bottom Fig. 2d is smoothed and it follows that the upper is over smoothed. Our result Fig. 2e is shown best smoothing surface from the input with various signal to noise ratio as a result of the different smoothing strength.

4 Conclusions and Future Work

We propose the method using the probability distribution each observed point that involves the accuracy of 3D imaging device. The range image of on-the-fly 3D imaging device contains a range of noise levels in a scene. The smoothing methods that use one smoothing parameter are not satisfied to the on-the-fly range image. We solve the smoothing the range image using Gaussian observation model and unit tangent vector equality; we formulate to linear system with regularization technique for on-the-fly 3D imaging devices. The experiments demonstrated that the method smooths out the surface with various signal to noise ratio, and the surface inherit appropriate curvatures to forms of surfaces. Future work will study the combine of the moving least squares to solve the linear system with regularizer instead of the least squares. Other study is unity of polynomial basis in place of raw observed point.

Acknowledgments The authors gratefully acknowledge the support of Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0025512) and reviewers' comments.

References

1. Boyer KL, Srikantiah R, Flynn PJ (2002) Saliency sequential surface organization for free-form object recognition. *Comput Vis Image Underst* 88(3):152–188
2. Habak C, Wilkinson F, Zakher B, Wilson H (2004) Curvature population coding for complex shapes in human vision. *Vis Res* 44(24):2815–2823
3. Moreno AB, Sanchez A, Frias-Martinez E (2006) Robust representation of 3d faces for recognition. *Int J Pattern Recognit Artif Intell* 20(8):1159–1186
4. Carr JC, Beatson RK, McCallum BC, Fright WR, McLennan TJ, Mitchell TJ (2003) Smooth surface reconstruction from noisy range data. In: *Proceeding of the 1st international conference on Computer graphics and interactive techniques in Australasia and South East Asia*, pp 119–126. ACM, Melbourne, Australia (2003)
5. Levin D (2003) Mesh-independent surface interpolation. In: *Geometric modeling for scientific visualization*. pp 37–49
6. Perona P, Malik J (1990) Scale-space and edge detection using anisotropic diffusion. *IEEE Trans Pattern Anal Mach Intell*, 12(7):629–639
7. Tasdizen T, Whitaker R, Burchard P, Osher S (2002) Geometric surface smoothing via anisotropic diffusion of normals. In: *Proceeding of the conference on visualization*. IEEE Computer Society, Boston pp 125–132 2002
8. Davis TA Algorithm 8xx: Suitesparseqr, a multifrontal multithreaded sparse qr factorization packag., submitted ACM TOMS
9. Davis TA: Multifrontal multithreaded rank-revealing sparse qr factorization, submitted ACM TOMS
10. Gray A (1993) *Modern differential geometry of curves and surfaces*. Studies in advanced mathematics. CRC Press, The Netherland

Group-Aware Social Trust Management for a Movie Recommender System

Mucheol Kim, Young-Sik Jeong, Jong Hyuk Park
and Sang Oh Park

Abstract This paper presents an interactive movie recommender system for constructing an intelligent home network system. The proposed model is based on a group-aware social trust management, one of the new paradigms for personalized recommendation. In this paper, we show the concept model of group-aware social networks for the proposal and a prototype implementation.

1 Introduction

A Social network expresses the concept of psychological and social relationships between individuals or groups as networks. It is a graphically represented social structure consisting of nodes that are tied by one or more specific types of inter-dependency [1–3]. For more than 50 years, various sociology and psychology-based

M. Kim · S. O. Park (✉)
School of Computer Science and Engineering, Chung-Ang University,
221, Heuk Seok-dong, Dongjak-gu, Seoul, Korea
e-mail: sj1st@cs.cau.ac.kr

M. Kim
e-mail: mucheol.kim@gmail.com

Y.-S. Jeong
Department of Computer Engineering, Wonkwang University,
Chonbuk, Iksan, Korea
e-mail: ysjeong@wku.ac.kr

J. H. Park
Department of Computer Science and Engineering, Seoul National University
of Science and Technology, Seoul, Korea
e-mail: parkjonghyuk1@hotmail.com

studies have analyzed this structure. Since the late 1990s and the growth of the World Wide Web (WWW), typical social network services (blogs, collaborative filtering systems, online gaming, etc.), key concepts in the new user-oriented web (web 2.0), have resulted in untold user generated multimedia contents [4–6]. The explosive growth of multimedia data leads to increased time and effort being required to search for content. Recently, personalized recommender systems that suggest product items such as films, music and books according to the online consumer's tastes have attained a lot attention and there has been increasing research interest in such recommender systems [7].

A recommender system involves making automatic predictions about the interests of a user using information filtering techniques. Typically, the collaborative filtering approach and the content-based approach are used to filter information about a user. Collaborative-filtering approaches compute similarity between users based on users' preferences and recommend items which are highly rated by similar users. Content-based approaches recommend items with similarity between items and do not use any preference data [8]. A social network is mainly employed in domains where the human notions of trust and reputation are significant, such as security systems, recommender systems and online transaction systems [9]. In particular, a group-aware social network that effectively models interactions between group based influences and behavioral patterns is well suited to the collaborative filtering approach that collects taste/preference information from many users.

This paper proposes an interactive recommender system for movies that operates in intelligent home network environments. The proposed system supports interactions between users and service providers by exploiting a social network that is created based on the users' preferences.

The rest of this paper is organized as follows: [Sect. 2](#) presents the architecture of the proposed interactive recommendation system. [Section 3](#) describes the group-aware social trust management approach. In [Sect. 4](#), we implement a proposed prototype system. Finally, conclusions and recommendations for future work are given in [Sect. 5](#).

2 Related Work

Studies on social networks have been continued in the social science and psychology fields, and with the development of the Web since 2000s, researchers have become interested in online social networks. Many studies on social networks are developing mainly in three areas: the development of social network models, analysis methods, and applications.

There have been many studies that attempted to specify the degrees of relationships by defining social trust models [10–12]. Golbeck [10] proposes a trust model that is appropriate for online community using the social and personal preferences of users. Kim and Han [12] proposed a trust model that incorporates in the existing trust model the element of uncertainty that the user's trust cannot be

ascertained. Meanwhile, various studies related to the analysis of social networks proposed methods to effectively analyze and mine social networks which are expressed as relational networks. Barabasi and co-workers [13, 14] focus on analyzing graphs expressed as relational networks. Using graph analysis and trust models, [15] has been conducted on visualization. Furthermore, [16–18] propose maintenance methods for dynamically changing relationships between users in social networks. [19–21] propose methods to predict the newly created links for relationship maintenance by applying various classification techniques [22] based on the attributes of users and their existing link relationships. In addition, studies intended to apply social networks to diverse areas such as e-mail spam detection [23–25] and recommendation systems [10, 26–29] have been actively conducted. Studies related with search and recommendation systems utilizing social networks have proposed methods to filter social network information using indirect information [10, 27, 28]. However, these filtering approaches bring about results that add to sparsity problems. Therefore, recent studies of social networks should focus on solving these problems not only by filtering approaches but also by extending relations in social networks [26, 29].

3 Recommender Architecture

Figure 1 depicts the architecture of the proposed interactive recommender system based on a cognitive social network model. The proposed system consists of two main modules: a group-aware social trust management module and an interactive recommendation module. The group-aware social trust management module collects user profiles and users' behavioral interaction data from dynamic media sources and incorporates the recognized user interests or preferences in a social network. The interactive recommendation module analyzes user preferences in the organized social network and service provider's content so as to recommend content items that are likely to be of interest to the user. Information associated with the social network and with the service provider's content is retrieved from databases the Social Network DB (SN DB) and the Movie Content DB, respectively. The user's choices with regard to personalized recommendations are fed back to the group-aware social trust management module so that the recommender system is updated for improvement (i.e., the system learns over time from customers).

4 Group-Aware Social Trust Model

This section describes the group-aware social trust management module that is the core of the proposed interactive recommender system. The proposed system dynamically captures user interests and preferences by creating and maintaining a social network that is built based on a collection of consumer profiles.

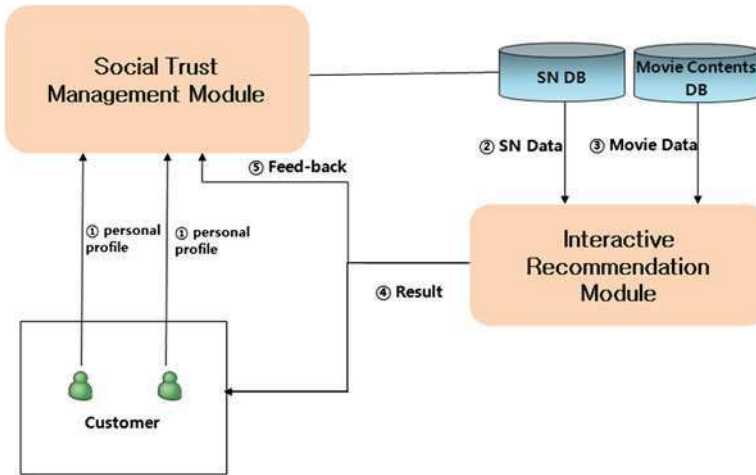


Fig. 1 The proposed recommender system architecture

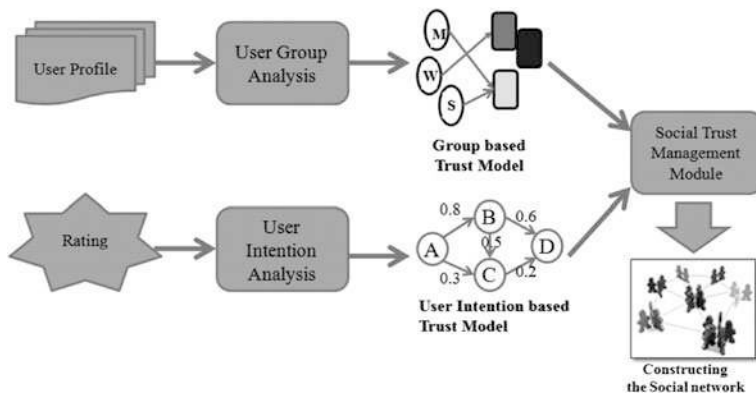


Fig. 2 The Proposed group-aware social trust management process

The group-aware social trust management module in Fig. 1 is composed of three subcomponents—a User Group Analysis Manager, a User Intention Analysis Manager and a Social Trust Management Module (see Fig. 2). The User Group Analysis Manager extracts attributes from the user profile and organizes user groups with the user characteristics. The Influence Determinant Manager defines a personalized influence model. The User Intention Analysis Manager extracts user behavioral information which represents user interactions and intentions. The Social Trust Management Module creates and maintains a social network.

Information attributes that are extracted from user profiles and past user interaction data to identify user preferences are: gender, job, age and user ratings. The User Group Analysis Manager generates a personalized influence model based on the extracted information. The Social Network Manager combines the

Fig. 3 Implementation of the proposed system



interaction information from the User Intention Analysis Manager and the group based influence information from the User Group Analysis Manager and constructs a social network.

5 Implementation

A prototype of the proposed recommender system was implemented in Java so that it can be applicable in a variety of domains (e.g., web or mobile application environments). In addition, a dataset gathered in MovieLens [30], a movie recommendation website, was used to provide sufficient user profile and rating data in the experiment.

In the proposed system, a user can enter his/her profile (gender, job and age). Based on the user's profile, the proposed system captures the user's preferences and recommends films in which the user might be interested, as shown in Fig. 3. Films that have high user preferences are displayed along with a poster, a brief description of the film, the average rating score and the number of users who have rated the film.

6 Conclusions

In this paper, an interactive recommender system that makes personalized recommendations of movies in home networks is described. The proposed recommender system employs the group-aware social network model that is

regarded as a promising technique to capture the dynamics of socially-mediated information transmission in today's social networking environments. This social network model can systematically analyze user preferences that are in rapid and constant change and can represent their influence in social networks. The paper presents a conceptual model and a prototype of the proposed interactive movie recommender system.

Acknowledgments This research was supported by the IT R&D Program of MKE/KEIT [10035708, "The Development of CPS (Cyber-Physical Systems) Core Technologies for High Confidential Autonomic Control Software"].

References

1. Yager RR (2008) Granular computing for intelligent social network modeling and cooperative decisions. In international IEEE conference "intelligent systems", vol 1, pp 3–7
2. Adar E, Re C (2007) Managing uncertainty in social networks. *Data Eng Bullet* 30:23–31
3. Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. *Science* 323:892–895
4. Rijke M, Weerkamp W (2008) Search and discovery in user-generated text content. In *LNCSE* 4956, pp 714–715
5. Langville A, Meyer C (2006) *Google's pagerank and beyond: the science of search engine rankings*. Princeton University Press, New Jersey
6. Staab S (2005) Social networks applied. *IEEE Intell Syst* 20:80
7. Kim J, Jeong D, Baik D (2009) Ontology-based semantic recommendation system in home network environment. *IEEE Trans Consumer Electron* 55(3):1178–1184
8. Debnath S, Ganguly N, Mitra P (2008) Feature weighting in content based recommendation system using social network analysis. In: *Proceedings of the WWW'08*, pp 1041–1042
9. Golbeck J, Rothstein M (2008) Linking social networks on the web with FOAF: a semantic web case study. In: *Proceedings of the AAAI'08*, pp 1138–1143
10. Golbeck J (2009) Trust and nuanced profile similarity in online social networks. *ACM Trans Web* 3(4):1–33
11. Golbeck J, Hendler J (2006) Film trust: movie recommendations using trust in web-based social network. In *IEEE consumer communications and networking conference*
12. Kim S, Han S (2009) The method of inferring trust in web-based social network using fuzzy logic. In *international workshop on machine intelligence research*, pp 140–144
13. Barabasi AL, Jeong H, Neda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. *Physica A* 311:590–614
14. Dorogovtsev SN, Mendes JFF (2002) Evolution of networks. *Adv Phys* 51:1079–1187
15. Singh L, Beard M, Getoor L (2007) Visual mining of multi-modal social networks at different abstraction levels. *11th international conference information visualization*
16. Bae J, Kim S (2009) A global social graph as a hybrid hypergraph. In: *fifth international joint conference on INC, IMS and IDC*, pp 1025–1031
17. Monclar RS, Oliveira J, Souza JMD (2009) Analysis and balancing of social network to improve the knowledge flow on multidisciplinary teams. *13th international conference on computer supported cooperative work in design*, pp 662–667
18. Bourqui R, Gilbert F, Simonetto P, Zaidi F, Sharan U, Jourdan F (2009) Detecting structural changes and command hierarchies in dynamic social networks. In *advances in social network analysis and mining*, pp 83–88
19. Hasan MA, Chaoji V, Salem S, Zaki M (2006) Link prediction using supervised learning. In *SDM 06 workshop on link analysis, counterterrorism and security*

20. Saito K, Nakano R, Kimura M (2007) Prediction of link attachment by estimating probabilities of information propagation. In LNAI 4694, pp 235–242
21. Liben-Nowell D, Kleinberg J (2007) The link-prediction problem for social networks. *J Am Soc Inf Sci Tec* 58:1019–1031
22. Li C, Biswas G (2002) Unsupervised learning with mixed numeric and nominal data. *IEEE Trans Knowl Data Eng* 14:673–690
23. Yeh C-F, Mao C-H, Lee H-M, Chen T (2007) Adaptive e-mail intention finding mechanism based on e-mail words social networks. In the 2007 workshop on large scale attack defense, pp 113–120
24. Yoo S, Yang Y, Lin F, Moon I-C (2009) Mining social networks for personalized email prioritization. In KDD'09, pp 967–975
25. McCallum A, Wang XR, Corrada-Emmanuel A (2007) Topic and role discovery in social networks with experiments on enron and academic email. *J Artif Intell Res* 30:249–272
26. Huang Z, Zeng D, Chen H (2004) A link analysis approach to recommendation under sparse data. In the tenth Americas conference on information systems, pp 1–9
27. Debnath S, Ganguly N, Mitra P (2008) Feature weighting in content based recommendation system using social network analysis. In World Wide Web conference, pp 1041–1042
28. Walter FE, Battiston S, Schweitzer F (2008) A model of a trust-based recommendation system on a social network. *Auton Agent Multi-Ag* 16:57–74
29. Kim M, Seo J, Noh S, Han S (2010) Reliable social trust management with mitigating sparsity problem. *J Wirel Mobile Netw Ubiquitous Comput Dependable Appl* 1:86–97
30. Papagelis M, Plexousakis D, Kutsuras T (2005) Alleviating the sparsity problem of collaborative filtering using trust inferences. *Trust Management, Proceedings 2005*, vol 3477, pp 224–239

Collaborative Filtering Recommender System Based on Social Network

Soo-Cheol Kim, Jung-Wan Ko, Jung-Sik cho
and Sung Kwon Kim

Abstract In recent years, the use of social network services is constantly increasing. A social network service (SNS) is an individual-centered online service that provides means for users to share information and interact over the Internet. In a SNS, recommender systems supporting filtering of substantial quantities of data are essential. Collaborative filtering (CF) used in recommender systems produces predictions about the interests of a user by collecting preferences or taste information from many users. The disadvantage with the CF approach is that it produces recommendations relying on the opinions of a larger community (i.e., recommendations are determined based on what a much larger community thinks of an item). To address this problem, this article exploits social relations between people in a social network. That is, the recommender system proposed in this article takes into account social relations between users in performing collaborative filtering. The performance of the proposed recommender system was evaluated using the mean absolute error.

Keywords Collaborative filtering · Recommendation system · Social network

S.-C. Kim · J.-W. Ko · J.-S. cho · S. K. Kim (✉)
Computer Science and Engineering, Chung-Ang University, Seoul, Korea
e-mail: sskim@cau.ac.kr

S.-C. Kim
e-mail: sckim@alg.cse.cau.ac.kr

J.-W. Ko
e-mail: jwko@alg.cse.cau.ac.kr

J.-S. cho
e-mail: mfg@alg.cse.cau.ac.kr

1 Introduction

In today's online environments, there exist a variety of social networks made up of individuals or groups called "nodes", which are tied by one or more specific types of interdependency. Like real-world social structures, people are connected to each other in an online social network through many kinds of social relations. For example, various communities and organizations are freely created and run in web-based social networking sites such as *Twitter*, *Facebook*, *Epinions*, *Myspace* and *Cyworld*.

In a SNS, a large amount of information on users' behavior, activity or preferences is created. Note that not all of such information is trustworthy because anybody, who might intentionally or unintentionally supply false information, can participate in a SNS. Hence, recommender systems that help users find information by providing recommendations play a significant role in a SNS [1, 2].

Recommender systems use a specific type of information filtering approach such as content-based filtering, demographic filtering and collaborative filtering. Collaborative filtering used in the recommender system proposed in this article recommends items or users. In item predictions (filtering), items that like-minded users rated as of great value are measured for similarity to identify the set of items to be recommended. This technique does not support the social process of asking a trustworthy friend for a recommendation. The disadvantage of the collaborative filtering approach is that recommendations are made depending on the opinions of others irrespective of their trustworthiness. This approach produces standardized (non-specific) recommendations because the items that are favored by a larger community are constantly recommended, used, and reviewed while other items have little chance to be considered. In such an approach, a truly personalized view of an item using the opinions most appropriate for a given user is less likely to be developed. To resolve this problem, the proposed recommender system finds trustworthy users using social relations in an online social network and performs collaborative filtering with the users weighted by trustworthiness.

In the proposed collaborative filtering recommender system, the Friend of a Friend (FOAF), breadth-first search (BFS) and user's social recognition in the social network are used to connect the users (nodes) of an online social network. The *Epinions* dataset was used to implement the proposed recommender system. In the social network created using the *Epinions* dataset, social relations between users are analyzed and trustworthy users are found by computing the distance between users. The proposed system performs collaborative filtering using the identified trustworthy users [3].

The rest of the article is organized as follows. [Section 2](#) gives a brief description of social networks, recommender systems and collaborative filtering. [Section 3](#) presents the proposed recommender system that exploits social relations between users in a social network in order to improve the performance of the traditional collaborative filtering system. In [Sect. 4](#), the proposed collaborative

filtering recommender system based on social networks is compared to the conventional collaborative filtering system. Finally, [Sect. 5](#) concludes the article.

2 Related Work

2.1 Social Network and Friend of a Friend (FOAF)

A social network service provides means to connect with friends and to share opinions with others. Most social network services are web based and focus on building social relations between people, who share interests and activities. Friendship and social recognition created on social networking sites are important social factors to be considered in recommender systems. In friendship, distant friends linked via the FOAF as well as direct friends are considered. Social recognition, the value that an individual gets from the social network, is determined by the number of friends that the individual has in the network. Friendship and social recognition can be used to identify trustworthy users for a given user in a social network [4].

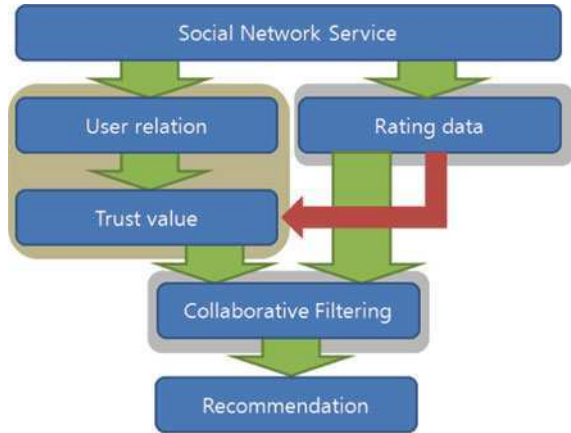
2.2 Recommender System and Collaborative Filtering

A recommender system recommends items or users that are likely to be of interest to the user based on predefined similarity measures. The recommender system proposed in this article recommends items using the collaborative filtering technique. For item recommendations, the collaborative filtering technique first looks for like-minded users and makes predictions (filtering) about the interests of the user using the ratings from those like-minded users [5].

2.3 Breadth First Search (BFS)

In computer science, breadth-first search (BFS) is a graph search algorithm that begins at the root node and explores all the neighboring nodes. Then for each of those nearest nodes, it explores their unexplored neighbor nodes, and so on, until it finds the goal. The BFS can be used to create a social network graph as the set of nodes reached by the BFS form the connected component containing the starting node. The recommender system proposed in this article employs the BFS to create a graph made up of users in a SNS and computes trustworthiness between users in the created graph.

Fig. 1 Proposed recommender system architecture



3 Proposed Method

Figure 1 shows the conceptual structure of the proposed recommender system. In the proposed system, the conventional collaborative filtering technique is enhanced by analyzing social relations between users in a SNS and identifying trustworthy users that are referred to for item recommendations. The red arrowed line in Fig. 1 highlights that the *BFS* algorithm adopted in the proposed system determines the trust value by taking into account both user relations and rating data.

3.1 Identification of Trustworthy Users in a Social

There can be many kinds of ties between the nodes in a social network that is a directed graph. Figure 2 depicts four types of ties: fan, friend, follower and member. The ‘fan’ relationship represents that a given user trust another user, whereas the ‘follower’ relationship represents that the given user is trusted by another user. The ‘friend’ relationship represents that the concerned two users have mutual trust (a two-way tie). In the ‘member’ relationship, the organization to which a given user belongs is considered trustworthy. Figure 3 illustrates the identification of trustworthy users based on the social network graph components (nodes and ties) and rating data.

The $U \cdot U$ array in Fig. 3 represents the trust level between users and the $U \cdot I$ array represents the user’s rating score for the items. The adopted *BFS* algorithm searches through every connected node of a given user in the directed graph. When there is more than one user (node) at the same depth and directed toward a same node, the user that has rated more items is chosen by referring to the $U \cdot I$ array.

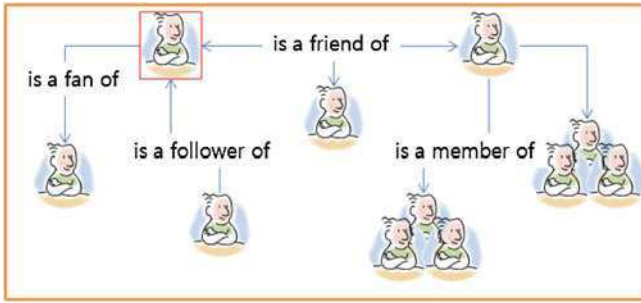


Fig. 2 Relationships in a social network

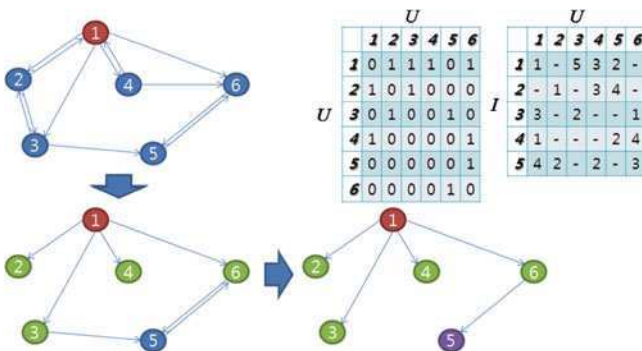


Fig. 3 Computation of trustworthy users in a social network

$$BFS_{(x,y)} = BFS_{(x,Max(y_pre))} + 1 \tag{1}$$

$$aware_{(y)} = \frac{fan_y + follower_y}{2(n - 1)} \tag{2}$$

$$TrustS_{(x,y)} = \frac{|N| - BFS_{(x,y)}}{|N|} \times \alpha + aware_{(y)} \times \beta \tag{3}$$

In Eq. 1, the distance between users is measured using the **BFS** algorithm. $BFS_{(x,y)}$ denotes the measured distance between user x and y . $Max(y_pre)$ in Eq. 1 denotes that when there is more than one user node at the same depth and directed toward a same node, the one with a higher number of rated items is chosen to compute $BFS_{(x,y)}$. Equation 2 counts **fan** and **follower** of a given user and divides the counted number by the number of users so as to compute social recognition that the user has earned in the social network. In Eq. 3, the social relation measures obtained in Eqs. 1 to 2 are transformed into a normalized value in the range

between 0 and 1. $TrustS_{(x,y)}$ represents the trustworthiness between user x and y —value 1 indicates that a given user has a close relationship with the other user who is highly recognized in the social network.

3.2 Recommendation System and Collaborative Filtering

$$S_{(x,y)} = \frac{\sqrt{\sum_{a=1}^n f_{(x,a)} f_{(y,a)} \times TrustS_{(x,y)}}}{\sqrt{\sum_{a=1}^n (f_{(x,a)})^2} \sqrt{\sum_{a=1}^n (f_{(y,a)})^2}} \quad (4)$$

In Eq. 4, $S_{(x,y)}$ represents similarity between items rated by user x and y . a denotes the items examined for similarity and n is the total number of the items. $f_{(x,a)}$ and $f_{(y,a)}$ denote the ratings of item a by user x and y , respectively. The conventional collaborative filtering is extended by adding **weight** denoting the weight given to the similarity computation according to the trust level between users.

$$U_{(x,a)} = \bar{r}_x + \frac{\sum_{y=1}^n S_{(x,y)} f_{(y,a)}}{\sum_{y=1}^n S_{(x,y)}} \quad (5)$$

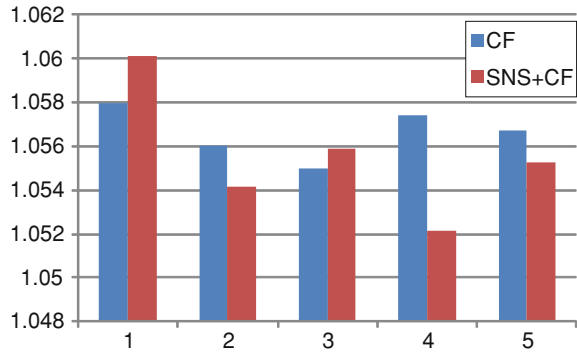
$U_{(x,a)}$ is the predicted rating (preference) of item a by user x (item a has not yet been rated by user x). \bar{r}_x denotes the average rating of items by user x . $S_{(x,y)}$ is the measured item similarity associated with user x and y . $f_{(y,a)}$ represents the rating of item a by user y . n denotes the number of neighboring nodes to be considered.

4 Experiments and Evaluation

4.1 Experimental Data

To perform the experiments with SNS data, the dataset from Epinions.com, a general consumer review site, was used. In the *Epinions* dataset, the number of users was 49,290 and the number of items was 139,738. The number of ratings of the items was 664,824. The *Epinions* dataset has 487,181 social ties. The numerical ratings of an item are in the range $\{1, 5\}$. In terms of social relation, value 1 represents that there is a relationship between users. The absence of the value indicates that there is no relationship. Social relations between users are directed (i.e., they are represented with a directed edge in the graph).

Fig. 4 Performance comparison (proposed vs. conventional collaborative filtering)



4.2 Experimental Method

In order to increase data accuracy, the volume of the *Epinions* dataset was reduced to 1/1000, and data for training and testing was randomly divided (the ratio of the data used for training to testing was 8:2). This operation was repeated five times in the experiments. The performance of the proposed social network-based recommender system was compared to that of the traditional collaborative filtering system. There are several ways to evaluate a recommender system. In this work, the mean absolute error (MAE) was used.

$$MAE = \frac{\sum_{i=1}^n |r_{a,i} - \bar{r}_{a,i}|}{n} \quad (6)$$

$r_{a,i}$ denotes the actual rating of an item by the user and $\bar{r}_{a,i}$ denotes the user's rating predicted by the recommender system. n is the number of items evaluated. The recommender system is 'good' (i.e., prediction is accurate) as the resulting value is close to 0.

4.3 Performance Evaluation

To evaluate the performance of the proposed recommender system, it was compared to the conventional collaborative filtering system. The performance of the proposed and conventional collaborative filtering systems was represented in *MAE*, and it was measured five times (e.g., the operation of randomly dividing the dataset for training and testing was repeated five times) (Fig. 4).

Overall, the *MAE* values are greater than 1, which indicates that the performance of the compared recommender systems is not high. In the first and third performance measures, the proposed system has lower performance (higher *MAE* values) than the conventional collaborative filtering system. On the other hand, the

proposed system performs better than the conventional collaborative filtering system in the second, fourth and fifth measures. The performance evaluation here shows that the proposed recommender system is a solution differed from the traditional collaborative filtering system.

5 Conclusion

This article proposes a social network-based recommender system to solve the problem of relying on the opinions of a larger community in the traditional collaborative filtering technique. In the proposed system, trustworthy users identified by analyzing social relations between users in a social network are used to recommend items. A drawback of the proposed recommender system is that social relations in the range $\{0, 1\}$ and the range $\{-1, 0\}$ are not clearly distinguished due to the use of weight values in the range between 0 and 1. In addition, the BFS algorithm adopted in the proposed system exhaustively searches the entire graph, so it takes a long time to yield recommendations. Trustworthy users that the proposed system identifies for item recommendations differ substantially from similar (or like-minded) users found in the traditional recommender system to make recommendations. In the future, a way to reduce the computational load of the proposed recommender system will be studied to be applicable in mobile environments.

Acknowledgments This work was supported by Basic Science Research Programs through the National Research Foundation of Korea (NRF) grand funded by the Korea government (MEST) (No.2010-0013121).

References

1. Leskovec J, Huttenlocher D, Kleinberg J (2010) Predicting positive and negative links in online social networks. In: WWW 2010, April
2. Adamic L, Adar E (2005) How to search a social network. In: HP Labs
3. Massa P, Avesani P (2007) Trust-aware recommender systems. In: RecSys'07, October
4. Symeonidis P, Tiakas E, Manolopoulos Y (2010) Transitive node similarity for link prediction in social networks with positive and negative links. In: RecSys2010, September
5. Zhang Z, Wang X-M, Y-X Wang (2005) A P2P global trust model based on recommendation. In: Proceedings of the fourth international conference on machine learning and cybernetics, August

Considerations on the Security and Efficiency of RFID Systems

Jung-Sik Cho, Soo-Cheol Kim, Sang-Soo Yeo
and SungKwon Kim

Abstract The RFID system is a contactless automatic identification system using small, low-cost RFID tag. The RFID system can be applied in various fields. For the widespread use of RFID systems, security threats such as user privacy violation and location privacy violation must be addressed. As the major advantage of RFID systems is increased efficiency, RFID security schemes should be designed to prevent security threats while maintaining the efficiency of the RFID systems. This paper identifies concerns regarding RFID security and efficiency that must be considered in building an RFID security scheme.

Keywords RFID system · Privacy · Forgery · Hash · Authentication

This work was supported by Basic Science Research Programs through the National Research Foundation of Korea (NRF) grand funded by the Korea government (MEST) (No.2010-0013121).

J.-S. Cho (✉) · S.-C. Kim · S. Kim
Division of Computer Science and Engineering,
Chung-Ang University, Seoul, Republic of Korea
e-mail: mfg@alg.cse.cau.ac.kr

S.-S. Yeo
Division of Computer Engineering, Mokwon University,
Deajeon, Republic of Korea
e-mail: sangsooyeo@gmail.com

1 Introduction

The RFID system is a contactless automatic identification system using small, low-cost tags. A typical RFID system consists of tags, readers and a back-end server. The tag, generally attached to objects such as products, the human body and animals, has unique identification information (the tag's *ID*). The reader can acquire the identification information from the tag via short-range radio frequency communication. The reader transmits the identification information to the back-end server, and can recognize the information of an attached object. The back-end server manages the identification information contained in the tag, and passes it to the reader [1–3].

Inherent weaknesses in low-cost RFID systems pose security threats such as privacy violation and forgery [5]. They can be solved if the appropriate cryptographic mechanism is applied during communication between the tag and reader. But, as a tag is small and low-cost, the hardware resources are limited. Therefore, it is difficult to apply a traditional cryptographic algorithm to the RFID system [3]. These situations are blocking the wide-spreading of the RFID system.

Presently, there are many researches to solve privacy violation and forgery under RFID system characteristics [3, 5, 6]. However, most previous schemes cannot fully resolve security threats in RFID systems [6]. It is because there are some problems in the generation method of response message by the tag and the transfer procedure. The previous authentication schemes use the random numbers for the indistinguishability and untraceability. But, as these values are exposed in the communication procedure of challenge-response between the tag and the reader, the adversary can easily detect the security value of tag through the eavesdropping and the traffic analysis. Furthermore, for the meaningless request from the adversary, the tag generates the same or easily analyzable response message. As a result, the adversary can identify the output value of specific tag.

The RFID system requires the back-end server to retrieve all tags in the system in order to identify a single tag, and an ideal goal is that they have constant-time tag retrieval complexity. However, most of the existing RFID tag authentication schemes have linear-time tag retrieval complexity, and it is very hard to reduce the retrieval complexity to be logarithmic in the number of tags, or even to be constant. As tag retrieval complexity decreases, it is easier for an adversary to gain access to tag information.

2 RFID Security Threats and Requirements

When a robust security scheme is not applied, security threats involving the tag recognition process of an RFID system are as follows.

- The tag's *ID* is transmitted to the reader via radio frequency communication, without any processing [1].

- The tag transmits its own *ID* when there is a regular query in any reader [1].
- Communication between a back-end server and reader is secure. Communication involving a reader and tag is insecure, because it is based on radio frequency [1].

These characteristics can cause serious information leakage in the RFID system. And the adversary can engage in various illegal behaviors by using the acquired information. Representative attack means used by adversary include attacks aimed at privacy violations such as eavesdropping, traffic analysis and location tracking attacks and attacks aimed at forgery such as replay attacks, spoofing attacks and physical attacks.

In recent years, many studies have been conducted to develop security schemes that prevent RFID security threats. The performance of the proposed RFID security schemes is evaluated against security requirements for RFID systems. Most commonly adopted RFID security requirements in practice are confidentiality [4], indistinguishability [4], forward security [4] and mutual authentication [5].

When an RFID security scheme satisfies confidentiality, indistinguishability and forward security, it is considered “resilient to privacy violation” If the scheme satisfies mutual authentication, it is considered “resilient to forgery”.

3 Related Researches Security Analysis

Previous RFID tag authentication schemes proposed to resolve security threats in RFID systems have some common vulnerabilities with regard to hash functions and *static IDs*. Five vulnerabilities are identified as follows.

- *Intended Request or Meaningless Request* [5]: It is a type of active attack and a method for the location tracking and the traffic analysis. It is closely related with the items below. To acquire the information from the tag, the adversary can send the intended requests or meaningless requests to the tag instead of eavesdropping. The problem is that in some protocol the adversary can expect the response message of the tag and can make the location tracking through it. And, to get the information of the tag, some intended request can be sent.
- *Acquisition of Tag Information with the Same Complexity as the Back-end Server* [5]: The adversary can acquire the response message of tag through the eavesdropping. Or, the adversary can acquire it through the above intended request. And the adversary will try to acquire the information by executing the brute-force attack. At this time, it is necessary to judge whether the attack is effective for the adversary and fatal to the RFID system. First of all, in ordinary *static-ID* based schemes, the computational complexity to recognize the tag at the back-end server is $O(n)$ (here, n is the number of tags). If the cost for the adversary to acquire the tag information through the brute-force attack is $O(n)$, the same as the back-end server, then it can be judged that the attack is effective

regardless of the bit length. Even if it can hardly be made in real time, this attack is available when considering the present performance of computer.

- *Excessive Growth of Computational Complexity for the Back-end Server to recognize the Tag* [5]: If the back-end server requires the excessive computational complexity to recognize the tag, it takes too much time and the efficiency falls down.
- *Response Message of Tag dependent on Random Number* [5]: Recent researches usually make the response message with the random number sent from the reader to the tag and the random number generated by the tag itself. At this time, these two random numbers can be exposed to the adversary through the eavesdropping. The random number of the reader is known in the request procedure and that of the tag is exposed in the response procedure to the back-end server. In this case, the random numbers become good information for the traffic analysis and the brute-force attack by the adversary. Furthermore, as mentioned earlier, when the adversary sends the intended random number as a request, the tag information may be exposed and the trace becomes possible.
- *Synchronization Problem and Location Tracking* [5]: To solve the problem due to the use of *static-ID*, many researches adopt the scheme to update the tag's *ID* or the secret value after the mutual authentication of the back-end sever and the tag. But, in this procedure, the mutual disagreement between the back-end sever and the tag may occur due to the unexpected accident or attack by the adversary.

The vulnerabilities listed above can be serious threats to location privacy. Due to these vulnerabilities, the tag *ID* might be exposed to the adversary and the efficiency of the RFID systems decreases. Therefore, RFID tag authentication schemes should be designed sufficiently considering the above situations.

4 Consideration

Previously proposed RFID security schemes prevent user privacy violations to some extent but they are vulnerable to location privacy violations. This is because they do not address security vulnerabilities related to the random number adopted in the schemes. In addition, some of them have security weakness because efficiency is prioritized over security. This section presents a number of concerns that must be taken into account in designing an RFID security scheme in order to provide strong security and efficiency for RFID systems.

Consideration 1

- The adversary should not be able to extract or guess any information by sending an *Intended Request*.
- The tag should send a different response message in each session irrespective of secret value updates.

The two security concerns in Consideration 1 are related to the two vulnerabilities (*'Intended Request or Meaningless Request'* and *'Synchronization Problem and Location Tracking'*) described in the previous section. The best way to address these concerns is using random numbers, but other security concerns related to the use of random numbers in an RFID security scheme are presented below.

Consideration 2

- The RFID tag makes use of random numbers in creating response messages via hash operation.
- The adversary can perform the brute-force attack using the exposed random number (this concern is widely recognized in previous security schemes).

The vulnerability *'Response Message of Tag dependent on Random Number'* occurs when random numbers are used in creating tag response messages. The adversary can capture the random numbers generated by the RFID reader and tag by eavesdropping communications between RFID reader and tag. Once the random numbers are exposed, among the information used in hash operations, only information that the adversary is not aware of is the tag ID or secret value. The tag ID or secret value remains identical until it is updated, so the adversary can find it out by performing brute-force attacks. Another vulnerability related to such brute-force attacks is *'Acquisition of Tag Information with the Same Complexity as the Back-end Server'*. That is, the complexity of the brute-force attack by the adversary is equivalent to the complexity of tag retrieval at the back-end server. To avoid security threats related to the random number, the following concerns should be considered.

Consideration 3

- In hash operations for reader authentication, the random number generated by the RFID reader should be used along with other tag information as a parameter. The random number generated by the reader should not serve any role in protecting tag messages. If it does, the adversary can find out tag message information using the exposed random number.
- The random number generated by the tag should not be transmitted as it is over the network. It should be processed (encrypted) using a predefined operation that is agreed between the back-end server and tag. In performing such an operation, additional secret values (keys) shared by the back-end server and tag can be used as a parameter.

When the concerns in Consideration 3 are built into a security scheme, security threats related to the use of random numbers described earlier can be avoided. However, there is another concern to consider—efficiency in the back-end server. When a security scheme is built to meet the concerns in Consideration 3, the back-end server needs to perform additional operations to acquire the tag's random number. Computational complexity increases as much as the bit length of the random number. For example, suppose that the computational complexity of tag

retrieval at the back-end server is $O(n)$. For the m -bit random number, the complexity increases to $O(mn)$. In a scheme that stores the information of the previous session for synchronization, the complexity increases up to $O(2mn)$.

The following are concerns related to the problem of decreased efficiency at the back-end server that occurs when the concerns in Consideration 3 are built into an RFID security scheme to eliminate security threats.

Consideration 4

- Security schemes that prioritize efficiency over security send static values to the back-end server for constant-time tag retrieval. These schemes are good in terms of efficiency but highly vulnerable to location privacy violation.
- To be resilient against privacy violation and forgery, most schemes use both tag *ID* and random number in hash operations. In such schemes, the complexity of tag retrieval at the back-end server is $O(n)$. Note that the back-end server performs the hash operation n times.
- There is a trade-off between security and efficiency. Neither of them can be ignored in building a “good” RFID security scheme, so attempts to provide strong security while maintaining efficiency as best as possible are made.
- To build a security scheme that provides high security and efficiency, the hash operation for tag retrieval at the back-end server is replaced with a more lightweight operation. This does not allow constant-time tag retrieval but its tag retrieval accelerates compared to the tag retrieval made using the hash operation. In such a scheme, when the tag creates response messages, it produces messages for tag retrieval and authentication separately. Messages for retrieval are processed using the operation lighter (faster) than the hash operation, whereas messages for authentication are processed using the hash operation. This enables the back-end server to perform the computationally expensive hash operation only once.

If a hash-based RFID tag authentication scheme is built by considering the RFID security and efficiency concerns identified in this section, it can overcome the weakness in previous RFID security schemes.

5 Conclusion

RFID system is the technology approaching us on the basis of advantages such as low-cost and contactless automatic identification. But, due to the most fundamental characteristics that it is small, low-cost and uses the radio frequency, the RFID system can cause the privacy violation and the forgery. A considerable amount of research has been conducted to provide robust security in RFID systems but there are some security threats that are not addressed in previously proposed RFID security schemes. In particular, vulnerabilities related to the random number used in most existing security schemes and efficiency decreases in return for

enhanced security are not resolved in existing security schemes. Hence, this paper identified concerns (considerations) related to these two aspects so that better RFID tag authentication schemes can be developed by taking into account the identified concerns. In the future, methods to further improve the efficiency of RFID security solutions will be studied.

References

1. Finkenzeller K (2002) RFID Handbook 2nd edn. Wiley
2. EPC Radio-Frequency identity protocols class-1 generation-2 (2008) UHF RFID protocol for communications at 860 MHz–960 MHz Version 1.2.0. EPCglobal Inc
3. Juels A (2006) RFID security and privacy a research survey. *Sel Areas Commun* 24(2):381–394
4. Cho J, Kim S, Yeo S (2011) RFID System security analysis, response strategies and research directions. In: Ninth IEEE international symposium on parallel and distributed processing with applications workshops, IEEE Comput Soc, pp 371–376
5. Syamsuddin I, Dillon, T, Chang, E, Han S (2008) A survey of RFID authentication protocols based on hash-chain method. In: Third international conference on convergence and hybrid information technology–ICCIT 2008, vol.2, pp 559–564
6. Yeo S, Kim S (2005) Scalable and flexible privacy protection scheme for RFID systems. European workshop on security and privacy in Ad hoc and sensor networks—ESAS'05, LNCS, vol. 3813, Springer, Heidelberg, pp 153–163

A Development Framework Toward Reconfigurable Run-time Monitors

Chan-Gun Lee and Ki-Seong Lee

Abstract Time-critical systems are usually loaded with run-time monitors to observe their temporal requirements because there can be timing violations which may trigger fatal damages to people or systems. Since the timing constraints of run-time monitor are non-trivial, it is prone to complicate modifications as well as implementations. We propose a run-time monitor which facilitates to reconfigure monitoring conditions at design time. As the monitoring concerns are well separated in design time, we can expect the system to mitigate complexity in implementation. Our timing monitor is modeled by using xUML in early stage of development process, and specifications of timing constraints are represented by RTL - like expression. The modeled monitor is transformed into the AOP code by MDA approach. We demonstrate the effectiveness of our approach by showing a case study and analyzing our work.

Keywords Time-critical system · Run-time monitor · Timing constraint · Reconfigurable monitor · MDA · xUML

C.-G. Lee (✉) · K.-S. Lee
Department of Computer Science and Engineering,
Chung-Ang University, Seoul, Korea
e-mail: cglee@cau.ac.kr

K.-S. Lee
e-mail: goory00@gmail.com

1 Introduction

In time-critical systems, as well as functional correctness, temporal correctness is also an important requirement. Typically, such requirements are specified as timing constraints. Since the violation of timing constraints may cause fatal consequences on a person or system, it should be monitored in run-time.

For implementing a run-time monitor, an in-lined reference approach is commonly used. It is a method that monitoring codes and observed system codes are executed on the same process. This approach can monitor the system precisely and quickly, however an implementation of the monitor is non-trivial and difficult. Therefore, when the monitoring constraints should be modified, a cost for adapting change of constraints increases highly.

In this work, we propose a reconfigurable monitor for time-critical systems. We abstract timing constraints and define timing monitor model at design time. Our approach ultimately provides a full separation of functional concerns and non-functional concerns such as timing monitoring. As the abstracted monitor model could be simply attached to system architecture, the system designer is able to compose the run-time monitor easily.

In order to have this flexibility of timing monitor design, we have to consider of mitigating load for implementation. In this regard, our system architecture model which has timing monitor, can be transformed to source code by using MDA. Moreover, as generated monitor code is formed to AOP, a separation of monitor concern is maintained in code level as well as architecture level. Especially our runtime monitor is specified by Real Time Logic (RTL) [1]-like expression for considering time-critical properties, we can measure non-trivial timing relations such as deadline or delay time.

The rest of the chapter is organized as follows. In [Sect. 2](#), we discuss monitoring researches on MDA perspective. [Sect. 3](#) presents the overview of our approach and monitor model. Then we details implementation in [Sect. 4](#), evaluates our work in [Sect. 5](#). Finally in [Sect. 6](#) we conclude paper.

2 Related Work

Model-Driven Architecture is a software development methodology which emphasizes the role of the models. The design and specification of a system are platform independently modeled by standardized format and they are transformed into the source code by tool chains.

There have been series of studies for checking the correspondence between the requirement specification defined in the model-design process and the actual implementation. The previous work by Engels et al. [2] illustrated the mechanism to derive the test cases by converting the designed model through graph algorithms. They also suggested to add the corresponding assertions by using Java Modeling Language into the test code. In [3] Gargantini et al. proposed a method of generating

Abstract State Machine(ASM) models from the UML. They presented how to perform the verification and validation of the ASM models for various scenarios. Regretfully, despite the importance of monitoring for complex timing properties, little efforts have been taken until now. In [4] Saudrais et al. addressed the importance of temporal properties. They proposed to generate the monitoring code from Timed Automata model which includes UML model and state definitions. Our own previous work [5] extended the study of MDA toward the monitoring of time-critical systems and enabled the monitor to deal with real-time events and complex timing constraints efficiently.

We note that it is necessary to enhance maintainability of the run-time monitor in order that it actively applies modification of timing constraints which may occur due to various reasons.

3 Reconfigurable Run-time Monitor

In this chapter we present run-time monitor which facilitates to reconfigure monitoring condition at design time. The important aspect is to provide a full separation of functional and non-functional concerns. Where the non-functional concerns are well separated in implementation as well as design time, we can expect the system to mitigate complexity. In this regard, we designed a separate model which has non-functional requirements. When we need to modify the non-functional requirements, we have only to reconfigure specifications, and then remains are done automatically by model transformation [5]. As our work is targeting for time critical system, the non-functional requirements in which we concentrate, are classified with real time properties such as start, end, period, deadline, delay, jitter, resolution and response time etc. [6].

3.1 Generation Flow of Run-time Monitor

Our reconfigurable run-time monitor is generated as followed sequences.

- *System architecture modeling* Functional requirements are designed.
- *Monitor modeling* We design a timing monitor. The monitor model can be attached to system architecture model. In order to measure non-trivial timing relations in run-time, we use RTL-like expressions in monitor model.
- *Monitor script extracting* When we generate the functional system codes by using MDA tool, our monitor model is extracted to annotation script.
- *Generating monitor code* The monitor script is generated into AOP code by using our code generator. The monitor is highly modularized by AOP, separation of non-functional concern is well maintained from design to implementation. When the weaving is done, it checks timing constraints of running system as in-lined reference monitor.

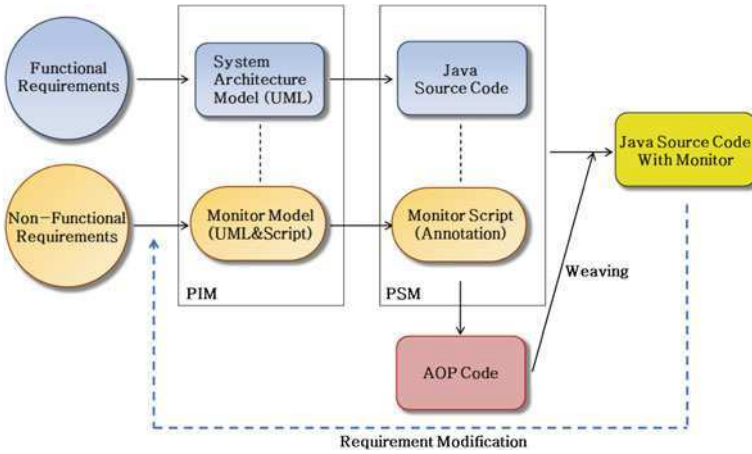


Fig. 1 Reconfigurable run-time monitor generation flow

- *Reconfiguring monitor* When we need to modify timing constraints, we can go back to previous monitor modeling step and modify monitor model. Because remaining steps are done automatically by using tool chain, we can reconfigure the monitor easily.

Figure 1 briefly shows development process of our approach.

3.2 Non-Functional Monitor Modeling

We designed our monitor model by extending Executable UML(xUML) [7] which can be directly generated to suited code for the target platform. We extended stereotypes of xUML to support our monitor and defined constraints model. By using *AnnotationType*, it is possible to compose our monitor with annotation code script. Also, considering separation from functional concerns, we included information for applying AOP. Figure 2 presents the monitor model example.

In order to specify timing constraints, we defined monitor model which should have a script for timing condition as shown in right-hand side of Fig. 2. We extended the constraint specification of annotation type which has been addressed in JavaMOP [8]. In this way we can support specification for complex timing constraints. This script model presents *events*, *condition* and *action* for the timing constraints. The temporal logic in the *condition* is specified by RTL-like expression which is a variant of RTL [1, 9]. RTL-like expressions describe the temporal relations of the real-time systems. Next statement shows a simple example. It is possible to denote a deadline or a delay time among various events. Therefore, we can express non-trivial temporal properties and detect violations.

$$\text{Deadline Constraint Example : } @(A, i) + 3 \geq @(B, i) \quad (1)$$

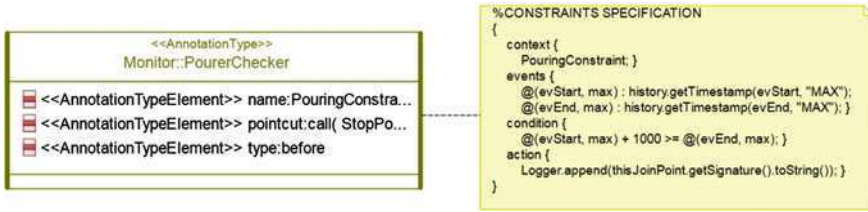


Fig. 2 Example of monitor model

The monitor model combines with the system architecture model. In reconfiguration step, we can modify constraints formula easily for adapting changed requirements. Moreover, considering performances of running systems, we can recompose monitors to tighten or weaken monitoring. Even if we fully reconfigure monitors, it may not influence to the functional system code because monitor and system concerns are separated in code level as well as design level.

4 Implementation

We designed a simple Factory Automation System model for a case study which was inspired by [10]. In this system, a *Pourer* puts products out periodically into a jar on a *Conveyor*. When products in the jar are moved by the *Conveyor*, a *WeighingMachine* checks weights whether it is proper amount. We assume that defective products may come out when the *Pourer* does not work in time, because jars pass continuously along the *Conveyor*. In order to guarantee the *Pourer* observes the rule, we designed a *PourecChecker* which monitors temporal accuracies. Using IBM Rational Rhapsody 7.6 - MDA tool, we composed and transformed a model. The transformed monitor model was analyzed by JavaCC [11] and generated to AOP code. We applied AspectJ [12] for AOP and performed weaving using AJDT which is eclipse plug-in tool for AspectJ. Figure 3 shows an architecture model of the Factory Automation System which has timing monitor at right-hand side. Ultimately, following code is the generated monitor code.

```

public aspect PourecChecker {
    pointcut PouringConstraint(): call(StopPouring());
    before(): PouringConstraint() {
        if(history.getTimestamp(evStart, 'MAX') + 1000 >=
            history.getTimestamp(evEnd, 'MAX')) {
            //Constraint Satisfaction.
        }
        else{
            Logger.append(thisJoinPoint.getSignature().
                toString());
        }
    }
}
    
```

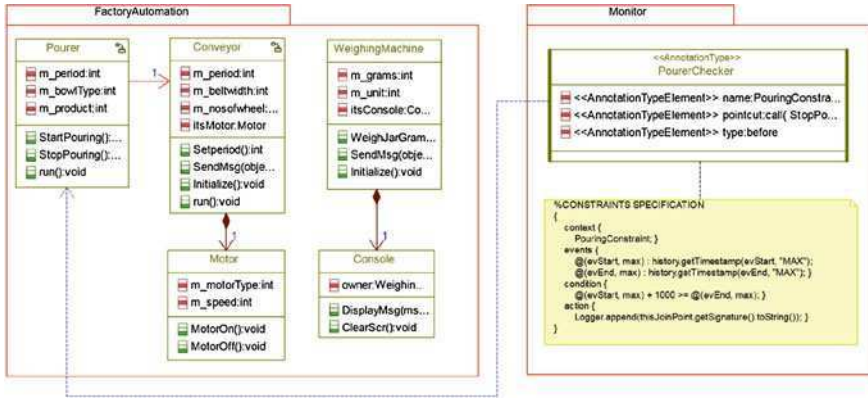


Fig. 3 Factory Automation System with Timing Monitor

5 Evaluation

The proposed run-time monitor has advantages of both MDA and AOP. The *PourerChecker* is generated automatically from its design model, and the actual monitor code is well modularized as shown above. Although the previous work such as java-MOP handles logical temporal conditions, it falls down the ability of expressing real-time properties. However our monitor supports various real-time conditions such as deadlines or delay time by using RTL-like expressions. Especially we can modify the monitoring requirements easily in design level, and it is applied to codes automatically. Therefore, our monitor has high modifiability. This enables the monitor to reconfigure composition when it needs modifications such as changes of a number of monitor, timing condition and observed monitoring position etc.

6 Conclusion

We proposed reconfigurable run-time monitoring system which specifies the timing constraints in its design time and generates the monitor automatically by MDA approach. In order to represent non-trivial temporal relationships among event instances, we supports constraint model based on RTL-like expressions. Especially, when we need to modify the timing requirements, we have only to reconfigure specification, then modification is reflected to monitor by model transformation. For the future work we are planning to apply our monitor system to time-critical domain platform such as Real-Time Specification for Java(RTSJ). Although RTSJ fundamentally supports temporal predictability, its timing

exception handler only covers a unit of single thread. We will extend our monitor to handle complex timing events for considering multiple threads.

Acknowledgments This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (No. 20110013924) and a grant (CR070019M093174) from Seoul R&BD Program.

References

1. Gargantini A et al (2008) A model-driven validation and verification environment for embedded systems. In: Proceedings of SIES
2. Mok AK, Liu G (1997) Efficient runtime monitoring of timing constraints. In: Proceedings of RTAS
3. Gough C-G et al. (2007) Real-Time Java: writing and deploying real-time java applications. 17, 93.<http://www.ibm.com/developerworks/java/library/j-rtj5>
4. Lee C-G et al (2007) Monitoring of timing constraints with confidence threshold requirements. IEEE Trans Comput 56(7)
5. Freitas EP et al. (2007) Using aspect-oriented concepts In the requirements analysis of distributed real-time embedded systems. In: Proceedings of IESS, pp 221–230
6. Chen F, Rosu G (2005) Java-MOP: a monitoring oriented programming environment for java. In: Proceedings of TACAS
7. Engels G et al (2006) Model-driven monitoring: an application of graph transformation for design by contract, In: Proc. of ICGT, vol. 4178, pp 336–350
8. Lee K -S, Lee C -G (2011) Model-Driven monitoring of time-critical systems based on aspect-oriented programming. In: Proceedings of SSIRI
9. Saudrais S et al (2007) From formal specifications to QoS monitors. J Object Technol 6(11):7–24
10. AspectJ WebSite (2010) <http://www.eclipse.org/aspectj/>
11. JavaCC WebSite (2010) <http://javacc.dev.java.net>
12. xUML WebSite (2011) <http://www.kc.com/XUML>

Part VI
Personal Computing Technologies

Web Based Application Program Management Framework in Multi-Device Environments for Personal Cloud Computing

Hyewon Song, Eunjeong Choi, Chang Seok Bae
and Jeun Woo Lee

Abstract Recently, various researchers focus on cloud computing services to be personally provided to service users as a mobile device is smarter. In order to facilitate providing this personal service, we propose a web based application program management (wAPM) framework in this paper. At first, we explain the architecture for the wAPM framework with its function block, and describe the process for managing various application programs installed in multi-devices of users based on the wAPM framework. Moreover, we implement an application, App Manager, using Android devices, and a web server, App Management Server, to manage application programs of registered users, which use the App Manager with their devices. Finally, we show the results from experiments with the implemented App Manager and App Management Server.

Keywords Personal cloud computing · Application synchronization

H. Song (✉) · E. Choi · C. S. Bae · J. W. Lee
Electronics and Telecommunications Research Institute (ETRI),
138 Gajeongno, Yuseong-gu, Daejeon 305-700, Korea
e-mail: hwonsong@etri.re.kr

E. Choi
e-mail: ejchoi@etri.re.kr

C. S. Bae
e-mail: csbae@etri.re.kr

J. W. Lee
e-mail: ljwoo@etri.re.kr

1 Introduction

As the Cloud Computing technology is widely accepted in IT, many researchers study the Cloud Computing as a new service paradigm. Some of them focus on the cloud computing service, Personal Cloud Computing Service, to be personally provided to users since its potential users prefer mobile devices to facilitate their ubiquitous life [1, 2]. The Personal Cloud Computing enables a user with multi-devices, e.g., smart phones, tablet PCs, smart TVs, and so on, to share the data stored in the devices including personal information such as addresses, telephone numbers, e-mails, etc., scheduling information in Calendar application, e-mail data, files such as pictures, videos, documents, and so on. In order to support to make the data sharing available among multi-devices belonging to the user, there are various researches to study the data sharing using synchronization method. [3–6] They consider the data synchronization to provide the consistency of data to users whenever the users access any data in their devices.

In [3], the authors represent a middleware for synchronization, Syxaw (Synchronizer with XML-awareness), in a mobile and a resource-constrained environment. The Syxaw interoperates transparently with resources on the World Wide Web, and provides a model of synchronization including a synchronization protocol and a XML based reconciliation model. Similarly, [5] focuses on the data synchronization using a middleware for synchronization, Polyjuz. The Polyjuz enables sharing and the synchronization of data across a collection of personal devices that use formats of different fidelity in [5]. In addition, Wukong in [4] is a file service supporting heterogeneous backend services, allows ubiquitous and safe data access. However, they do not consider application program installed in user's devices. Besides data consistency, it is needed to provide a consistent environment for executing application in devices.

In this paper, we propose a web based application program management (wAPM) framework to facilitate providing the personal cloud service among multi-devices. First of all, we explain the architecture for the wAPM framework with its function blocks. Additionally, we describe the basic process for managing various application programs installed in multi-devices of users based on the wAPM framework based on the architecture. Moreover, we implement an application, App Manager, using Android devices, and a web server, App Management Server, to manage application programs of registered users, which use the App Manager with their devices. Finally, we show the results from experiments, as a feasibility test with the implemented App Manager and App Management Server.

2 Web Based Application Program Management (wAPM) Framework

As mentioned in a previous section, the wAPM is an application program management framework in multi-devices environments for Personal Cloud Computing. Also, the wAPM is a user-convenient and device adaptive framework for executing application in diverse devices as well as accessing data given by various contents of devices. The wAPM provides (1) application program management among devices belonging to a specific user using web based synchronization and a push process, and (2) device adaptive application program management based on diverse information of users and their devices besides application programs.

2.1 Application Program Management Architecture with Function Blocks

The wAPM framework consists of two basic components: Application Manager in devices and Application Management Server. The Application Manager (AM) is installed in devices belonging to a specific user and communicates with the Application Management Server (AMS) to provide consistency of application programs to the user. The AMS is a web server to support main functions for application program management, for example, maintenance of application programs among devices, information management for users and their devices besides application programs, etc. Figure 1 shows the basic architecture for proposed Web based Application Program Management (wAPM) framework with function block in Personal Cloud Environments.

As shown in Fig. 1, the AMS has six modules including Sync Management, Push Management, User Management, Application Management, User Device Management and Information Base. The Sync Management and Push Management module are to manage synchronization and push process between AMS and AM in devices. Also, the Push Server is related to the Push Management module in AMS. The Information Base contains information of users and their devices as well as applications. This information in the Information Base is managed by User Management, User Device Management and Application Management module, individually.

Similarly, the AM also has six modules containing Sync Handler, Push Handler, Application Handler, Configuration Manager, Information Manager, and Information Base. The Sync Handler and Push Handler module are to control synchronization and push process in devices. Also, the Application Handler module deals with target application programs for AM including web applications and native applications. In addition, the Information Base contains information of

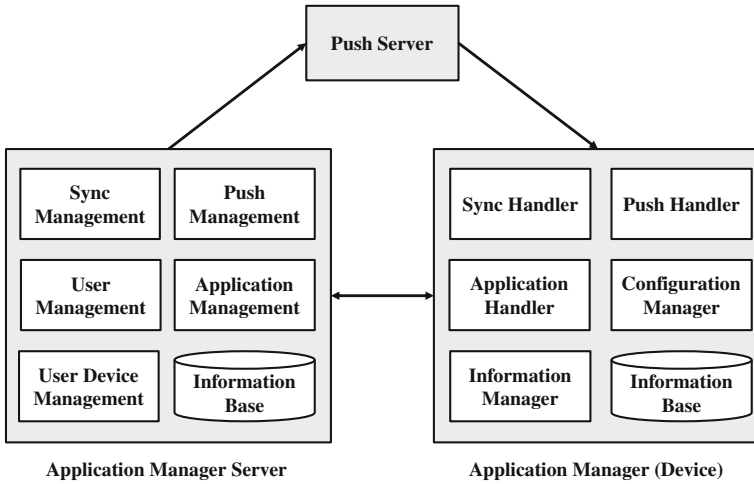


Fig. 1 Functional architecture for application management in personal cloud environments

users and their devices as well as applications, and AM can access and use the information in the Information Base throughout the Information Manager module.

2.2 Application Program Management Process

Based on above architecture, we propose the application program management process. The process is consists of three steps: (1) first synchronization with a server after changing event trigger, (2) handling with an information base and sending a push message, and (3) second synchronization with a server after push trigger. Figure 2 shows the application program management process based on wAPM framework.

In first step, a user changes application program status of a device such as installing a new application program or deleting an existing application program. After then, the AM in the device recognizes the change, and performs the first synchronization process with the AMS. In second step, the AMS updates own Information Base for managing information of users and their devices besides application programs, and searches on a device list belonging to the user in order to determine adequate devices to which the change is applied. After selecting devices to send push message, the AMS requests sending push message to the Push Server, and the selected devices belonging to the user receive the push message because of the change of one device. In third step, the device receiving the push message performs the synchronization process with the AMS for the changed application program. Finally, the devices belonging to the user can maintain the consistency of application programs.

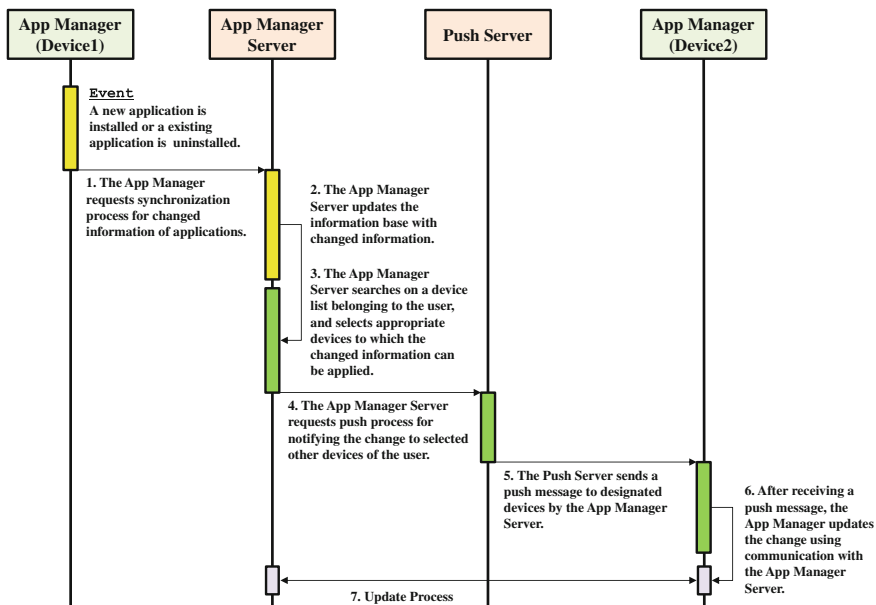


Fig. 2 Application program management process

3 Implementation and Results

As mentioned above, we implement the application based on the proposed framework as a use case using Android based mobile devices, such as smart phones and tablet devices, and web servers. Additionally, we set up the test environment including two mobile devices with the Application Manager, 1 Application Management Server, and 1 Push Server, and test the feasibility of the implemented results—application software of AM and AMS. Finally, we show the results of the test of feasibility briefly.

3.1 Implementation and Experiments

In order to implement the AM application, we use the mobile device with Android 2.2, Proyo, and naturally use the Android Development Tool with Eclipse. Also, we construct a web server for the AMS. In a case of the Push Server, it can be implemented by either integrating to the AMS or separating from the AMS. In this paper, we choose the separated Push Server, and use the existing C2DM (Cloud to Device Messaging) Server provided by Google. The Fig. 3 describes an experimental environment and its scenario, and the implemented wAPM process for AM in a device.

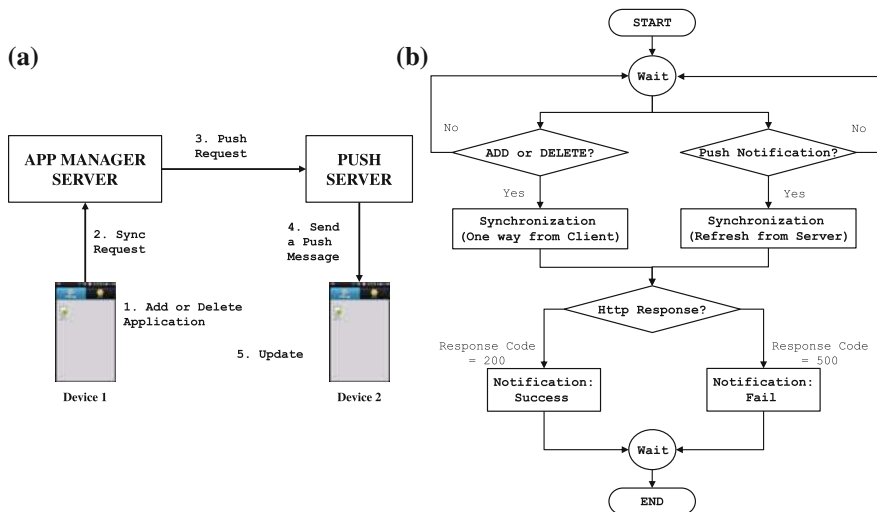


Fig. 3 a Experimental environments and scenario. b WAPM process for AM in a device

The implemented AM application and AMS can support a basic function of wAPM framework for maintaining consistency of applications in multi-devices. For testing the feasibility implemented application software, we set up the experimental environment as shown in Fig. 3a. Also, we test the feasibility according to the basic scenario. At first, an application program is added or deleted in the Device 1. After then, the AM in the Device 1 requests a synchronization process for changed the application program to the AMS. The AMS performs synchronization for changed information throughout its Synchronization Management module, and chooses a proper device belonging to same user to send a push message. After then, the AMS requests sending the selected device, Device 2, a push message to the Push Server. The Push Server sends the push message to the Device 2, and then, the AM in the Device 2 handles the push message and updates the changed information. Finally, the application program added or deleted in the Device 1 can be added or deleted in the Device 2 after the wAPM process.

The Fig. 3b describes the implemented wAPM process for AM in Android devices. This process is implemented as a service daemon in the device, and the triggering event is a change of application program list in the AM or a push alarm message from C2DM server. After trigger, this daemon executes a synchronization process with AMS to which it has already registered. According to a result from the synchronization process, the result notification message is served to a user, and then, the daemon goes back a waiting state again.

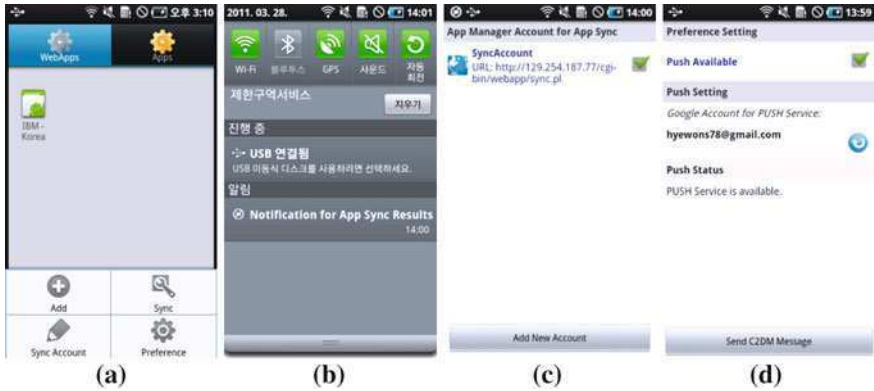


Fig. 4 Implementation results (in Android device)

3.2 Results

Figure 4 shows the implemented AM in the device. Figure 4a represents an ordinary view for the AM, which consists of two categories, Web Application and Native Application. Figure 4b shows the notification action when receiving a push message, and Fig. 4c is a view for registration to the AMS. At last, Fig. 4d is for configuration management and a user can set the preference information related to functions of the AM throughout this activity view.

During above experiment, when we change the application list in the Device 1 as adding or deleting an application program, we confirm the change can be applied to the Device 2 automatically. Namely, we confirm the wAPM process is automatically performed well according to proposed steps as shown in Fig. 4b.

4 Conclusion

In order to support consistency of application programs among multi-devices in a Personal Cloud Computing environment, we propose a web based application program management (wAPM) framework to facilitate providing the personal service. The wAPM framework is constructed with Application Management Server and Application Manager, and provides synchronization and push alarm process for managing various application programs installed in multi-devices of users. Moreover, in this paper, we implement an application, App Manager, using Android devices, and a web server, App Management Server, to manage application programs of registered users, which use the App Manager with their devices. Finally, we show that the implemented App Manager and App Management Server can support feasible personal service to provide consistency of application programs among devices throughout the results from experiments.

Acknowledgment This work was supported by the IT R&D program of MKE/KEIT. [K10035321, Terminal Independent Personal Cloud System].

References

1. Ambrust M, Fox A, Griffith R, Joseph AD, Kats RH, Konwinski A, Lee G, Patterson DA, Rabkin A, Stoica I, Zaharia M (2009) Above the clouds: a Berkeley view of cloud computing. UCB/EECS-2009-28
2. Ardissono L, Goy A, Petrone G, Segnan M (2009) From service clouds to user-centric personal clouds. In: IEEE international conference on cloud computing, pp 1–8
3. Lindholm T, Kangasharju J, Tarkoma S (2009) Syxaw: data synchronization middleware for the mobile web. *J Mob Netw Appl* 14(5):661–676
4. Mao H, Xiao N, Shi W, Lu Y (2010) Wukong: toward a cloud-oriented file service for mobile internet devices. In: IEEE international conference on services computing (SCC), pp 498–505
5. Ramasubramanian V, Veeraraghavan K, Puttaswamy KPN, Rodeheffer TL, Terry DB, Wobber T (2010) Fidelity-aware replication for mobile devices. *IEEE Trans Mob Comput* 9(12):1697–1712
6. Yang H, Yang P, Lu P, Wang Z (2008) A syncML middleware-based solution for pervasive relational data synchronization. In: IFIP international conference on network and parallel computing, pp 308–319

Hands Free Gadget for Location Service

Jinho Yoo, Changseok Bae and Jeunwoo Lee

Abstract The paper is related to how to implement the gadget system which includes the position-aware technology. This paper proposes position services which consist of indoor positioning and outdoor positioning. This system uses sensor devices for position recognition and communication device for the transmission of data. This research provides position calculation method for position recognition. In addition to these functionalities, this system supports low power using the low power policy of main processor.

Keywords Embedded · Hands free · Low power

1 Introduction

As more and more hardware technology improves its performance, the hardware reduces its size and price and increases complexity. This paper proposes the gadget as a position aware service system using some position sensors and communication module. The position aware service system can support subscriber's location service. This technology is used in applications like theme parks, expos and shows.

J. Yoo (✉)
Division of Information and Communication,
Baekseok University, Cheonan, Korea
e-mail: yoojh@bu.ac.kr

C. Bae · J. Lee
Next-generation Computing Research Department, ETRI, Daejeon, Korea
e-mail: csbae@etri.re.kr

J. Lee
e-mail: ljwoo@etri.re.kr

This gadget mainly supports location services in an enclosed space and a broad area. This paper will explain position aware system specification, position aware method and service scenario.

2 System Configuration

System configuration consists of hardware system overview and software implementation block. The hardware system overview includes hardware blocks and explains their roles. Software implementation block divides the whole software into small blocks and implements their functionalities of small blocks.

2.1 System Overview

This system produced by this research is for location services and has system hardware and software components for its services. At first, the system includes the modules for position services. Main processor controls their modules for their services. System provides its storages and several input/output peripherals needed for the applications executing on the main processor. Overall system configuration is viewed in Fig. 1.

2.2 System Modules

System modules consists of main processor block, position recognition block, GPS device management block and memory management block.

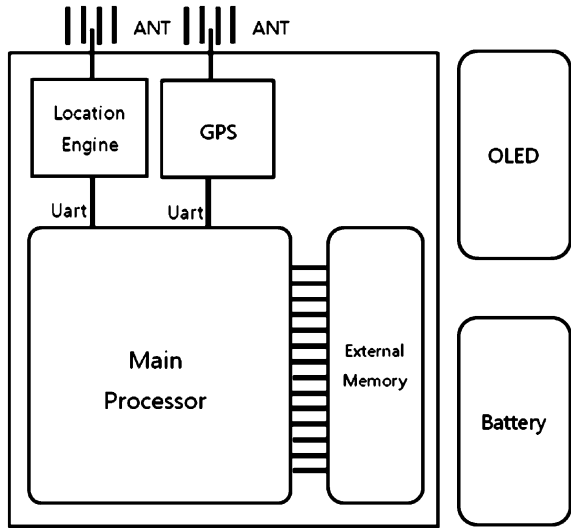
2.2.1 Main Processor Block

The program related to main processor includes system startup, drivers software for each device. System startup program is responsible for processor initialization, memory initialization and board support initialization modules. The scheduler assigns time period for process functions and executes the program according to scheduling policy.

2.2.2 Position Recognition Block

This block has the basic modules for position recognition. This block stores the current positions consistently. This block refer to last position stored when the gadget enters into blind area. This block calculates the values from cc2431 and applies filter algorithm to their values.

Fig. 1 System overview



2.2.3 GPS Device Management Block

GPS device makes one output position data per one second. This block uses this output position data. This block provides the position data from GPS device to application programs.

2.2.4 Memory Management Block

This block manages RAM and flash memory. System needs to store position data, user account information and so on. This block manages store functions for saving permanent data and does recovery from saved data when system clashes.

3 Position Recognition Module

Position recognition modules include two modules which are indoor position recognition module and outdoor position recognition module. Outdoor position recognition module adopted GPS module whose name is GSD4e and indoor position module adopted zigbee cc2431 module which includes hardware location engine. Main functions of this system are location service and network service for transmission of data.

3.1 Outdoor Position Recognition Module

Outdoor location-aware module uses GPS and its related modules. This research have a choice of GSD4e chip. The CSR made this chip which supports low power modes and mainly used in mobile application like smart phones. Nowadays it is easily available in the position recognition using GPS chips. This chip uses small footprint which is best fit for mobile devices.

GSD4e has position values output which are displayed by WGS-84 format (World Geodetic System, 1984) [1]. This format can be translated into ECEF (Earth-Centered, Earth-Fixed) which is Earth-Centered and fixed coordinates system. And also can be translated into ENU (East, North, Up) [2]. We use conveniently these coordinates. Let the outputs of the GPS be latitude Φ , longitude λ , height h ,

$$\begin{aligned} X &= \left(\frac{a}{\chi} + h\right) \cos \phi \cos \lambda, \\ Y &= \left(\frac{a}{\chi} + h\right) \cos \phi \sin \lambda \\ Z &= \left(\frac{a(1 - e^2)}{\chi} + h\right) \sin \phi \end{aligned} \quad (1)$$

Here, $\chi = \sqrt{1 - e^2 \sin^2 \phi}$

We can make the equations of ECEF coordinates system as followings when the output values changes slightly,

$$\begin{aligned} dx &= \left(\frac{-\alpha \cos \lambda \sin \phi (1 - e^2)}{\chi^3} - h \cos \lambda \sin \phi\right) d\phi \\ &\quad - \left(\frac{\alpha \sin \lambda \cos \phi}{\chi} + h \sin \lambda \cos \phi\right) d\lambda \\ &\quad + \cos \phi \cos \lambda dh \\ &\quad + \left(\frac{1}{4} \alpha \cos \phi \cos \lambda (-2 - 7e^2 + 9e^2 \cos^2 \phi) \right. \\ &\quad \quad \left. - \frac{1}{2} h \cos \lambda \cos \phi\right) d\phi^2 \\ &\quad + \left(\frac{\alpha \sin \lambda \sin \theta (1 - e^2)}{\chi^3} + h \sin \lambda \sin \theta\right) d\theta d\lambda \\ &\quad - \cos \lambda \sin \theta dh d\theta \\ &\quad + \left(\frac{\alpha \cos \lambda \cos \theta}{2\chi} - \frac{1}{2} h \cos \lambda \cos \theta\right) d\lambda^2 \\ &\quad - \sin \lambda \cos \theta dh d\lambda + O(d\theta^3) + O(dhd\theta^2) \end{aligned}$$

$$\begin{aligned}
dy = & \left(\frac{-\alpha \sin \lambda \sin \phi (1 - e^2)}{\chi^3} - h \sin \lambda \sin \phi \right) d\phi \\
& - \left(\frac{\alpha \cos \lambda \cos \phi}{\chi} + h \cos \lambda \cos \phi \right) d\lambda \\
& + \sin \phi \cos \lambda dh \\
& + \left(\frac{1}{4} \alpha \cos \phi \sin \lambda (-2 - 7e^2 + 9e^2 \cos^2 \phi) \right. \\
& \quad \left. - \frac{1}{2} h \sin \lambda \cos \phi \right) d\phi^2 \\
& + \left(\frac{\alpha \cos \lambda \sin \theta (1 - e^2)}{\chi^3} + h \cos \lambda \sin \theta \right) d\theta d\lambda \\
& - \sin \lambda \sin \theta dh d\theta \\
& + \left(\frac{\alpha \sin \lambda \cos \theta}{2\chi} - \frac{1}{2} h \sin \lambda \cos \theta \right) d\lambda^2 \\
& - \cos \lambda \cos \theta dh d\lambda + O(d\theta^3) + O(dhd\theta^2) \\
dz = & \left(\frac{\alpha(1 - e^2) \cos \theta}{\chi^3} - h \cos \theta \right) d\phi + \sin \phi dh \\
& + \cos \phi dh d\theta \\
& + \left(\frac{1}{4} \alpha \sin \phi (-2 - e^2 + 9e^2 \cos^2 \phi) - \frac{1}{2} h \sin \phi \right) d\phi^2 \\
& + O(dhd\theta^2)
\end{aligned}$$

Here, $d\theta$ is $d\phi$ or $d\lambda$

The differences of ECEF coordinates by rotation makes an effect on the coordinates value of the ENU coordinates system. The orientation of ENU coordinates system is made by the rotation of ECEF coordinates system. At first, if it rotates λ

on z axis, ϕ on y axis, then we can get the equation like (2).

$$\begin{bmatrix} de \\ dn \\ du \end{bmatrix} = \begin{bmatrix} -\sin \lambda & \cos \lambda & 0 \\ -\sin \theta \cos \lambda & -\sin \theta \sin \lambda & \cos \theta \\ \cos \theta \cos \lambda & \cos \theta \sin \lambda & \sin \theta \end{bmatrix} \begin{bmatrix} dx \\ dy \\ dz \end{bmatrix} \quad (2)$$

And, the Eq. 2 can be substituted with the Eq. 2 and we can get the equation as followings.

$$\begin{aligned}
de = & \left(\frac{\alpha}{\chi} + h \right) \cos \theta d\lambda \\
& - \left(\frac{\alpha(1 - e^2)}{\chi^3} + h \right) \sin \theta d\theta d\lambda + \cos \theta d\lambda dh
\end{aligned}$$

$$\begin{aligned}
dn &= \left(\frac{\alpha(1-e^2)}{\chi^3} + h \right) d\theta + \frac{3}{2} \alpha \cos \theta \sin \theta e^2 d\theta^2 \\
&\quad + dh d\theta + \frac{1}{2} \sin \theta \cos \theta \left(\frac{\alpha}{\chi} + h \right) d\lambda^2 \\
du &= dh - \frac{1}{2} \alpha \left(1 - \frac{3}{2} e^2 \cos \theta + \frac{1}{2} e^2 + \frac{\alpha}{h} \right) d\theta^2 \\
&\quad - \frac{1}{2} \left(\frac{\alpha \cos^2 \phi}{\chi} - h \cos^2 \theta \right) d\lambda^2
\end{aligned}$$

3.2 Indoor Position Recognition Module

This research uses cc2431 on which hardware location engine embedded. indoor location service requires rough position in indoor area. We filters and estimates the position from the several position sensor data. cc2431 hardware location engine has location aware algorithm. The input parameter of this algorithm is the value of RSSI (Received Signal Strength Indicator). The algorithm makes an result of coordinates on their coordinates system. We can get the position values from the result of coordinates (Fig. 2).

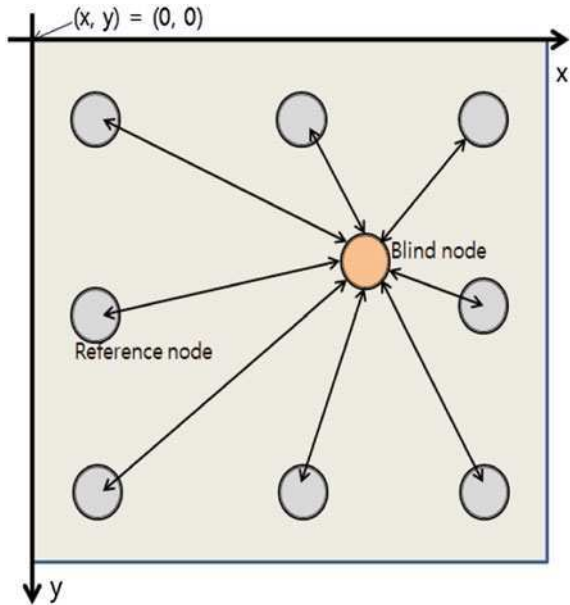
This figure shows a simplified system for location detection. Reference node is a static node placed at a known position [3]. For simplicity this node knows its own position and can tell other nodes where it is on request. A reference node does not need to implement the hardware needed for location detection, it will not perform any calculation at all. A Blind node is a node built with cc2431. This node will collect signals from all reference nodes responding to a request, read out the respective RSSI values, feed the collected values into the hardware engine, and afterwards it reads out the calculated position and sends the position information to a control application.

This research supports two dimensions indoor position recognition. Two dimensions can be extended to three dimensions which use x, y and z. The cc2431 hardware location engine makes inaccurate values of the RSSI. We use filter algorithm for doing prediction and estimation over and over again. At last we get position values which is nearby true value. Actually two dimensions position estimation is sufficient for the service of this research. However, three dimensions is needed for indoor broad space like stepped exhibition hall.

The received signal strength is a function of the transmitted power and the distance between the sender and the receiver. The received signal strength will decrease with increased distance as the equation below shows.

$$RSSI = -(10n \log_{10} d + A)$$

Fig. 2 Indoor two dimension position recognition



- n: signal propagation constant, also named propagation exponent.
- d: distance from sender.
- A: received signal strength at a distance of one meter.

The average RSSI value is simply calculated by requiring a few packets from each reference node each time the RSSI value are measured and calculated according to the equation below.

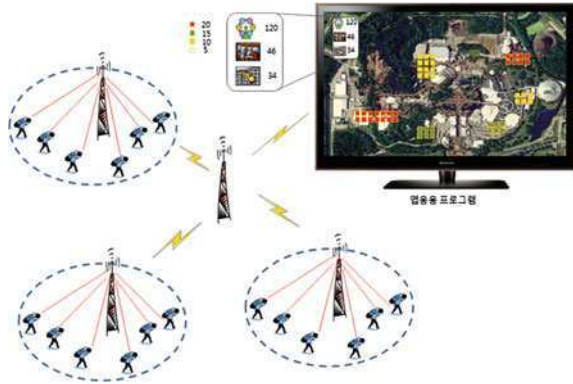
$$\overline{RSSI}_n = \frac{1}{n} \sum_{i=0}^{i=n} RSSI_i$$

If a filter approximation shall be used, this can be done as shown below. In this equation the variable α is typically 0.75 or above. This approach ensures that a large difference in RSSI values will be smoothed.

$$RSSI_n = \alpha * RSSI_n + (1 - \alpha) * RSSI_{n-1}$$

This research does not need position recognition with rapid speed changes. We want rough position at proper time and we will calculate estimated position using special filter algorithm [4].

Fig. 3 Overall service concept



4 Position Aware Service

4.1 Overview and Scenario

The gadget calculates user's position using indoor/outdoor position recognition module and sends its position to position database server over network. The position server can make many applications using user's position data in amusement park or expo. This technology is applicable to electronic ticket system.

The whole system consists of gadget, coordinator access point and server. The gadget signs up for server through coordinator access points. The gadget reports its position to server periodically. The server makes some position services using gadget's reported position data like Fig. 3.

The position-aware service of this research is most pertinent to administration of distribution, traffic control, harmful detection, information appliance, health system, avoiding missing child and ticket management system.

5 Conclusion

Up to now we have looked at the configuration and service of position recognition system. The position recognition can be applied to the applications of the flow control with calculating the degree of congestion in real time. We can analysis the migratory routes from the stored position data and have some applications from the analyzed data. This research proposed the position recognition methodology for theme park, expo etc. In this study we have looked at technical components of overall system for implementation. We need the filter algorithm and the compensation for irregular sensor values depending on the situation in indoor positioning. We can find more precise position data using the map information of the corresponding area.

References

1. Grewal MS, Weill LR, Andrews AP (2002) Frontmatter and index, in global positioning systems, inertial navigation, and integration. Wiley, New York
2. Zhang J, Zhang K, Grenfell R, Deakin R (2003) Realtime GPS orbital velocity and acceleration determination in ECEF system. In: Proceedings of the 16th international technical meeting of the satellite division of the Institute of Navigation (ION GPS/GNSS 2003), Portland, OR, September 2003, pp 1288–1296
3. Aamodt K (2006) CC2431 Location engine, Application Note AN042, SWRA095
4. St-Pierre M, Gingras D (2004) Comparison between the unscented Kalman filter and the extended Kalman filter for the position estimation module of an integrated navigation information system. Intelligent vehicles symposium, 2004 IEEE, 14–17 June 2004, pp 831–835

Biologically Inspired Computational Models of Visual Attention for Personalized Autonomous Agents: A Survey

Jin-Young Moon, Hyung-Gik Lee and Chang-Seok Bae

Abstract Perception is one of essential capabilities for personalized autonomous agents that act like their users without intervention of the users in order to understand the environment for themselves like a human being. Visual perception in humans plays a major role to interact with objects or entities within the environment by interpreting their visual sensing information. The major technical obstacle of visual perception is to efficiently process enormous amount of visual stimuli in real-time. Therefore, computational models of visual attention that decide where to focus in the scene have been proposed to reduce the visual processing load by mimicking human visual system. This article provides the background knowledge of cognitive theories that the models were founded on and analyzes the computational models necessary to build a personalized autonomous agent that acts like a specific person as well as typical human beings.

Keywords Visual attention · Personalized · Autonomous agent

J.-Y. Moon (✉) · H.-G. Lee · C.-S. Bae
Electronics and Telecommunication Research Institute,
218 Gajeongno, Yuseong-gu, Daejeon 305-700, Korea
e-mail: jymoon@etri.re.kr

H.-G. Lee
e-mail: leehj@etri.re.kr

C.-S. Bae
e-mail: csbae@etri.re.kr

1 Introduction

Nowadays, we expect that personalized autonomous agents such as intelligent software agents, humanoid robot, or 3D avatars, act like us for themselves on behalf of us as we have seen them doing in that way in numerous science-fiction films, like Avatar. As the personalized autonomous agents should be able to decide what to do under external conditions from the environment according to decision criteria derived from their users, perception is one of essential capabilities for the agents in order to understand the environment. Visual perception in humans plays a major role to recognize their current situation and interact with objects and entities within the environment by interpreting their visual sensing information. One of important technical obstacles of visual perception is to efficiently process enormous amount of visual stimuli continuing from visual organ in real-time. Human visual systems decide where to focus in the scene through visual attention in order to reduce the processing load of the visual information in the brain. Attention is cognitive process of selectively concentrating on one aspect of the environment while ignoring other things [1].

Visual attention of humans is influenced by not only exogenous saliency originated from physical stimuli regardless of difference between individuals but endogenous influence including goal, intention, emotion, and pre-knowledge of their own. During attention, they attend physically salient regions or object under the bottom-up control in a parallel fashion at the pre-attentive stage and they are controlled by top-down cues at the attentive stage. The accurate mechanism, however, between bottom-up and top-down attention has not been uncovered [2].

To build the personalized autonomous agents, the endogenous attention enables them to search for attended area or object in the scene like a general human being and the endogenous attention enables them to bias and maintain their attention according to cues derived from their current task which is assigned by their goal, intention, or emotion. About three decades, computational models of visual attention have been proposed to adopt visual attention for autonomous agents, computer vision, and image processing as well as to simulate and validate theoretical model based on psychological or neurobiological experiment. Among them, we will focus on models suggested in the engineering area to recognize their architecture and primary components for visual attention of the personalized autonomous agents.

The rest of article is organized as follows. [Section 2](#) introduces theoretical basis of the computational models. In [Sect. 3](#), we describe a general architecture of bottom-up attention model and integration of top-down cues. Finally, we conclude this paper in [Sect. 4](#).

2 Theoretical Basis of Virtual Attention Models

The metaphor spotlight has been widely used to explain spatial attention deployed across space and time. That means the spotlight discloses the hidden area in the dark by deploying the attention there. Numerous physiological experiments

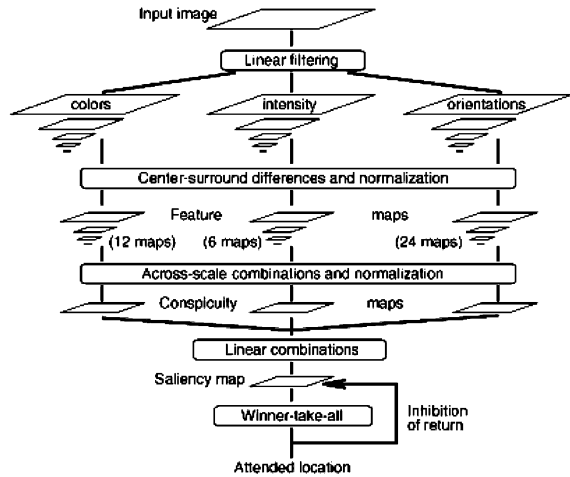
[2] have investigated the characteristics of spatial attention. Attention is shifted across space over time. In addition, the attention can be divided into four or five independent targetable beams and simultaneously allocated to the multiple targets. The metaphor Spotlight, however, has the invariant width of the beam and a zoom-lens model with a variable-width focus was proposed on that account. The size of focus is influenced by the overall load or the difficulty of a task. In addition, the attention should be transformed from 3D to 2D because of the depth in the metaphor Spotlight. Therefore, researchers started to insist that attention should be based on not space but object. Because real visual input is changed over time and looks different owing to occlusion or fragmentation, they suggest object files that represent identity or continuity of an object for object-based attention. Although a lot of researches are improving object-based attention, spatial location in attention plays the most important role for deploying and allocating attention [2].

The most influencing physiological theories for computation models of visual attention are Feature Integration Theory (FIT) [3], Guided search model [4], and CODE theory [5]. In [3], Treisman classified the visual search task according to parallel and sequential processing through human subject experiments. Although the response time to search a target distinguishable from distractors by a single feature is constant regardless of the number of the distractors, the response time to search a target separated by conjunction of two or more features is proportional to the number of distractors. On the basis of this result, the FIT insists that visual information processing should consist of the pre-attentive stage to generate each feature map in parallel and attentive stage to generate a master map after analyzing the feature maps. The FIT considers only bottom-up approach. In [4], Wolfe suggested Guided search model as an alternative model criticizing the early FIT model. In guided map, feature maps are generated by integration of bottom-up local difference and top-down information on the basis of current task. The feature maps are merged into a master map according to their own weights. Lastly, Contour Detector (CODE) theory was proposed in [5]. A scene is processed in parallel and a winner finishing the process first can be conspicuous according to a race model. In the CODE, attended areas are represented as objects by mixture of spatial and object-based attention.

3 Computation Models of Visual Attention

Most computational models of visual attention were biologically inspired and designed on the basis of physiological theories, like the FIT or the guided search. Basically, the models compute several features in parallel and then integrate them into a single Saliency Map (SM), which is a 2D array having high values at salient points compared to their surroundings, by weighted-sum of the features.

Fig. 1 A general architecture of bottom-up visual attention models from [6]



3.1 General Architecture of Bottom-up Models

In [6], Itti proposed the most favorite model that has widely inspired other models of visual attention. The model consists of feature extraction, Conspicuity Map (CM) generation, SM generation, searching for Focus of Area (FOA) steps.

In the step of feature extraction, feature maps about color, intensity, and orientation are calculated by center-surround difference operations individually. In the step of CM generation, feature maps are combined into three CMs for colors, intensity, and orientation respectively. A salient object appearing in some scenes can be masked by a less salient object appearing in almost scenes or noises. That is why the model should do normalization considering the local maximum. Through normalization, the area with the great disparity between local and global maxima gets more salient. In the step of SM generation, the CMs are merged into a SM by across-scale combination operations. In the across-scale combination, the model transforms scale four related to the scale center and the sum of point to point. In the step of searching for a FOA, the model uses Winner-Take-All (WTA) neural network to detect the most salient region. The neurons in the WTA are independent and a winner neuron takes attention. The selective attention is originated from the fire of the winner neuron. Inhibition of Return (IOR) suppresses repeated selection of some already attended areas within a certain period. The FOA was fixed-size circle in [6] but was variable-sized free-form in [7] and [8]. Like in [6], most computational models basically adopt color, intensity, and orientation as primary types of features because physiological and neurobiological experiments proved that they are primary features to visual sensing organ (Fig. 1).

The computational models select some other features for a specific reason. For example, skin color is used to obtain the pointing direction of a finger in [7] and detect face by using skin color database in [8]. However, a higher-level feature map like a facial map using symmetry or ellipse is needed to distinguish arm or

Table 1 Features used in the proposed models

	Primitive features			Motion	Depth	Other features	Higher level-features
	C	I	O				
[6]	O	O	O				
[7]						Edge and corner, skin color (only for detection)	Entropy, symmetry
[8]	O					Edge, skin color	Ellipse, symmetry
[14]	O	O	O				
[15]	O				O	Mean curvature, depth gradient	
[12]	O					Edge	Color contrast, symmetry, eccentricity
[9]			O	O	O (when available)		
[10]				O	O (Stereo disparity)	Horizontal image flow	
[16]	O	O	O				
[17]	O					Edge	Symmetry
[11]	O	O	O	O (not direction)		Edge	
[18]	O	O	O	O		Flicker	
[21]	O	O	O				
[19]	O	O	O				

legs from face in the skin color map. The ellipse and symmetry are also higher-level features that require a low-level edge feature. In addition, motion is adopted for a video (a sequence of images) to detect the spatial change of objects or entities. In [9] and [10], depth is employed as a target selection criterion on the assumption that close objects are more conspicuous than distant ones. The used features are enumerated in Table 1.

3.2 Integration with Top-down Cues

Most visual attention models extend the typical bottom-up computational models from [6]. Table 1 shows previous works how to extend the bottom-up computational models with top-down bias. As shown in Table 2, most models use knowledge of target objects, for example object files and object representation. A specific task of a target search like searching for a man with a red T-shirt imposed in [11]. In [12], the model is working at the different modes during finding a FOA.

Table 2 Integration of bottom-up and top-down approaches

	Experience	Knowledge of scene	Knowledge of objects	Task or goal
CM generation			[13] Adjusting weights during integration FMs into a CM through learned target representation	
SM generation			[12] Managing object files: List of objects information including time, location, features	[11] Using Context-dependent features
Global SM generation	[20] Bayesian probability (weight the location)	[21] Classifying three locations by extracting gist features of a scene	[8] Generating a color FM for tasks and integrating into a global SM [19] Setting weights according to a target in the learning mode	
Searching FOA			[19] Generating a top-down map by weighted-joining excitation and inhibition maps for a global SM	[12] For complex task, three behaviors (Search and track/ Explore/Detect change)

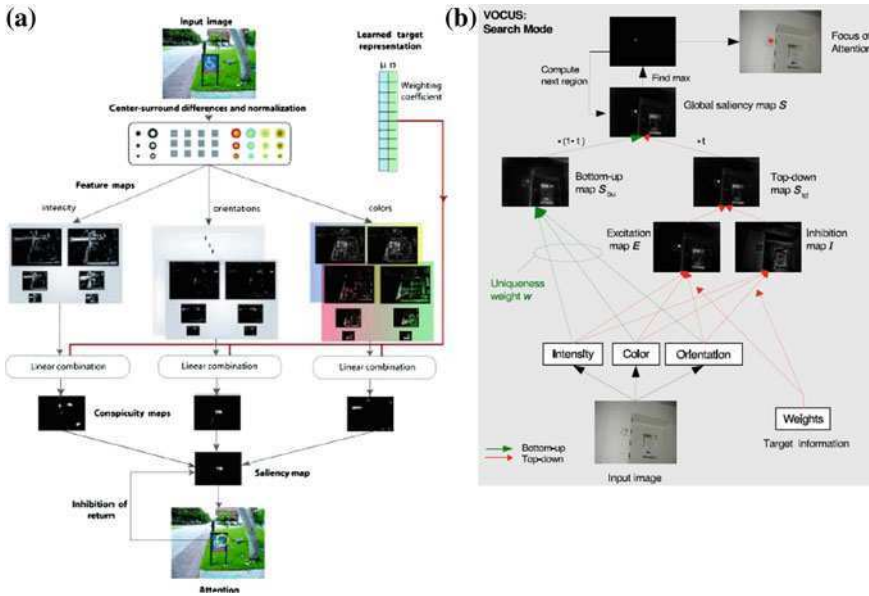


Fig. 2 Architectures of two visual attention models combining top-down cues. (a) a top-down cue biasing model from [13] (b) VOCUS top-down extension from [24]

Among the literature of visual attention models extending top-down cues, we compare two models of different approaches of extending top-down information in Fig. 2.

In Fig. 2a, the model from [24] adopts learned target representation including relevant features of a target object and weight coefficients and uses the representation during combining feature maps into a conspicuity map by linear combination. Due to the target representation, all scenes whose features are similar those of a target object get more salient. In [24], the influence of top-down cue using a target mask with its weight is limited and a less salient object cannot be detected. In contrast to the model from [24], a top-down SM used for a global SM is generated independently by combining excitation and inhibition maps in the VOCUS extension, which is shown in Fig. 2b. This architecture enables less salient objects to be attended by an independent top-down SM. Additionally, this model considers irrelevant features as well as relevant features which other models concentrate on.

4 Conclusion

This article gives an overview of computational models of visual attention from the theoretical basis to the technical fundamentals of their typical architectures and primary components.

The previous works of computational models of visual attention combining top-down and bottom-up were limited to the target search but a model for perception of agent needs a whole cognitive framework including intention, emotion, reasoning, and so on.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MEST) (NRF-M1AXA003-20100029793).

References

1. Anderson JR (2004) *Cognitive psychology and its implications*, 6th edn. Worth Publishers, New York, p 519
2. Wolfe JM (2000) Visual attention. In: deValois KK (ed) *Seeing*, 2nd edn. Academic Press, New York, pp 335–386
3. Treisman AM, Gelade G (1980) A feature-integration theory of attention. *Cogn Psychol* 12:97–136
4. Wolfe JM, Cave K, Franzel S (1989) Guided search: an alternative to the feature integration model for visual search. *J Exp Psychol Hum percept Perform* 15:419–433
5. Logan GD (1996) The CODE theory of visual attention: an integration of space-based and object-based attention. *Psychol Rev* 103:603–649
6. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans Pattern Anal Machine Intell* 20:1254–1259
7. Heidemann G, Rae R et al (2004) Integrating context-free and context-dependent attentional mechanisms for gestural object reference. *Mach Vis Appl* 16:64–73
8. Lee K, Feng J, Buxton H (2005) Cue-guided search: a computational model of selective attention. *IEEE Trans Neural Netw* 16(4):910–924
9. Courty N, Marchand E (2003) Visual perception based on salient features. In: *Proceedings of the 2003 IEEE/RSJ international conference on intelligent robots and systems*, Las Vegas, Nevada
10. Maki A, Nordlund P, Eklundh JO (2000) Attentional scene segmentation: integrating depth and motion. *Comput Vis Image Underst* 78:351–373
11. Moren J, Ude A, Koene A, Cheng G (2008) Biologically based top-down attention modulation for humanoid interactions. *Int J Hum Robot* 5(1):3–24
12. Backer G, Mertsching B, Bollmann M (2001) Data- and model-driven gaze control for an active-vision system. *IEEE Trans PAMI* 23(12):1415–1429
13. Navalpakkam V, Itti L (2005) Modeling the influence of task on attention. *Vis Res* 45:205–231
14. Hamker FH (2005) The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision. *J Compute Vis Image Underst Spec Issue Atten Perform* 100(1–2):64–106
15. Ouerhani N, Hügli H (2000) Computing visual attention from scene depth. In: *Proceedings of the 15th international conference on pattern recognition (ICPR'00)*, vol 1, pp 375–378
16. Peters C, Sullivan CO (2003) Bottom-up visual attention for virtual human animation. In: *Proceedings of the 16th international conference on computer animation and social agents (CASA)*
17. Park SJ, Shin JK, Lee M (2002) Biologically inspired saliency map model for bottom-up visual attention. In: *Proceedings of the BMCV*, pp 418–426
18. Itti L, Dhavale N, Pighin F (2003) Realistic avatar eye and head animation using a neurobiological model of visual attention. In: *Proceedings of SPIE 48th annual international symposium on optical science and technology*, pp 64–78

19. Frintrop S, Backer G, Rome E (2005) Goal-directed search with a top-down modulated computational attention system. In: Proceedings of the of the annual meeting of the German association for pattern recognition DAGM 2005. Lecture notes in computer science (LNCS), Springer, pp 117–124
20. Oliva A et al (2003) Top-down control of visual attention in object detection. In: IEEE proceedings of the international conference on image processing, IEEE, vol I, pp 253–256
21. Siagian C, Itti L (2007) Rapid biologically-inspired scene classification using features shared with visual attention. *IEEE Trans Pattern Anal Mach Intell* 29:300–312
22. Ouerhani N (2003) Visual attention: from bio-inspired modeling to real-time implementation. PhD thesis, Institut de Microtechnique Universit'e de Neuchatel, Switzerland

Mobile Health Screening Form Based on Personal Lifelogs and Health Records

Kyuchang Kang, Seonguk Heo, Changseok Bae
and Dongwon Han

Abstract This paper proposes mobile health screening form based on personal lifelogs as an individual life history and health records linked to continuity of care record. To compose mobile health screening form, we use four categories of data based on biometric screening values, lifestyle patterns, a disease history and an interactive questionnaire. From consumer's perspective, this work may contribute to promote a patient-centered personal healthcare service rather than doctor-oriented conventional medical service. To make more intelligent screening form for the next study, we need to allow for interpretation of relationship between data such as data mining technique.

Keywords Healthcare · Health screening form · Lifelog and data mining

K. Kang (✉) · S. Heo · C. Bae · D. Han
Electronics and Telecommunications Research Institute,
161 Gajeong-dong Yuseong-gu, Daejeon, Korea
e-mail: k2kang@etri.re.kr

S. Heo
e-mail: h7530@etri.re.kr

C. Bae
e-mail: csbae@etri.re.kr

D. Han
e-mail: dwhan@etri.re.kr

1 Introduction

Screening, in medicine, is a strategy used in a population to detect a disease in individual without signs or symptoms of that disease. Unlike what generally happens in medicine, screening tests are performed on persons without any clinical sign or disease [1].

Today, almost all the people have chance of a regular health screening supported by a company or the government every year. As preparing the screening, people may fill up the health screening questionnaire in general.

In addition this regular screening, we may also need question and answer process while we visit and talk to a doctor irregularly. While visiting the doctor, people may describe their symptoms accurately within a short time about 5 min. Because of short consulting time, however, the patient who lacks the expressive power or medical knowledge has difficulty to deliver accurate information to the doctor. In case of people with a chronic disease, they may visit a doctor regularly and describe the current symptom and status in a daily life.

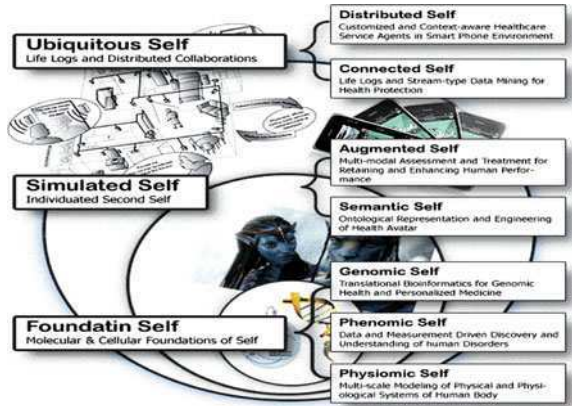
As one solution to describe their own condition and current status of body in current health screening environment, we propose a mobile health screening form leveraging personal biometric information, lifestyles, disease histories and symptoms in this paper. Therefore, this proposal aims to generate the health screening form automatically by filling it up automatically with personal lifelogs and health records. However, because this work is on the beginning stage, we focus on presenting conceptual approach and fast prototype implementation in this paper. To make more intelligent screening form for the next stage, we need to allow for interpretation of relationship between data such as data mining technique.

The remainder of this paper is organized as follows. In [Sect. 2](#), we present related works and the background of this work. [Section 3](#) shows requirements and basic concept and design of a mobile health screening form. In [Sect. 4](#), we present prototype and discussion. Finally, we conclude this paper in [Sect. 5](#)

2 Related Works

In the medical field, we need to gather as much personal information as possible about patients in order to achieve high-quality diagnosis and treatment. Until now, the personal information used in diagnosis and treatment has basically been gathered and used only within medical facilities. It consists of clinical records, test results, medical images, and other such information. However, the medical information that can be gathered within a medical facility is very limited. It would seem that there is a large amount of data that would be useful for medical purposes within the voluminous and varied data gathered in daily life, most of which is

Fig. 1 Conceptual diagram of health avatar



spent outside medical facilities, but such lifelog data has gone unused in most cases [2].

In Korea, SBI (Systems Biomedical Informatics) research center aims to create personalized ‘health avatar’, representing individuals genomic through phenomic reality (or ‘digital self’) using multi-scale modeling and data driven semantics for the purpose of personalizing healthcare [3].

‘The health avatar platform’ will be created as an agent space and health data integration pipeline. ‘Health avatar platform’ will create a space for interacting plug-in intelligent health agents and data analysis toolkits and provide a data and access grid for heterogeneous clinical and genomic data. The health avatar platform will function as an infra-structure for the development and evaluation of intelligent health applications for personalized medicine.

Figure 1 shows the conceptual diagram of health avatar and this work is conjunction with the ‘Connected Self’ of health avatar project supporting lifelogs and stream-type data mining for health protection.

From the perspective of a health screening questionnaire, Chris et al. [4] proposed an adaptable health screening questionnaire that is computer-based lifestyle questionnaire allowing individual doctors to modify the questionnaire to their requirement. Akan et al. [5] developed electronic screening tool providing a graphical user interface with audio outputs for users who may be functionally or computer illiterate. However, these previous trials are only subsidiary function of the mobile health screening form proposed in this paper.

From the lifelog utilization point of view, NTT have studied several subjects [6, 7] enabling lifelogs to be used in the practical service implementation. In case of these NTT’s previous work, we can coordinate these results as a lifestyle category of proposed mobile health screening form.

From a device point of view, there are several commercial lifelogging devices [8–11]. These devices can provide a mobile health screening form with the lifestyle data.

3 Mobile Health Screening Form

This section addresses requirements of a mobile health screening form and presents a basic concept and design.

3.1 Requirement

Basically, the mobile health screening form should reflect personal information, lifestyle patterns, illness history. Furthermore, it is compatible with a conventional paper-based screening questionnaire.

From a personal information perspective, the mobile health screening form may provide with user's biometric screening values such as height, weight, blood pressure and so on.

In the point of lifestyle patterns, the mobile health screening form should provide a scheme that measuring and predicting life patterns by means of life-logging of the user.

For tracking cares of the user, we also should allow for the disease history and track them.

Finally, we also should support compatibility for a conventional health screening questionnaire used in medical center in general.

In order to satisfy these requirements, we can summarize components of the mobile health screening form as four categories of data as followings:

- Biometric screening value: height, weight, blood pressure, heart rate, blood glucose, total cholesterol, HDL cholesterol, LDL cholesterol, triglycerides.
- Lifestyle pattern: sleeping patterns, diet patterns and physical activities.
- History of disease: links to a health portal server or the facility of the family doctor through standardized format such as CCR [12].
- Health screening questionnaire: provides an interactive question and answer tool.

The CCR standard [12] which used in the category of disease history is a patient health summary standard. It is a way to create flexible documents that contain the most relevant and timely core health information about a patient, and to send these electronically from one caregiver to another. It contains various sections such as patient demographics, insurance information, diagnosis and problem list, medications, allergies and care plan. These represent a "snapshot" of a patient's health data that can be useful or possibly lifesaving, if available at the time of clinical encounter.

3.2 Concept and Design

In this section, we describe the concept and design of the mobile health screening form.

Fig. 2 Basic concept of digital health screening form and its operation environment

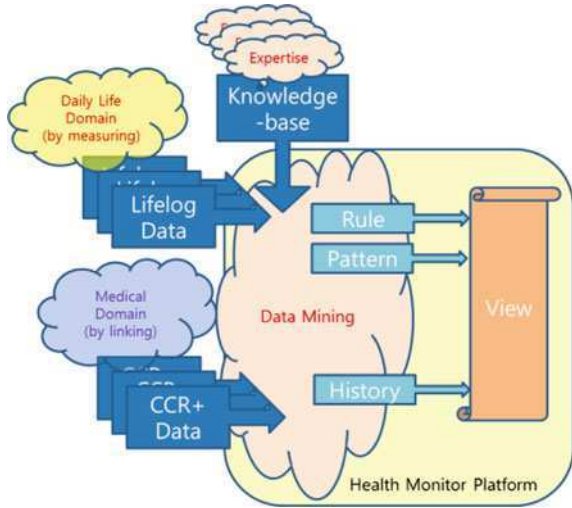


Figure 2 shows the basic concept of digital health screening form and its operation environment.

Basically, the health screening form consists of data linked to a medical domain such as CCR data and data measured by daily life domain such as lifestyle. Then, these data are mixed and mined with knowledge from various expertises. From these processing, we can summarize the user’s life as ‘rule’, ‘pattern’, and ‘history’. These elaborated data can be seen to user through a smart device and transferred to a care giver for an advanced care.

4 Prototype and Discussion

This section presents the initial prototype of the proposed health screening form including a conventional paper questionnaire.

We implemented the prototype on an Android platform that the OS version is 2.2 (Froyo) and tested it on the phone and the tap device.

Figure 3 shows the screenshot of the implemented prototype running on Galaxy Tap device.

The prototype of the mobile health screening form consists of 4 subsidiary categories.

- Biometric category: it represents basic vital index. Currently five items are used (height, weight, heart rate, blood pressure, blood glucose)
- Lifestyle category: it relates with human’s life style such as sleeping, eating and activity. Therefore, these items can be filled by user’s lifelog captured by various devices.



Fig. 3 The screenshots of the prototype implementation. a Biometric category. b Lifestyle category. c History category. d Questionnaire category

- History category: this item is linked to a health portal providing a health record represented CCR standard specification.
- Questionnaire category: it is interactive question & answer system to find various symptoms in the normal living.

This is ongoing work and we are on the initial phase of study. Therefore, we need more studies from many aspects of points such as; (1) how to mine and extract valuable information from lifelogs, (2) how to collaborate with the conventional home & mobile healthcare devices, (3) how to interchange the data between the mobile health screening form and the traditional medical filed: we need more study about standard data interchange language.

5 Conclusion

We described the concept of a mobile health screening form and a fast prototype implementation. To compose the mobile health screening form, we use four categories of data based on biometric screening values, lifestyle patterns, histories of disease and an interactive questionnaire.

Because this work is an initial phase to make a mobile health screening form, we lack of an elaborate algorithm or a breakthrough idea. However, we think this work will contribute to make a personalized and mobilized health screening form reflecting personal lifestyles and histories.

In the future, we plan to continue our research efforts in this filed with the aim of making an intelligent screening form. So we need to allow for an interpretation of relationship between data by means of data mining algorithm.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0018265).

References

1. Wikipedia, Screening (medicine). [http://en.wikipedia.org/wiki/Screening_\(medicine\)](http://en.wikipedia.org/wiki/Screening_(medicine))
2. Ito T, Ishihara T, Nakamura Y, Muto S, Abe M, Takagi Y (2011) Prospects for using lifelogs in the medical field. NTT Tech Rev 9:1
3. National Core Research Center, Health Avatar. <http://healthavatar.snu.ac.kr>
4. Carey-Smith C, Powley D, Carey-Smith K (1993) An adaptable health screening questionnaire. In: Proceedings of artificial neural networks and expert systems, pp 259–260
5. Doruk Akan K, Farrell SP, Zerull LM, Mahone IH, Guerlain S (2006) eScreening: developing an electronic screening tool for rural primary care. In: Proceedings of system and information engineering design symposium, pp 212–215
6. Tezuka H, Ito K, Murayama T, Seko S, Nishino M, Muto S, Abe M (2011) Restaurant recommendation service using lifelogs. NTT Tech Rev 9:1
7. Watanabe T, Takashima Y, Kobayashi M, Abe M (2011) Lifelog remote control for collecting operation logs needed for lifelog-based services. NTT Tech Rev 9:1

8. Microsoft, Introduction to SenseCam. <http://research.microsoft.com/en-us/um/cambridge/projects/sensecam/>
9. ZEO, Sleeping Monitoring Device. <http://www.myzeo.com/>
10. Livescribe, Pen-shaped Gadget. <http://www.livescribe.com/ko/>
11. Evernote, Remember Everything. <http://www.evernote.com/>
12. ASTM E2369–05e1, Standard specification for continuity of care record (CCR). <http://www.astm.org/Standards/E2369.htm>

Remote Presentation for M Screen Service in Virtualization System

Joonyoung Jung and Daeyoung Kim

Abstract We have designed and developed the instant computing using virtualization system with Xen. In this system, several remote I/O devices should connect with the virtualization station dynamically. Multiple virtual machine's (VM) are run in the virtualization station and multiple I/O devices connect with one of multiple VM's. So, user can make a computing environment with I/O devices nearby. In this paper, we propose the remote presentation protocol for M (multiple) screen service.

Keywords Remote presentation · M screen · Virtualization system · Multicast

1 Introduction

Modern computers are sufficiently powerful to use virtualization to present the illusion of many smaller virtual machines (VMs), each running a separate operating system instance [1]. The virtualization system has been made several companies such as Citrix, VMware and Microsoft.

This work was supported by the IT R&D program of MKE/IITA, [2008-S-034-01, Development of Collaborative Virtual Machine Technology for SoD].

J. Jung (✉)

Electronics and Telecommunications Research Institute,
161 Gajeong-dong, Yuseong-gu, Daejeon 305-700, Republic of Korea
e-mail: jyjung21@etri.re.kr

D. Kim

Chungnam National University, 220 kung-dong Yuseong-gu,
Daejeon 305-764, Republic of Korea
e-mail: dykim@cnu.kr

Virtualization system is that a lot of users connect to one server and use applications independently [2]. There has been a technology for a desktop or server virtualization. XenDesktop developed by Citrix System uses para-virtualization technology to improve I/O performance [3]. xVM developed by Sun Microsystems includes desktop virtualization, server virtualization and data center automation technology. It is operated in Solaris environment and supports Microsoft window OS, linux and Solaris as guest OS [4]. PnP-X developed by Microsoft extends the PnP (Plug & Play) function, that is, the network device is managed by PnP [6]. VMware View developed by VMware manages all devices at center and support virtual desktop to user [5].

These technology uses network protocol. For example, Microsoft, N-computing and Citrix make a network protocol to connect virtual server and client such as RDP (remote desktop protocol), UXP (user eXension protocol) and ICA (independent computing architecture) [6–9].

The user can integrate and manage various ubiquitous computing devices as virtual device resource in virtualization system. This system offers the best computing environment to user by combining virtual device resources nearby. The user can enjoy the consistent computing environment anywhere in the virtualization system. However these technologies, such as RDP, UXP and ICA, don't include the M screen service mechanism.

After connecting between the virtualization server and remote I/O devices, the screen data of VM should be sent a remote monitor device or remote multiple screens (M Screen) at the same time.

In this paper, we propose protocol and mechanism for sending screen data to M screens simultaneously.

The rest of the paper is organized as follows. In Sect. 2, we show the virtualization system for instant computing. In Sect. 3, we propose the network system independent M screen. Conclusions are presented in Sect. 4.

2 Virtualization System for Instant Computing

2.1 Structure

We have made the “Virtualization system unionizing remote I/O devices” as shown in Fig. 1. This system has several local service zones (LSZs) and the LSZ is consisted of the virtualization station and several remote I/O clients. User can make computing environment using VM in the virtualization station and remote I/O clients nearby.

The remote I/O client device (RICD) is found in LSZ automatically and is used to construct the computing environment for user. It is a ubiquitous device such as smart-phone and tablet PC that has network function and computing power. It has a protocol to connect with the virtualization station dynamically. It also has I/O resources called I/O clients such as a keyboard, a mouse and a monitor.

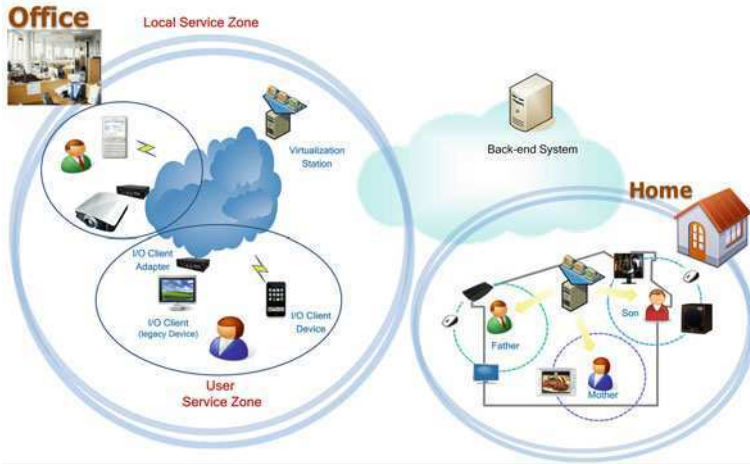


Fig. 1 Virtualization system unionizing remote I/O devices for supporting user friendly computing environment

The remote I/O client adapter (RICA) is a device to help a legacy I/O device that doesn't have network and computing power such as a legacy monitor. The legacy I/O device connects with the virtualization station dynamically through the RICA. A RICA can connect with several legacy I/O devices and each legacy I/O device can connect with each VM of the virtualization station independently. The user can make a computing environment by using various legacy I/O devices and RICD's.

The virtualization station that is in LSZ manages VM's and remote I/O clients, and users. The back-end system manages user's information to maintain user's computing environment even if the user moves to another LSZ.

2.2 *Dynamic Connection Protocol*

To make a dynamic instant computing environment, the remote I/O client should connect with the virtualization station dynamically and automatically. First of all, the virtualization station should find remote I/O client automatically. Second of all, each remote I/O client connects with the virtualization station dynamically.

The structure of virtualization station, RICD and RICA is shown in Fig. 2. Multiple VM's are working in the virtualization station and MRCS(Multiple I/O Resource Connection Service) server tries to connect with remote I/O client automatically. Zone Management manages I/O clients, user and connection state.

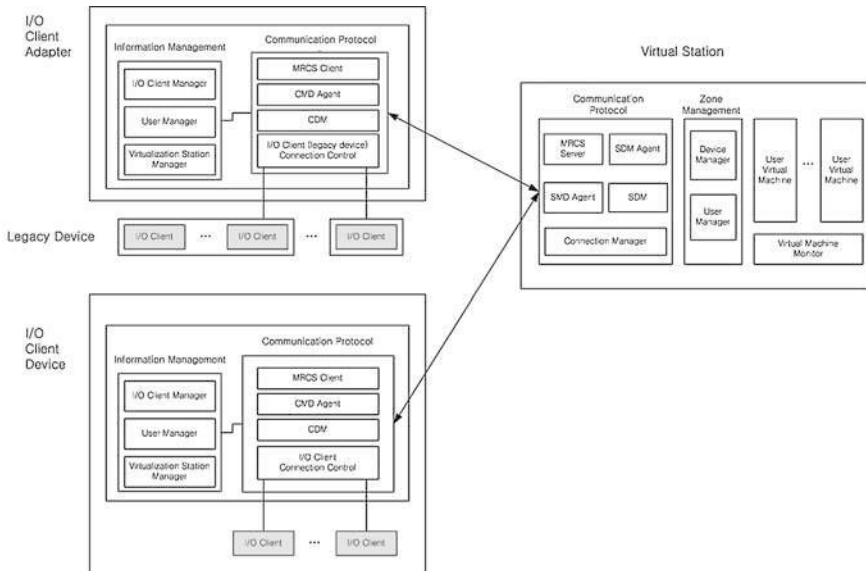


Fig. 2 Connection structure between the virtualization station and the remote I/O client device/adapter

Remote I/O client connect with the virtualization station through MRCS client automatically. I/O Client Connection Control (ICCC) is aware of attachment and detachment of remote I/O clients automatically. Information Management manages I/O clients, users and the virtualization station. The RICD and the RICA can connect with several remote I/O clients.

For automatic connection, the virtualization station has several modules such as advertisement message module, solicitation message module, I/O client profile module and keep alive module in MRCS server.

Get IP address module searches the IP address of the virtualization station automatically. The advertisement message module makes an advertisement message to announce the connection information and broadcasts it in LSZ. The solicitation message module receives the solicitation message from I/O client devices (adapters) and parses the solicitation message to know the contents of it. The connection manager module in the virtualization station manages the connections with multiple I/O clients for control message. If multiple I/O clients request the connection with the virtualization station, the virtualization station should connect with multiple I/O clients and manage connections stably. The profile module in the virtualization station receives the I/O client profile and user profile from I/O clients and user device such as USB memory stick. The Keep Alive module checks periodically if I/O client device (adapter) are disconnected or not.

3 Network System Independent M Screen

There are some applications that multiple users would like to share same screen data, such as an education. For this, the image data of one guest OS is sent to M screens simultaneously.

Multicast is the efficient delivery of data to a group of destinations simultaneously. With multicast, the screen data are delivered as much as possible only once over each link of the network, creating copies only when the links to the destinations split. However, some protocols are needed for multicast. For example, IGMP is used by IP hosts and adjacent multicast routers to establish multicast group memberships. There are several different multicast routing protocols, such as the Distance-Vector Multicast Routing Protocol (DVMRP), Multicast Open Shortest First (MOSPF) and Protocol-Independent Multicast (PIM). The ubiquitous device may have limited resource, so it may not have a multicast protocol such as IGMP. The small office that has a few subnets is hard to support multicast protocol because of router devices and a network operator. So we propose the hybrid M screen method that doesn't use multicast protocol such as IGMP and PIM in the virtualization system however it uses multicast IP address in local network. We call this as "the network system independent M screen in the virtualization system".

3.1 Structure

When a remote I/O client for a screen device is selected for instant computing environment, the SMD (Station Multiple Display) Agent should know whether the remote I/O client is used for M screen device or not. If the remote I/O client is used for first screen device in VM, the unicast connection is made between CDM (Client Display Module) and SDM (Station Display Module) for the screen data. However, if the remote I/O client is used for one of M screen devices in VM, the multicast connection is made between CMD (Client Multiple Display) Agent and CDM in same subnet for the image data. The detail operation will be explained in protocol below.

The SMD Agent sends the control message concerning with M screen to the CMD (Client Multiple Display) Agent of the first screen device.

The SDM Agent receives the I/O client allocation/deallocation message from the SMD Agent and then makes or releases the SDM.

The SDM in the virtualization station connects with the CDM in the I/O client device through TCP connection and then sends the image data to the CDM.

The CMD (Client Multiple Display) Agent receives the M screen message from the SMD Agent through the MRCS Client. It becomes a multicast server for M screen in local network.

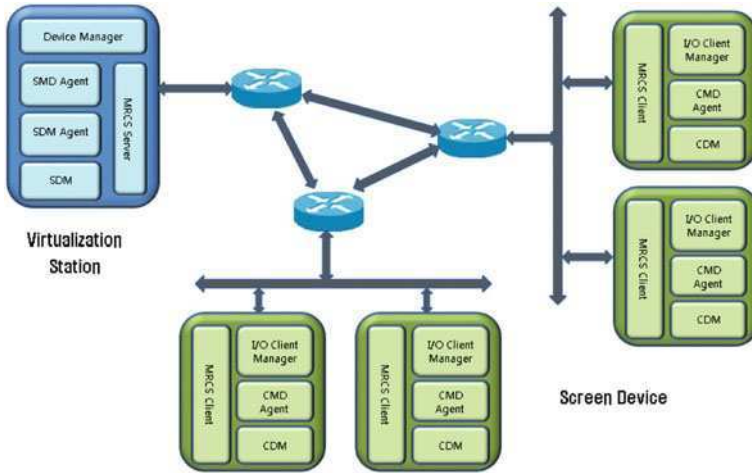


Fig. 3 M Screen System Structure

The CDM (Client Display Module) will send image data to a screen device and another M screen device in local network when the first screen device is used as a multicast server in local network (Fig. 3).

3.2 Protocol

Allocation. The allocation of M screen device is below. The Device Manager knows what a remote I/O client is chosen as a M screen device when user chooses remote I/O clients for making an instant computer. The Device Manager sends information about a screen device allocation to SMD Agent. The SMD Agent receiving allocation information judges whether that screen device is chosen for M screen device or not. The SMD Agent sends the screen device information to the SDM Agent if the screen device is chosen for the single screen device. The SDM Agent makes the SDM and then the SDM connects with the CDM and then sends the image data to CDM. The CDM receiving the screen data sends it to the screen device.

The SMD Agent judges whether the chosen screen device is located in the same subnet with the existing screen device or not if the screen device is chosen for the M screen device.

The SMD Agent sends the screen device information to the SDM Agent if the chosen screen device is located in another subnet. The SDM Agent makes the SDM and the SDM connects with the CDM and then sends the image data to CDM. The CDM receiving the screen data sends it to the screen device.

The SDM Agent judges the chosen screen device is first device for M screen devices when the chosen screen device is located in same subnet with existing screen device. The SMD Agent makes a multicast address and sends the multiple screen information for multicast server to the CMD Agent of the existing screen device when the chosen screen device is the first device for M screen device. The SMD Agent sends M screen information for multicast client to the CMD Agent for the chosen screen device. The CDM received image data from the SDM sends it to the screen device and M screen devices in local subnet. The CDM received image data from the CDM of multicast server sends the image data to the screen device.

The SDM of virtualization station will be multicast server for M screen devices if the M screen devices and the virtualization station are located in the same local network. The SMD Agent sends multiple screen information to the SDM Agent, the existing screen device and the new chosen screen device. The SDM Agent make the SDM sent multicast image data to local network. The CDM of existing screen device disconnects unicast connection with the SDM and receives multicast image data from the SDM and then sends it to the screen device. The CDM of the new chosen screen device receives the multicast image data from the SDM and sends it to the screen device.

Deallocation. The deallocation of M screen device is below. The Device Manager sends the deallocation information about screen device to the SMD Agent.

The SMD Agent receiving the deallocation information judges whether the screen device is an M screen device or not. The SMD Agent sends the deallocation information to the SDM Agent if the screen device is used as a single screen device. The SDM Agent send deallocation information to the CDM and then the CDM disconnects unicast connection with the SDM to stop receiving the image data.

The SMD Agent sends the deallocation information to SDM Agent if the deallocation screen device is located in the same subnet with the virtualization station. The SDM Agent receiving the deallocation information sends the deallocation information to the SDM and the CDM of the deallocation screen device. The connection between the SDM and the CDM is disconnected to stop sending/receiving the image data.

The SMD Agent decreases the multicast client for M screen device if the deallocation screen device is not a multicast server. The SMD Agent sends a Leave Request Message (LRM) to CMD Agent of the deallocation screen device if the multicast client number is one or more. The CMD Agent receiving the LRM stops receiving the multicast image data.

The SMD Agent sends a Leave Request Message (LRM) to CMD Agent of the deallocation screen device and Server Stop Request Message (SSRM) to the CMD Agent of the multicast server device if the multicast client number is zero and the M screen device isn't in the same network with the virtualization station. The CMD Agent receiving the SSRM stops sending multicast image data and the CMD Agent receiving the LRM stops receiving multicast image data.

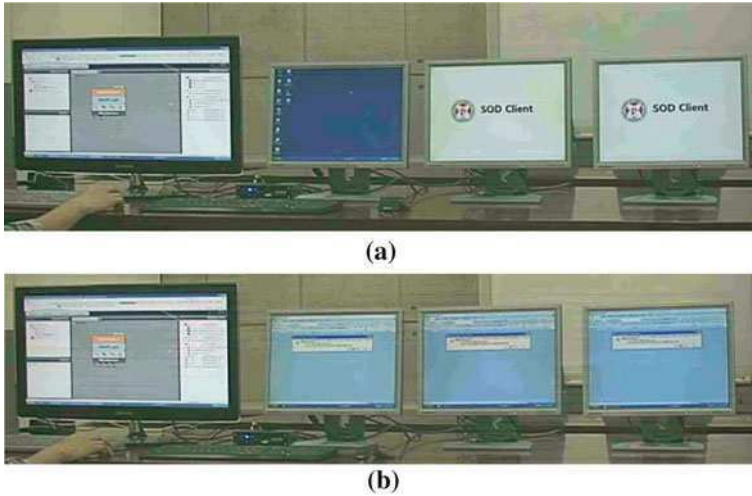


Fig. 4 M screen system. (a) one screen device is allocated, (b) three screen devices are allocated

The SMD Agent sends a Leave Request Message (LRM) to CMD Agent of the deallocation screen device and Server Stop Request Message (SSRM) to the SDM Agent if the multicast client number is zero and the M screen device is in the same network with the virtualization station. The SDM Agent receiving the SSRM sends the deallocation information to SDM and then the SDM stops sending multicast image data. The CMD Agent receiving the LRM stops receiving multicast image data.

The SMD Agent connects with the multicast client device using unicast connection if the deallocation screen device is a multicast server of subnet and the multicast client number is one. The SDM sends image data to the CDM of a multicast client device. The CDM of the multicast client connects with the SDM and stops receiving the multicast image data. The CDM receives image data from the unicast connection with the SDM. The CDM of the multicast server (deallocation device) disconnects with the SDM and stops sending the multicast image data to subnet.

The SMD Agent chooses a new multicast server in subnet if the multicast client number is two or more. The SDM connects with the CDM of new multicast server using unicast and sends image data to it. The SDM disconnects with the CDM of old multicast server. The CDM of old multicast server disconnects with the SDM and stops sending multicast image data to the M screen devices in subnet. The CDM of new multicast server connects with the SDM and stops receiving multicast image data from old multicast server and receives image data from the SDM using unicast and then sends image data to M screen devices in the same subnet using multicast address.

3.3 Implementation

We make an M screen system as seen in Fig. 4. The left screen device is used for managing the virtualization system. The others are used for M screen devices.

First of all, you can see that one screen device is allocated in Fig. 4a. If a user would like to allocate more screen devices, the user can allocate more screen devices using management tool. You can see that three screen devices are allocated in Fig. 4b.

4 Conclusions

We propose the remote presentation protocol for M screen service in the virtualization system. This protocol is one of essential technologies because a VM is connected with M screen devices dynamically and send image data to them simultaneously. This protocol can be used for education system. However, we use the UDP packet for sending image data to M screen devices in same subnet. If the image data packet is lost, the image of M screen devices is broken. So we need a further study to solve this problem.

References

1. Barham P, Dragovic B, Fraser K, Hand S, Harris T, Ho A, Neugebauer R, Pratt I, Warfield A (2003) Xen and the art of virtualization. In: SOSP '03 ACM symposium on operating systems principles. pp 164–177
2. Nieh J, Yang SJ, Novik N (2000) A comparison of thin-client computing architectures. Technical report CUCS-022-00, Nov 2000
3. Citrix White paper (2009) Desktop virtualization with citrix XenDesktop, 06 Jan 2009
4. Sun White paper (2008) Sun xVM virtualization portfolio: virtualizing the dynamic datacenter, Aug 2008
5. VMware White paper Solving the desktop dilemma : with user-centric desktop virtualization for the enterprise
6. Microsoft White paper (2007) PnP-X: plug and play extensions for windows, 11 Jan 2007
7. Ncomputing White paper (2002) Technology white paper, Feb 2002
8. Microsoft White paper (2002) Remote desktop protocol (RDP) features and performance, Dec 2002
9. Tristan Prichardson (2009) The RFB protocol version 3.8, Mar 2009

Lifelog Collection Using a Smartphone for Medical History Form

Seonguk Heo, Kyuchang Kang and Changseok Bae

Abstract In this paper, we present a lifelog system which collects user's daily information for medical care. To achieve this goal, we propose to use a smartphone. It has many kinds of sensors like GPS, accelerometer, and magnetic sensor. In this reason, users can easily obtain their lifelog information whenever they need. By using the stored information, users can find their life habits. The stored information can be used in medical care, too. Experiments show that smartphone has good performance as a life log collector device.

Keywords Lifelog · Smartphone · Medical history form · Mobile devices

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST)(No. 2011-0018265).

S. Heo (✉) · K. Kang
University of Science and Technology, 217 Gajeong Ro,
Yuseong Gu, Dajeon, Korea
e-mail: h75304@etri.re.kr

K. Kang
e-mail: k2kang@etri.re.kr

C. Bae
Electronics and Telecommunications Research Institute,
218 Gajeong Ro, Yuseong Gu, Dajeon, Korea
e-mail: csbae@etri.re.kr

1 Introduction

In recent years, mobile devices such as smartphone have shown great development. These performance (processing ability, battery capacity, sensors accuracy) has increased significantly, and become prevalent. These changes affect the user's lifestyle like Social Networking Site (SNS), life logging, Diary. Smartphone is one of the most useful for the lifelog system. Users can easily record their own life because it has many kinds of sensors and portable. Daily information of these individuals is defined as lifelog. For example, there are sleep, meal, habits, physical exercise, and individual health. Lifelog has been used in many parts. LifeBlog [1] is a good example. It is a multimedia diary using user's photos, videos, and other information (like text message). It is now discontinued.

Objective judgment is very important in medical field. The doctor determines your prescription, considering your status (like weight) and lifestyle (when sleep, how much eat). It has some problem. Patients give their life information to doctor. It is uncertain information because of their subjective judgment. These minor errors sometimes caused large medical accident. In this paper, we proposed lifelog system for medical decision. We focused 'what lifelog information', 'how collect', and 'how appear'.

2 Related Work

There have been many studies for the acquiring of lifelog. "MyLifeBits [2]" is well known example. There are many kinds of ways to collect lifelog. Most of all, multimedia data recording way (Video, Sound, Camera) is good example. It can be used immediately without some processing. And, it is collected by tools like wearable computer [3]. But, User should search where to lifelog in the large multimedia data [4, 5]. Furthermore, wearable computer is big and heavy. For that reason, mobile devices and various sensors are used for collecting lifelog information. Mobile devices are very portable and easy to customizing. If using one sensor to collect lifelog, an error may grow depending on the environment. When using sensor, the association of various sensor can be reduced error [6, 7]. Lifelog types vary greatly. Therefore, it must decide the scope of collection. Collection methods vary depending on the purpose [8–10].

3 Contribution

There were continuously studies in collecting lifelog. After due consideration about this studies, our study needs to meet a number of requirements to successfully pass. Smartphone is suitable device, which satisfy following requirements.

- Mobility
- Having various sensors
- Easy to connect other devices
- Easy to process data

In this paper, lifelog for Medical History Form is defined lifestyle information and physical information. Physical information is data such as weight, height, heart rate, blood-sugar level, and muscle ratio. This is important data in medical care. Thus, data should be collected by medical devices. These devices should interact with Smartphone. And, physical information should not be modified by anyone.

Lifestyle information is data such as caloric intake, sleeping hours, and exercise hours. This information is able to modify, because it can be changed by lifestyle. User can input data with manually or using DB. It is different between physical information and lifestyle information.

These information is converted DB, after they are collected by Smartphone. it can be processed by data mining algorithm. Using these methods is more objective and efficient then pre-methods.

4 Experiment

This experiment is focused to collect lifestyle information. User can collect this information with sensors (like GPS, accelerator) that is included in Smartphone. We collected three important data in lifestyle information. These are meal information, sleep information, and location information.

For this experiment, I used the smartphone on android 2.3 O.S (Gingerbread).

The target smartphone specification is presented in Table 1.

Lifelog contains a lot of daily log. But I collected this information due to the limitation of smartphone. There are lots of information which I get from daily life. But those three informations have many proportion than others. How I collected and gathered the results of each are shown in Fig. 1.

4.1 Meal Infomation

Meal information contain food calories, name, meal start time, meal stop time, and food type. If necessary, nutrients of the diet is the information of interest as well. In order to collect this information, two method can be used. One method is automatically input form the database on a diet. The other method is to manually input. In this experiment, second method was used (Fig. 2).

Order for this section is following.

1. Input meal menu data (meal name, type, calories)
2. Take a picture

Table 1 Target smartphone specification

Samsung SHW-M130L (Galaxy U)	
Display	3.7" AMOLED Plus (800 × 480)
Processor	Samsung S5PC111 (1 GHz)
OS	Android v 2.3 (Gingerbread)
Memory	512 MB (RAM) + 650 MB (ROM)
Weight	131 g
GPS	O
Connectivity	Bluetooth technology v 3.0 USB v 2.0 (high-speed) Wi-Fi 802.11 b/g/n

Fig. 1 Three kind of information



3. Push the activation button
4. Push the inactivation button.

In this section, users have interest to two information. One is the amount of food that users have taken. Second is the time for taking foods. User can determine their eating habit by using these information.

4.2 Sleep Information

Sleep Information contain sleeping start time and sleeping stop time. Using this information, a sleep-related Lifelog information can be obtained. There are many ways to obtain this information, but we were collecting simply using the alarm clock.

This simple method can not get a lot of information, but information accuracy is improved. To obtain Sleep Information, wake-up time should be set first, and then push sleep-start button. After setting the alarm, sleep until the alarm goes off.



Fig. 2 Meal information result screen



Fig. 3 Sleep information result screen

When the set time, the alarm goes off until press the stop button. The data is stored when press the button (Fig. 3).

You can see two kinds of result screen, one is a time line graph and the other one is sleep hours bar graph. Using this information, average sleeping hours and sleep cycles can be determined.

4.3 Location Information

Location Information contain total movement distance, burned calories, and total movement time. We used smartphone GPS in order to get user's location data.

Fig. 4 Compare outdoor and indoor. a outdoors; b indoors

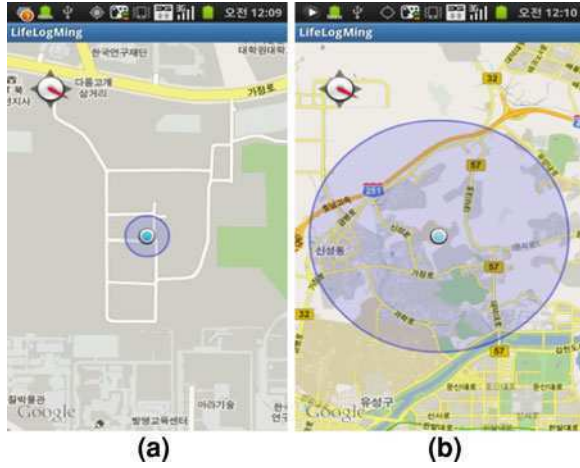


Fig. 5 Location information result

Smartphone GPS signal is strong outdoors but weak indoors. In this reason, it is not useful indoors. There are some methods to solve this problem, but we do not consider about this problem. Because, we gather location data only outdoors in this paper (Fig. 4).

When user moves, GPS data is changed. Using this data, user can see the distance and speed. We used formula to calculate the user calories. This formula has three parameters, which are fitness factor, weight, and total time. Fitness factor is changed by user speed (Fig. 5).

5 Conclusion

We gathered lifestyle data in experiment with collecting lifelog application. As you can find result from this experiment, we showed that we can collect someone's lifelog data by using Smartphone. This data can be useful for Medical History Form.

We proposed this Lifelog collection system for Medical History Form. We just focused on lifestyle data in this experiment. Therefore we need to study about collecting physical data by network connecting between medical device and Smartphone. After we collected two types of lifelog data, we will study about data mining algorithm for Medical History Form.

References

1. Nokia lifeblog, <http://www.nokia.com/lifeblog>
2. Gemmell J, Bell G, Lueder R, Drucker S, Wong C (2002) MyLifeBits: fulfilling the memex vision. In: ACM international conference on multimedia, Juan les Pins, pp 235–238
3. Aizawa K (2005) Digital personal experiences: capture and retrieval of life log. In: 11th international multimedia modeling conference, Melbourne, pp 10–15
4. Takata K, Ma J, Apduhan BO, Huang R, Jin Q (2008) Modeling and analyzing individual's daily activities using lifelog. In: ICES'08, Sichuanpp, pp 503–510
5. Doherty AR, Smeaton AF (2008) Automatically segmenting lifelog data into events. In: 9th international workshop on image analysis for multimedia interactive services, pp 20–23
6. Mizuno H, Sasaki K, Hosaka H (2007) Indoor–outdoor positioning and lifelog experiment with mobile phones. In: Proceedings of multimodal interfaces in semantic interaction (WMISI'07), pp 55–57
7. Minamikawa A, Kotsuka N, Honjo M, Morikawa D, Nishiyama S, Ohashi M (2007) RFID supplement mobile-based life log system. In: Applications and the internet workshops, p 50
8. Abe M, Morinishi Y, Maeda A, Aoki M, Inagaki H (2009) Cool: a life log collector integrated with a remote-controller for enabling user centric services. In: Proceedings of international conference on consumer electronics (ICCE'09), Las Vegas, pp 1–2
9. Hwang KS, Cho SB (2008) Life log management based on machine learning technique. In: Proceedings of IEEE international conference on multisensor fusion and integration for intelligent system, Seoul, pp 691–696
10. Strommer E, Kaartinen J, Parkka J, Ylisaukko-oja A, Korhonen I (2006) Application of near field communication for health monitoring in daily life. In: 28th international conference of the IEEE engineering in medicine and biology society, New York, pp 3246–3249

Simplified Swarm Optimization for Life Log Data Mining

Changseok Bae, Wei-Chang Yeh and Yuk Ying Chung

Abstract This paper proposes a new evolutionary algorithm for life log data mining. The proposed algorithm is based on the particle swarm optimization. The proposed algorithm focuses on three goals such as size reduction of data set, fast convergence, and higher classification accuracy. After executing feature selection method, we employ a method to reduce the size of data set. In order to reduce the processing time, we introduce a simple rule to determine the next movements of the particles. We have applied the proposed algorithm to the UCI data set. The experimental results ascertain that the proposed algorithm show better performance compared to the conventional classification algorithms such as PART, KNN, Classification Tree and Naïve Bayes.

Keywords Life log · Particle Swarm Optimization · Simplified Swarm Optimization

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (NRF-M1AXA003-2010-0029793).

C. Bae (✉)
Electronics and Telecommunications Research Institute,
218 Gajeong Ro, Yuseong Gu, Daejeon, Korea
e-mail: csbae@etri.re.kr

W.-C. Yeh
Advanced Analytics Institute, University of Technology Sydney,
PO Box 123 Broadway NSW 2007, Australia

Y. Y. Chung
School of Information Technologies, University of Sydney,
Sydney NSW 2006, Australia

1 Introduction

Due to the rapid development of wearable computing environments, we believe that it should make possible for continuous recording of various personal events using a wearable video camera, plus other miniature sensors. This type of wearable computer can be our secretary-agent. Therefore, the research on capture and retrieval of personal events in multimedia is emerging. Microsoft has done a bit of research on Digital Memory, but paid focus on the capture and storage of video media only and its SenseCam was just interested in images.

In order to continuously record various personal events in our life, the amount of the captured data will be very large and it is not easy to retrieve a particular event stored from the life-log server. For example, it may take another year to just watch the entire video captured in the life-log server for a one-year period. Therefore, we need an intelligent agent to understand and edit the captured context information automatically. In this paper, we propose a new data mining techniques that can study and learn the various characteristics of an important personal event of its user in order to predict and estimate the user's interests.

Feature selection and classification rule mining are two important problems in the emerging field of data mining which is aimed at finding a small set of rules from the training data set with predetermined targets. Feature selection is the process of choosing a subset of features from the original set of features forming patterns in a given dataset. The subset should be necessary and sufficient to describe target concepts, retaining a suitably high accuracy in representing the original features. The importance of feature selection is to reduce the problem size and resulting search space for learning algorithms [1]. The classification rule mining is aimed at finding a small set of rules from the training data set with predetermined targets [2].

Data mining is the most commonly used name to solve problems by analyzing data already present in databases. Many approaches, methods and goals have been tried out for data mining. Biology inspired algorithms such as Genetic Algorithms (GA) and swarm-based approaches like Ant Colonies [3] have been successfully used. Furthermore, a new technique which named Particle Swarm Optimization (PSO) has been proved to be competitive with GA in several tasks, mainly in optimization areas. However, there are some shortcomings in PSO such as premature convergence. To overcome these, we propose the modified Particle Swarm Optimization which named Simplified Swarm Optimization (SSO).

2 Related Researches

Support Vector Machines (SVMs) have been promising methods for data classification and regression [4], because it offers one of the most robust and accurate methods among all well-known algorithms. However, SVMs still have some

drawbacks: it can be abysmally slow in test phase; it has the high algorithmic complexity and extensive memory requirements of the required quadratic programming in large-scale tasks.

The Naïve Bayes method is a method of classification applicable to categorical data, based on Bayes theorem. Careful analysis of the Bayesian classification problem has shown that there are some theoretical reasons for the apparently unreasonable efficacy of Naïve Bayes classifiers [5]. An advantage of the Naïve Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

The Particle Swarm Optimization (PSO) comprises a set of search techniques, first introduced by Eberhart and Kennedy [6]. It belongs to the category of Swarm Intelligence methods; it is also an evolutionary computation method inspired by the behavior of natural swarms such as bird flocking and fish schooling. The details have been given in the following.

3 Proposed Algorithm: Simplified Swarm Optimization

The underlying principle of the traditional PSO is that the next position of each particle is a compromise of its current position, the best position in its history so far, and the best position among all existing particles. PSO is a very easy and efficient way to decide next positions for the problems with continuous variables, but not trivial and well-defined for the problems with discrete variables and sequencing problems. To overcome the drawback of PSO for discrete variables, a novel method to implement the PSO procedure has been proposed based on the following equation after C_w , C_p , and C_g are given:

$$x_{id}^t = \begin{cases} x_{id}^{t-1} & \text{if } rand() \in [0, C_w) \\ p_{id}^{t-1} & \text{if } rand() \in [C_w, C_p) \\ g_{id}^{t-1} & \text{if } rand() \in [C_p, C_g) \\ x & \text{if } rand() \in [C_g, 1). \end{cases} \quad (1)$$

In the traditional PSO, each particle needs to use more than two equations, generate two random numbers, four multiplications, and three summations in order to move to its next position. Thus, the time complexity is very high for the traditional PSO. However, the proposed SSO does not need to use the velocity, it only uses one random, two multiplications, and one comparison after C_w , C_p , and C_g are given. Therefore, the proposed SSO is more efficient than the other PSOs. Figure 1 is the flow chart diagram to explain the proposed process of individual update.

A classification rule contains two parts: the antecedent and the consequent. The former is a series of logical tests, and the latter gives the class while an instance is

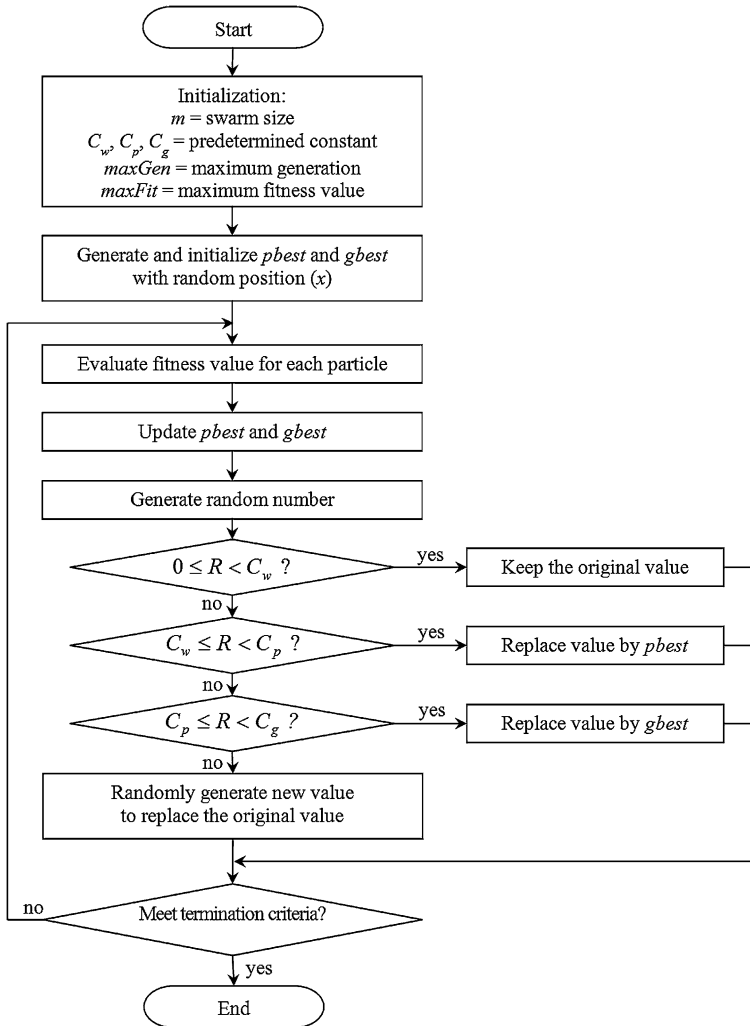


Fig. 1 The flowchart of the Simplified Swarm Optimization (SSO)

covered by this rule. These rules take the format as in Fig. 2. Where lower bound and upper bound are the attribute’s lowest and highest value, respectively. Each clause (dimension) is composed of an attribute, its lower bound and upper bound. The position representation of each individual (particle) contains N clauses (dimensions) except the last cell—Class X , which is the predictive class of the rule. To evaluate the quality of solutions, the fitness function has been taken into account. Its representation is defined as follows:

Attribute 1	<i>Lower Bound</i>	<i>Upper Bound</i>	...	Attribute <i>N</i>	<i>Lower Bound</i>	<i>Upper Bound</i>	Class <i>X</i>
----------------	------------------------	------------------------	-----	-----------------------	------------------------	------------------------	----------------

Fig. 2 Rule mining encoding

$$\text{The rule quality} = \text{sensitivity} \times \text{specificity} = \frac{TP}{TP + FN} \times \frac{TN}{TN + FP}. \quad (2)$$

where TP , FN , FP and TN are, respectively, the number of true positives, false negatives, false positives, and true negatives associated with the rule:

- (1) True Positives (TP) are the number of cases covered by the rule that have the class predicted by the rule;
 - (2) False Positives (FP) are the number of cases covered by the rule that have a class different from the class predicted by the rule;
 - (3) False Negatives (FN) are the number of cases that are not covered by the rule but that have the class predicted by the rule;
 - (4) True Negatives (TN) are the number of cases that are not covered by the rule and that do not have the class predicted by the rule.
- The goal of classification rule mining is to discover a set of rules with high quality (accuracy). To achieve this, appropriate lower bound and upper bound for each attribute (feature) are searched for. In initial stage, for each attribute we set its position of upper bound between a randomly chosen seed example's value and that value added to the range of that feature. Similarly, the value of lower bound is initialized at a position between the seed example's value and that value minus the range of that feature. The procedure is defined as:

$$\text{Lower bound} = k_1 * (S - R), \quad (3)$$

$$\text{Upper bound} = k_2 * (S + R), \quad (4)$$

where S is the corresponding attribute value of a randomly chosen instance; R is the range of corresponding attribute value for all training instances; k_1 and k_2 are two random numbers between 0 and 1.

After a rule has been generated, it is put into a rule pruning procedure. The main goal of rule pruning is to remove irrelevant clauses that might have been unnecessary included in the rule. Moreover, rule pruning can increase the predictive power of the rule, helping to improve the simplicity of the rule. The process of rule pruning is as follows:

- (1) Evaluate a rule's quality.
- (2) Tentatively removing terms from each rule and see if each term can be removed without the loss of rule quality. If yes, remove it. Then moves onto

Table 1 Comparison results of data mining experiment

Datasets	SSO	Part	KNN	Classification tree	SVM	Naïve bayes
Breast cancer	98.67	93.70	96.63	95.46	96.06	97.07
Glass	73.46	65.43	69.65	69.20	69.16	70.22
Diabetes	75.82	74.36	71.48	70.71	64.98	75.77
Iris	96.00	90.67	96.00	95.33	94.67	92.00
Zoo	98.09	94.18	96.09	91.09	91.09	91.18

the next term and eventually the next rule. This process is repeated until no term can be removed.

After we generate a rule set, a series of testing instances are used to measure its classification accuracy. For each instance, it will go through every element in rule set and get a prediction value for the corresponding class when it is covered by a rule. The prediction function is defined as follows:

$$\text{Prediction value} = \alpha * \text{rule quality} + \beta * \text{rule cover percentage} \quad (5)$$

where α and β are two parameters corresponding to the importance of rule quality and rule cover percentage, $\alpha \in [0, 1]$ and $\beta = 1 - \alpha$. The prediction value for each class is cumulative and the result is the class with highest prediction value.

4 Experimental Results

To thoroughly investigate the performance of the proposed SSO algorithm, we have conducted experiment with it on five widely used datasets taken from the UCI repository, such as Breast Cancer, Glass, Diabetes, Iris, and Zoo.

The experiment was carried out to compare predictive accuracy of discovered rule lists by well-known ten-fold cross-validation procedure [7]. Each data set is divided into ten partitions, each method is run ten times, using a different partition as test set and the other nine partitions as the training set each time. The two parameters α and β in Eq. 5 are important to the final validation accuracy. Slight adjustment could change the results significantly. In our experiment, we set $\alpha = \beta = 0.5$.

After the cross-validation of five data sets, we get the average validation accuracy of each data set. We compare these results with other five algorithms in Table 1. PART is WEKA's improved implementation of C4.5 rules. PART obtains rules from partial decision trees. It builds the tree using C4.5's heuristics with the same user-defined parameters as J48 [8]. KNN, Classification Tree, SVM and Naïve Bayes are four classic algorithms implemented by ORANGE. We compared SSO against these algorithms as they are considered industry standards. The results of the six algorithms are shown in Table 1. The comparison clearly states that the

competitiveness of SSO with other algorithms. It can be seen that predictive accuracies of SSO is higher than those of other five algorithms.

5 Conclusions and Future Works

This paper discusses the shortcomings of conventional PSO and proposes a novel algorithm namely Simplified Swarm Optimization (SSO) in order to overcome PSO's drawbacks. In SSO, a random number and three parameters (C_w , C_p , C_g) are required to discretely update a particle's position. We also combined SSO with K-Means clustering algorithm to deal with continuous variables.

We applied SSO to some areas such as feature selection and classification rule mining. Comparing with traditional PSO, SSO has stronger search capability in the problem space and can efficiently find minimal reductions of features. Experimental results states competitive performance of SSO. Due to less computing for swarm generation, averagely SSO is over 30% faster than PSO. Furthermore, we applied SSO to classification rule mining and achieved satisfactory results. The reason to select features is that feature selection can effectively improve the classification accuracy. The proposed algorithm has compared with other algorithms such as PART and KNN. The results show that the generated rule set from SSO has higher accuracy.

As our proposed SSO has proved to be superior to other traditional data mining algorithms, the proposed algorithm can be applied to the life-log media fusion applications

References

1. Wang XY, Yang J, Teng X, Xia W, Jensen R (2006) Feature selection based on rough sets and particle swarm optimization. *Pattern Recog Lett* 28: 438–446
2. Quinlan JR (1986) Induction of decision trees. *Mach Learn* 1:81–106
3. Vapnik V (1995) *The nature of statistical learning theory*. Springer, New York
4. Joachims T (1998) Text categorization with support vector machines. In: *Proceedings of 10th European conference on machine learning*, Chemnitz, Germany
5. Zhang H (2004) *The optimality of naive bayes*. AAAI Press,
6. Kennedy J, Eberhard RC (1995) Particle swarm optimization. In: *Proceedings of IEEE international conference on neural networks*, Piscataway, NJ, USA, pp 1942–1948
7. Weiss SM, Kulkowski CA (1991) *Computer systems that learn*. Morgan Kaufmann, San Mateo
8. Witten IH, Frank E, *Mining Data* (2005) *Data mining: practical machine learning tools and techniques*, 2nd edn. Morgan Kaufmann, San Francisco

The Design and Implementation of Web Application Management on Personal Device

Eunjeong Choi, Hyewon Song, Changseok Bae
and Jeunwoo Lee

Abstract This paper describes the implementation and design of managing web-based applications which is converted from a web application to a native application. According to existing technologies, a user should enter a web site address into a web browser to use a web application. Otherwise, developers implement a hybrid web application for a web application and upload it to application stores or markets, and a user should visit application stores or markets for smart phones, search a web application, and then install it. These methods are not easy to use or maintain the web application. In this paper, a user who uses the web application by typing the address of a Web-based application with handsets can convert it into a native application, install it on the terminal, and manage it as a native application more conveniently. In addition, Web-based applications installed on a device can be synchronized to the user's other devices via a personal cloud server. As a result, by easily converting a web application into a native application on the existing a web browser on a device, a user can easily use and manage web applications.

Keywords Web application · Smart phone · Personal computing

E. Choi (✉) · H. Song · C. Bae · J. Lee
Electronics and Telecommunications Research Institute (ETRI),
138 Gajeongno, Yuseong-gu, Daejeon 305-700, Korea
e-mail: ejchoi@etri.re.kr

H. Song
e-mail: hwonsong@etri.re.kr

C. Bae
e-mail: csbae@etri.re.kr

J. Lee
e-mail: ljwoo@etri.re.kr

1 Introduction

Native applications are developed and used on smart phones such as Android, iPhone and so on. Also, the uses of web applications increase on the smart phones. Web applications are useful to use because they are light weight and platform independent. However, web applications are not easy to use because users should know the correct addresses and type the address on a web browser on mobile devices using small on-screen keyboard. That's inconvenient for users. So, it will be convenient if web applications can be used as native applications by clicking icons on a home screen.

This paper describes the implementation and design of managing web-based application which is converted from a web application to a native application. In this paper, a Web application to convert a native application, terminal Application Manager to install and manage it, suggested techniques.

2 Related Work

Mozilla Prism is designed to create a better environment for running favorite web-based applications of users. Much of what we used to accomplish using an application running locally on our computers is moving into the web browser. Thanks to advances in web technology, these apps are increasingly powerful and usable. As a result, applications like Gmail, Facebook and Google Docs are soaring in popularity (Fig. 1) [1, 2].

A common way to use a web application easy is to make bookmark for the web pages. The bookmark of a web browser saves the web site addresses and title information to a desktop. These methods are just to save the information of a web site such as a title and a URL. Users can use the web applications by clicking icons. However, in this paper, the method is enhanced not only by saving the information of web application but also by making new viewers of the web applications.

3 Web Application Management

This paper describes the design of web application managements including the architecture and process of web application managements.

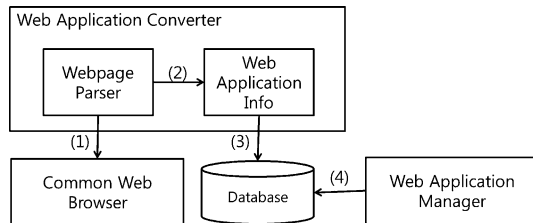
3.1 Architecture

Figure 2 shows an architecture for web application managements. There are several modules: (1) Web Application Converter including Webpage Parser block



Fig. 1 Mozilla prism concept [2]

Fig. 2 An architecture for web application managements

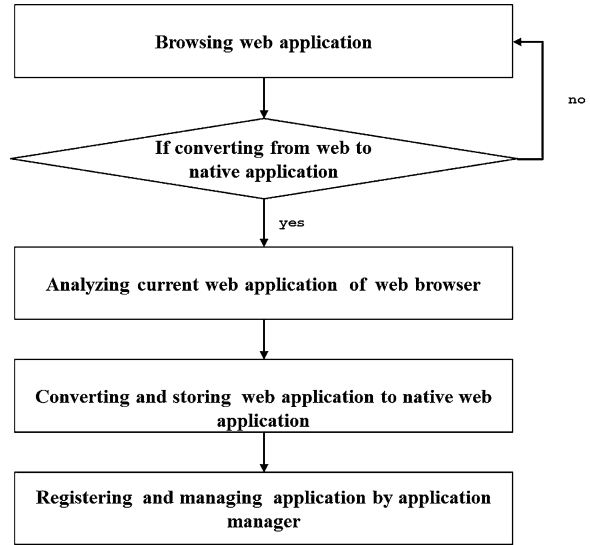


and Web Application Info block, (2) Common Web Browser, (3) Database, and (4) Web Application Manager. The Common Web Browser, Database, and Web Application Manager modules are external software interfaces. In this paper, Web Application Converter including Webpage Parser and Web Application Info blocks is developed. An implementation issue will be described in the next paper.

Web Application Converter. Web Application Converter is a module for converting a web application to a native web application. Web Application Converter includes Webpage Parser and Web Application Info blocks. The Webpage Parser block parses the downloaded web sites browsing on a common web browser, extracts information from the web application, and then saves them to a database.

Common Web Browser. A user can easily use a web application via a common web browser. While a user uses a web application by using this Common Web

Fig. 3 Procedure for web application managements



Browser, the user can convert it to a hybrid application which means an application with the web application information and a native viewer application for it.

Database. After parsing web sites of the common web browser, the information of the web application will be stored in a database by the Web Application Converter.

Web Application Manager. After converting a web application to a native application, the Web Application Manager will manage the application as a native application and execute it on a device. Also, the application manager synchronizes the applications to the personal cloud and the user's other devices.

3.2 Procedure

Figure 3 shows a procedure for web application managements.

The procedure for converting a web application to a native web application is as follows. The first step is that a user browses a web application via a common web browser. While a user uses a web application by using this Common Web Browser, the user can convert it to a native application. The second step is that a user selects if the current web application will be converted into a native application. If a user selects converting it, the Web Application Converter analyzes the current web application of the common web browser. The next step is that the Web Application Converter converts the web application into a native application, and then stores the information to the database. Finally, the last step is that the Web Application Manager registers the web application to a personal cloud server and manages it like other native applications.

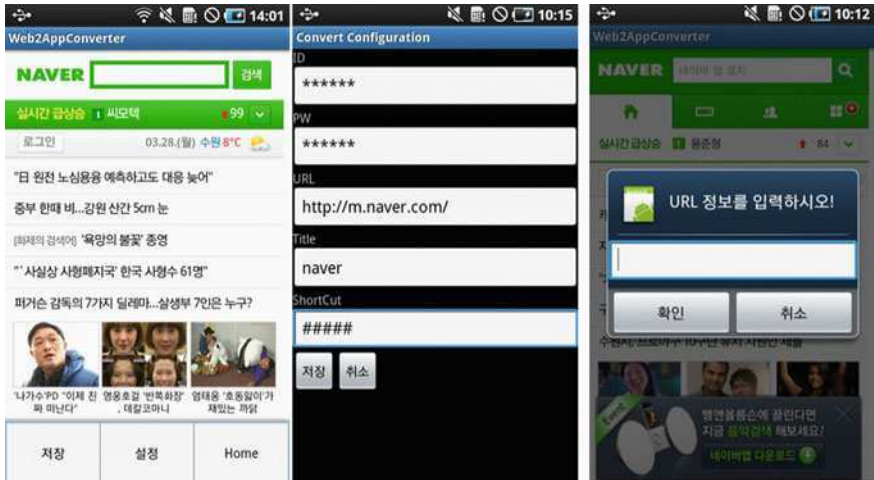


Fig. 4 An example of web application managements based on an android device

4 Implementation and Result

We implemented an example of web application managements on Android platform. The following example of Android-based web application managements shows: (1) Browsing the web application screen; (2) Web application information setup screen; (3) Web Application Search Screen (Fig. 4).

5 Conclusion

In this paper, a user who uses the web application by typing the address of a Web-based application with handsets can convert it into a native application, install it on the terminal, and manage it as a native application more conveniently. In addition, Web-based applications installed on a device can be synchronized to the user's other devices via a personal cloud server. As a result, by converting a web application into a native application on the existing a web browser on a device, a user can easily use and manage web applications.

Acknowledgment This work was supported by the IT R&D program of MKE/KEIT. [K10035321, Terminal Independent Personal Cloud System].

References

1. Leichtenstern T (2008) Launching web applications in Prism. *BORDERLESS Linux Mag* 90:52–53
2. Mozilla Prism, prism.mozillalabs.com/
3. Grønli T-M, Hansen J, Ghinea G (2011) Integrated context-aware and cloud-based adaptive home screens for android phones. *HCI 2011, LNCS*, vol 6762, pp 427–435
4. Google Mobile: Android basics: Getting to know the Home screen (2010), <http://www.google.com/support/mobile/bin/answer.py?answer=168445#1149468>. Last visited 5 Oct 2010
5. Göker A, Watt S, Myrhaug HI, Whitehead N, Yakici M, Bierig R, Nuti SK, Cumming H (2004) An ambient, personalised, and context-sensitive information system for mobile users. In: *Proceedings of the 2nd European union symposium on ambient intelligence*. ACM, Eindhoven, pp 19–24

Ad Hoc Synchronization Among Devices for Sharing Contents

Eunjeong Choi, Changseok Bae and Jeunwoo Lee

Abstract This paper describes ad hoc data synchronization among devices for sharing contents. The purpose of this paper is to share user data in heterogeneous environments, without depending on central server. This technology can be applied to synchronize personal data between a device and a personal cloud storage for personal cloud services. The ad hoc synchronization needs sync agent service discovery module, user authentication module, network adapter, and application data synchronization module. The method described in this paper is better than existing synchronization technology based on client-server in availability, performance, and scalability quality attributes.

Keywords Data synchronization · Ad hoc synchronization · Personal cloud service

1 Introduction

Nowadays many people get used to smart phones such as iPhone or GalaxyS based on Android and the types of devices are various including IPTV and PC. And then, people want to use same data on a same platform even though their devices are

E. Choi (✉) · C. Bae · J. Lee
Electronics and Telecommunications Research Institute (ETRI),
138 Gajeongno, Yuseong-gu, Daejeon 305-700, Korea
e-mail: ejchoi@etri.re.kr

C. Bae
e-mail: csbae@etri.re.kr

J. Lee
e-mail: ljwoo@etri.re.kr

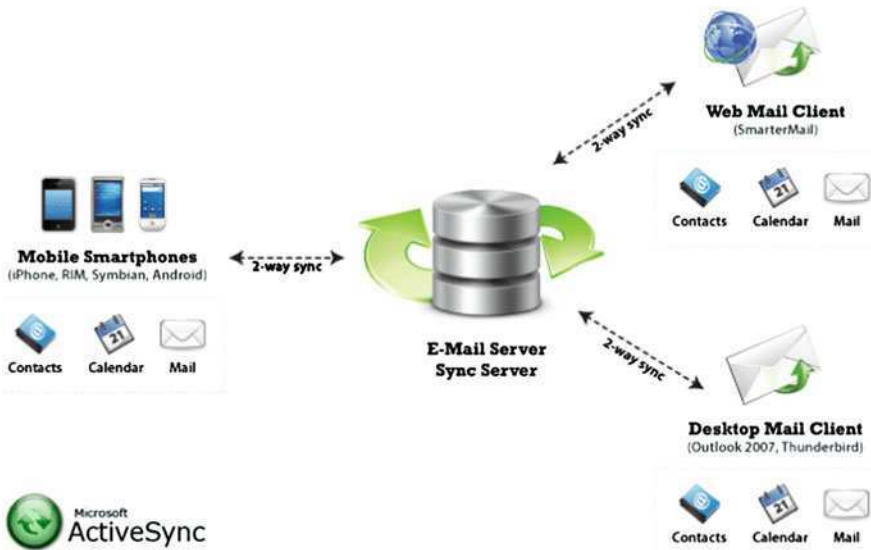


Fig. 1 Microsoft ActiveSync [3]

different. This means that the technology should be enhanced compared to existing technology which exchanges data between a mobile phone and PC using USB, etc. Now the data of a user can be uploaded or downloaded among various devices via server.

In this paper, ad hoc data synchronization among devices for sharing contents is described. The purpose of this paper is to share user data in heterogeneous environments, without depending on central server. This technology can be applied to synchronize personal data between a device and a personal cloud storage for personal cloud services.

In Sect. 2, existing data synchronization based on client–server and some of issues for it will be discussed. In Sect. 3, the proposed ad hoc data synchronization technique for data-sharing will be described. Finally, this paper will conclude in Sect. 3.

2 Related Work

Current representative application data synchronization services are Funambol [2] based on SyncML standard [1], ActiveSync of Microsoft [3], and MobileMe of Apple [4]. These application data synchronization services are based on data synchronization between a mobile phone and a central server. This means the mobile phone should be connected to a certain data server to synchronize the application data. Figure 1 shows the network topology of Microsoft

synchronization framework. As shown in the figure, the user data such as contacts or e-mail is uploaded or downloaded via a certain server on the Internet.

There are several problems in the existing application data synchronization technologies based on client–server data exchange.

Firstly, the data synchronization based on client–server causes data traffics on the network. The steps are as follows: (1) application data on a mobile phone should be stored on the storage of certain data server, (2) a user should retrieve the application data from the data server to use same data on a different devices. If the user is using iPhone and IPTV in a room, two devices need to upload and download application data via synchronization server on the Internet which is far from the devices even though the devices are adjacent with each other.

Secondly, the existing technology needs central server. This means users cannot synchronize application data if the network is offline with the central server.

Thirdly, the application data may be lost if the application data on a certain mobile phone was not uploaded to a server and then a user lost the device or the device is broken. As the first problem, the user cannot use the application data on a device which is generated from another device.

3 Ad Hoc Data Synchronization

The purpose of this paper is to share application data among devices without a central server. In this article, the architecture and procedure of ad hoc data synchronization are described in detail.

3.1 Architecture

Figure 2 shows a layered structure for ad hoc data synchronization on a device. There are four layers on a device: (1) Ad hoc Data Sync Agent layer, (2) User Authentication layer, 3) service discovery protocol layer, and 4) network layer such as Bluetooth, Zigbee, WiFi, and so on.

Ad hoc Data Synchronization Agent Layer. Ad hoc Data Synchronization Agent layer connects with other Ad hoc Data Synchronization Agent on a different device and then exchange application data each other.

User Authentication Layer. User Authentication Layer authenticates a user to try to connect with the device from another. The authentication is done by the user identification and password of each device.

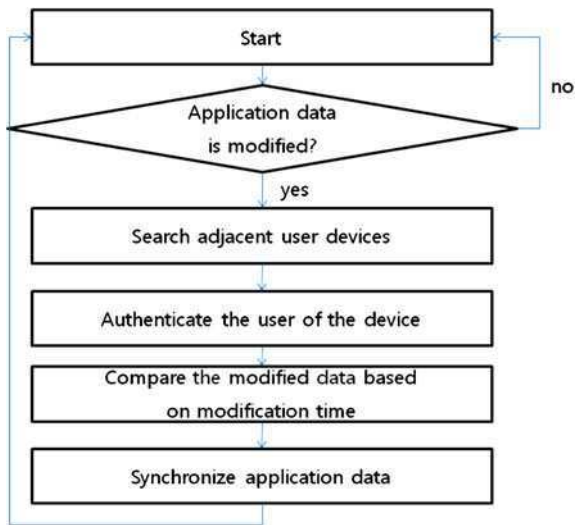
SDP Layer. This layer is service discovery protocol layer. Ad hoc Data Synchronization Agent tries to synchronize with adjacent devices. Service Discovery Protocol Layer finds network services to connect with the device.

Network Layer. There are several network services such as Bluetooth, Zigbee, WiFi, and so on.

Fig. 2 The architecture of ad hoc data synchronization



Fig. 3 The procedure of ad hoc data synchronization



3.2 Procedure

Figure 3 shows the procedure for ad hoc data synchronization. The steps to synchronize are as follows.

First, an application modifies its data and then the modification event is occurred.

Second, Ad hoc Data Synchronization layer captures the event and searches a user’s adjacent devices to connect with the device which the modification happens.

Third, if there is an Ad hoc data Synchronization Agent on the other devices, the User Authentication Layer tries to authenticate the user with user identification and a password.

Forth, the Ad hoc Data Synchronization Agent compares the data between two devices which is modified at last.

Fifth, the Ad hoc Data Synchronization Agent exchanges the data between two devices.

4 Conclusion

The technology described in this paper can be applied to synchronize personal data between a device and personal cloud storage for personal cloud services.

This paper describes the architecture and procedure for ad hoc data synchronization. There are four layers on a device: (1) Ad hoc Data Synchronization Agent layer, (2) User Authentication layer, (3) Service Discovery Protocol layer, and (4) Network layer such as Bluetooth, Zigbee, WiFi, and so on. Each layer is related to searching Ad hoc Data Synchronization Agent, authenticating a user, exchanging application data, and so on.

The method described in this paper is better than existing synchronization technology based on client-server in availability, performance, and scalability quality attributes. Devices can synchronize at any time with adjacent user devices even if the data server is broken. This method does not cause useless data traffic on the Internet.

Acknowledgment This work was supported by the IT R&D program of MKE/KEIT. [K10035321, Terminal Independent Personal Cloud System]

References

1. SyncML Initiative, <http://www.syncml.org>
2. Funambol, <http://www.funambol.com>
3. Microsoft ActiveSync, <http://www.microsoft.com/windowsmobile/en-us/help/synchronize/device-synch.mspx>
4. Apple MobileMe, <http://www.apple.com/mobileme/>
5. Sreeram J, Pande S (2010) Exploiting approximate value locality for data synchronization on multi-core processors. In: *ieee international symposium on workload characterization, IISWC'10*, Article No. 5650333
6. Su Z, Hou X (2010) Application of data synchronization based on ESB. In: *2nd IITA international conference on geoscience and remote sensing, IITA-GRS 2010*, vol 1, Article No. 5603013, pp 295–297

A Framework for Personalization of Computing Environment Among System on-Demand (SoD) Zones

Dong-oh Kang, Hyungjik Lee and Jeunwoo Lee

Abstract In this paper, we propose a framework for personalization of computing environment when a user moves from one System on-Demand (SoD) zone to another, which preserves the user's computing environment. In our approach, an image containing an operating system, user profiles and user's data are dealt with as components of personalization of virtual machines. We deal with how to assemble a virtual machine with personalization components efficiently. To show the feasibility of the proposed method, we apply the proposed approach to the personalization of virtual machines of a test bed of SoD service.

Keywords Personalization · Virtual machine · Computing environment · System on-Demand

1 Introduction

Recently in terms of seamless personal computing environment, some companies of cloud computing service provided desktop virtualization solutions, which allocate a virtual desktop to a user using system virtualization technologies and

D. Kang (✉) · H. Lee · J. Lee
Software Research Lab, Electronics and Telecommunications
Research Institute, Daejeon, Korea
e-mail: dongoh@etri.re.kr

H. Lee
e-mail: leehj@etri.re.kr

J. Lee
e-mail: ljwoo@etri.re.kr

remote desktop protocols [1–3]. Therefore, we have many chances to use virtual machines in daily life because of the development of these cloud computing technologies. That is, the virtualization technologies make system on-demand possible, which provides a user a computing system as the user demands [4]. In previous approaches to virtual machines, users usually made his or her virtual machine configurations which had fixed sets of configurations or selected a virtual machine configuration among configurations of virtual machines stored in a server. Therefore, previous approaches are difficult to give users the personalized virtual machine in level of user's profile.

In this paper, we propose a framework for personalization of computing environment for System on-Demand (SoD) service, which preserves the user's profiles and computing environment among multiple SoD zones. In our approach, an image containing an operating system, user profiles and user's data are dealt with as components of personalization of virtual machines. When a user enters into a SoD zone for the first time, the user's personalized image of the virtual machine is assembled based on the image of an operating system with the user profile and user's data. After that moment, the user can use his or her virtual machine based on the personalized image in the SoD zone. Because the user profile is downloaded from a personalization information management server before the virtual machine is booted and applied to the virtual machine during the user's data are downloaded, the user can get fast personalization of the virtual machine and use his or her virtual machine during the personalization process. Therefore, the system for SoD service using the proposed personalization framework of virtual machines gives the users fast personalized virtual machines as he or she demands when the user moves from one SoD zone to another. To show the feasibility of the proposed method, we apply the proposed approach to the personalization of virtual machines of a test bed of SoD service.

2 System On-Demand Service

The System on-Demand (SoD) service is the service that provides virtual personal computers which are optimized virtual machines using distributed u-computing devices around users as users demand. The SoD service can solve the limit of the place, devices and time of the traditional PC based computing technology in terms of the personal computing environment. The SoD service is provided within a SoD zone. The system of SoD service is composed of a SoD station, SoD servers and SoD clients in a SoD zone. SoD storage servers and a personalization information management (PIM) server can be optionally included in the system. The components of the system of SoD service are connected via network (Fig. 1).

The SoD station is the management server of SoD service, which has the control and information of SoD servers and SoD clients. The SoD servers provide computing capabilities of virtual machines, i.e., CPU, memory, hard disks, and network interface, etc. The SoD servers provide the kernel capability of virtual

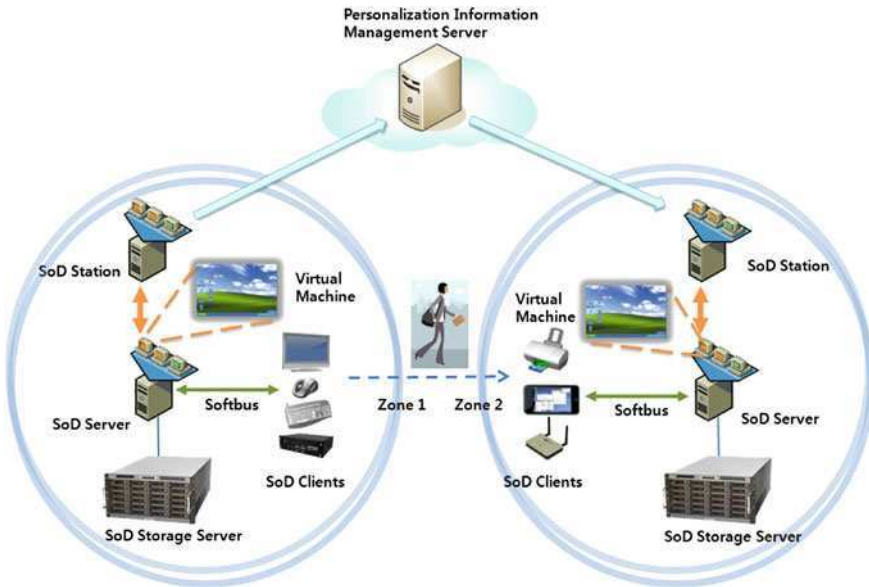


Fig. 1 Conceptual system configuration for SoD service

machines. The SoD clients are the devices which request the kernel capability of a virtual machine. Major SoD clients are the I/O devices like Human Interface Device (HIDs) and storage devices, etc. The SoD storage servers are the network storage servers to provide the images of virtual machines through network. The images include operating systems, application programs, and the user's data. The PIM server is the essential element of SoD system in terms of personalization of computing environment to store user profiles and manage personalization information. The PIM server can be outside a SoD zone and cooperates with SoD stations of multiple SoD zones.

3 Personalization Software

For the personalization of virtual machines, personalization software comprises a master computing environment manager, slave computing environment managers, domain 0 Virtual Machine (VM) personalization agents, domain U VM personalization agents, and a personalization information manager as shown in Fig. 2. Figure 3 shows the overall personalization mechanism of computing environment when a user migrates from one SoD zone to another. The master computing environment manager resides in a SoD server, and controls the behaviors of slave computing environment managers in SoD servers. The computing environment

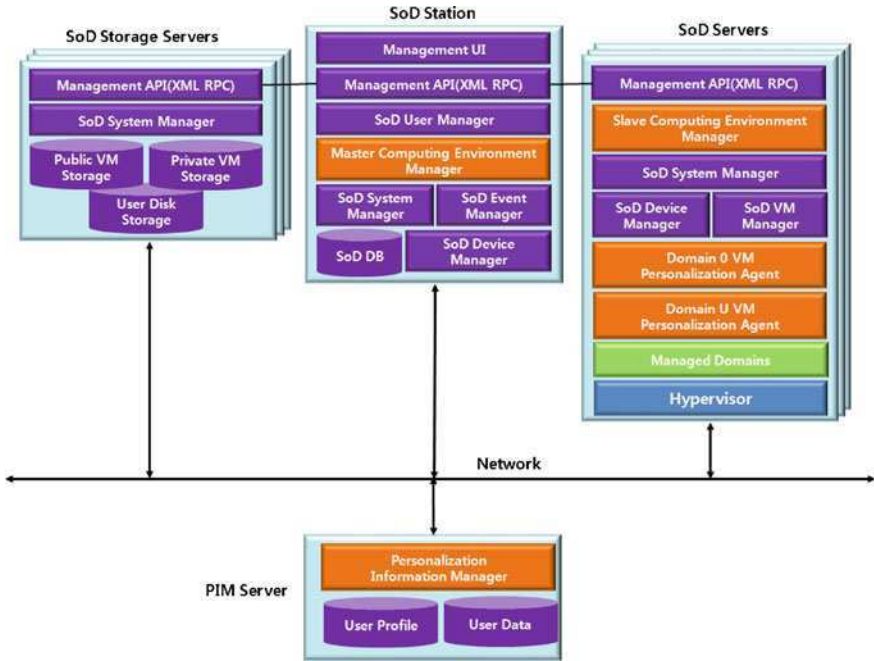


Fig. 2 Personalization software architecture

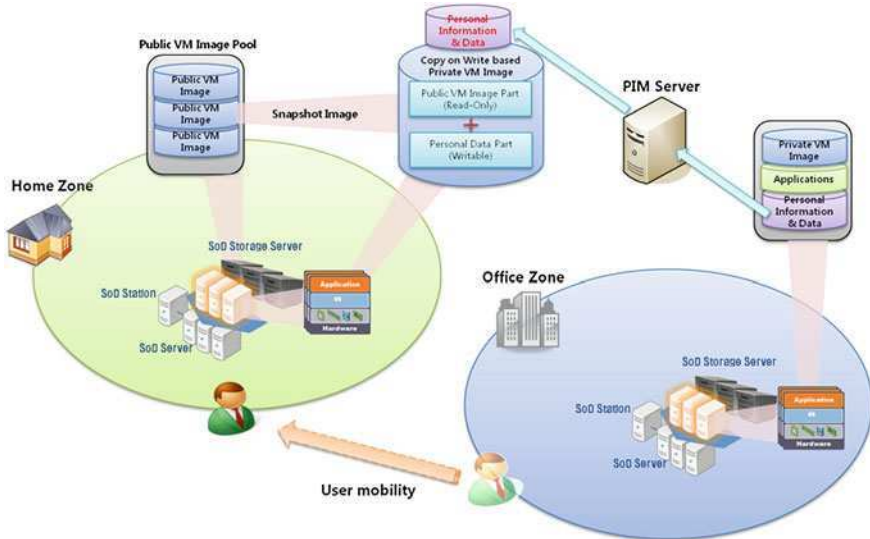


Fig. 3 Personalization mechanism of computing environment

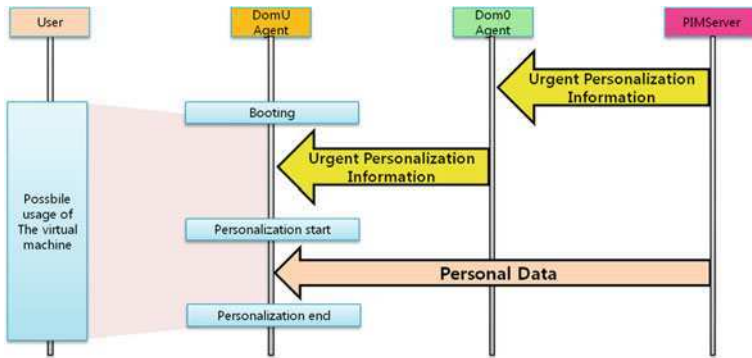


Fig. 4 The flow of personalization

managers deal with and provide information of users and SoD service system. The computing environment managers are connected via XML-RPC mechanism which is provided by the management Application Program Interface (API). A domain 0 VM personalization agent exists in a SoD server. It initializes personalization metadata when a user registers to SoD service for the first time. And, when the user enters into a SoD zone and wants to use his or her own virtual machine for the first time, it negotiates the personalization information manager and drags the user profile of the user into the SoD server before the virtual machine starts. When the virtual machine is booted, the domain U VM personalization agent in the virtual machine is activated in a virtual machine with a guest operating system and performs the personalization process like registry modification, file exchange, and user data transfer, etc. The personalization process may need some time because the user data of big size should be transferred. But, because the user profile is previously downloaded in the SoD server and transmitted to the personalization agent in the virtual machine, the user can use the virtual machine during the personalization process. When the user uses the virtual machine in a SoD zone, the personal image of the virtual machine is assembled and stored in a SoD storage server in the zone. After the first use of the virtual machine, the user can use the personalized image of the virtual machine stored in the zone. Therefore, the user can use the image without the further personalization process. Therefore, the user can experience fast personalization of virtual machines by the proposed technique. The process is depicted in Fig. 4.

4 Application to a Test Bed of SoD Service

We implement an all-in-one type SoD server system to include a SoD station, some SoD servers, SoD storage servers, a PIM server. And the SoD adaptor is used to convert legacy I/O devices to SoD clients in a SoD zone. A mobile SoD client is



Fig. 5 A test bed for SoD service

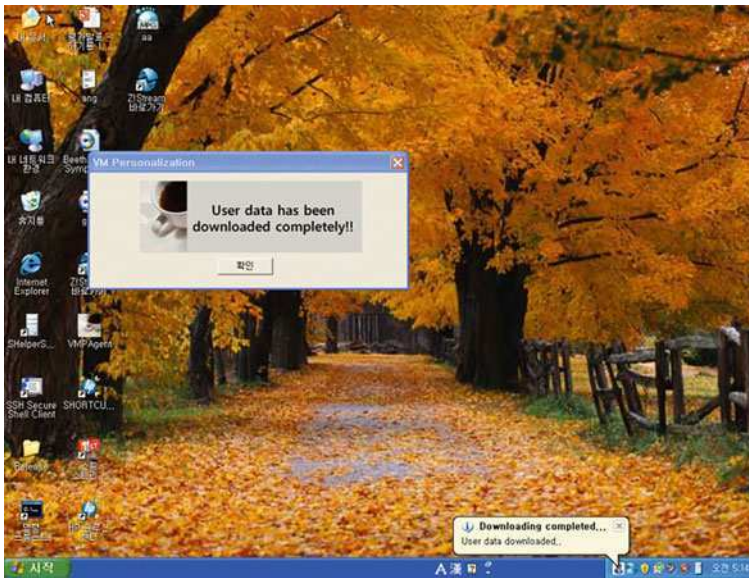


Fig. 6 The result of personalization

used like an iPad and an iPod in another SoD zone as depicted in Fig. 5. Figure 6 shows the result of personalization of computing environment when the personalization of a virtual machine is completed.

5 Concluding Remarks

Compared with the traditional approaches, the proposed approach can give faster personalization process of computing environment and a user can use his or her virtual machine during the personalization process. And, the proposed approach can be applied to SoD service for mobile workers or teleworkers and give more convenient usage to users. For the future research, we will study about how to relate the proposed method with I/O personalization of virtual machines.

Acknowledgments This work was supported by the IT R&D program of MKE/KEIT (2008-S-034-01, Development of Collaborative Virtual Machine Technology for SoD).

References

1. Wolf C, Halter EM (2005) *Virtualization from the desktop to the enterprise*. Apress, Berkeley
2. Matthews J et al (2008) *Running xen: a hands-on guide to the art of virtualization*. Prentice Hall, Saddle River
3. Ju Y et al (2008) VirtHome: a web-like mobile personalized virtual desktop computing space. In: *Proceeding of ISISE'08*, vol 2, pp 192–196
4. Kang D, Lee J (2009) Component based personalization technique of virtual machines for System on-Demand (SoD) service. In: *Proceeding of ICMIT 2009*, vol 2, pp 187–188

Part VII
Security and Application for Embedded
Smart Systems

Facsimile Authentication Based on MAC

Chavinee Chaisri, Narong Mettripun and Thumrongrat Amornraksa

Abstract In this paper, we propose a method to provide message authentication and integrity for a facsimile (fax) document using Message authentication code (MAC) based approach. The proposed method is divided into two parts; sender and receiver. Basically, at the sender side, a MAC value derived from the fax content and a predefined secret key is added to the document before sending it to the receiver via fax. At the receiver side, the modification of fax content can be detected by the use of the agreed MAC and secret key, and the MAC value added on the received fax. The experimental results, from the fax transmission over insecure communication channels, using different types of fax machine, font types and font sizes, demonstrate the promising results.

Keywords Message authentication · Data integrity · Facsimile · Message authentication code (MAC)

C. Chaisri · N. Mettripun (✉) · T. Amornraksa
Multimedia Communication Laboratory, Department of Computer Engineering,
King Mongkut's University of Technology Thunburi, 126 Pracha-uthit Rd,
Bangmod, Thungkru, Bangkok 10140, Thailand
e-mail: mettripun_n@hotmail.com

C. Chaisri
e-mail: ch.chavinee@gmail.com

T. Amornraksa
e-mail: t_amornraksa@cpe.kmut.ac.th

1 Introduction

Currently, facsimile (fax) machines are widely used in both analog and digital networks. However, the content within the received fax is suspicious since it can potentially be modified by malicious people. Hence, a detection method is greatly required to provide message authentication and integrity of a fax document. Actually, authentication in documents has been a topic of interest for many years, but for the subject document authentication via fax machine, it has not been much interest by most researchers. Some of them are listed here. Williams et al. [1] developed a method for spotting words in faxed document. This method allowed scale and translation invariant transformations to be used as one step of the signature recognition process. Their techniques provided a very robust means of identifying the words in a bitmapped fax documents. However, the authors did not consider the case of authentication for fax document. Musmann and Preuss [2] proposed comparison and valuation of different redundancy codes techniques for transmission via fax machine. In their experiments, the data transmission were carried out with one and two-dimensional. This method reduced the transmission errors and used less time in transmission, but the faxed document cannot be used for authentication purposes. Garain and Halder [3] proposed the methods of computationally extracting the security features from the document image as bank checks, and identifying the feature space if it was genuine or duplicate. Although his method provided document authentication, it cannot be used for document being sent via fax machine. Geisselhardt and Iqbal [4] proposed an authentication approach for hard copy document based on a preferably invisible encoded portion, and a method for generating such document in which the encoded portion allowed an optimized high capacity of data to be read with security or only few errors. Their method prevented the printed content on hard copy document against forgery attacks, and did not affect the aesthetic appearance of the document in the area of secret communication such as military communication. Unfortunately, it is not practical for real life communication because most of documents sending are performed via insecure communication channels which are more comfortable and faster. Hence, this method is not appropriate to implement with transmission via fax machine. In 2008, Kale et al. [5] proposed a system for compression and encryption of fax documents and error recovery over fax transmission. Basically, the size of document to be faxed was reduced by applying Joint Bi-level Image Experts Group (JBIG) compression technique. They also applied an encryption technique called Salsa20 to produce less effect on retransmission delays and less cost for fax communication. However, the encryption algorithm used i.e. Salsa20 has been proved to be insecure by cryptanalysis group in 2005 and 2008. In addition, their scheme did not consider any error or noise introduced during fax communication which resulted in some bit errors and some unclear parts of faxed document.

In this paper, we thus propose a method to provide message authentication and integrity for fax documents. Particularly, it is achieved by first applying a MAC

algorithm [6] to the fax content to generate a MAC value. Note that a MAC algorithm may be obtained by applying a hash algorithm i.e. MD5 [7] to the fax content to obtain a hash value, and then encrypt the result using a symmetric encryption algorithm i.e. DES to generate a MAC value [8]. Next, this MAC value is printed at the end of the fax content to produce the real fax document, and later used to verify the integrity and authentication of the faxed document. The verification process can be achieved simply by comparing the MAC values between the one printed on the faxed document and another from the computation process from the content on the faxed document at the receiver side. With our proposed method, the fax receiver can now verify the originality of the text-based content in the faxed document, and detect whether it is changed or not. We organize our paper as follows. In the next section, we describe details of our proposed method. In Sect. 3, sets of experiments are carried out and the results obtained are presented in order to verify the effectiveness of our proposed method. Finally we conclude the finding of our research in Sect. 4.

2 The Proposed Method

The model for fax sending and receiving between two different locations is described by the following steps.

1. Creating a fax document that can be used to detect message authentication and integrity of its content based on MAC algorithms.
2. Sending this fax document by an ordinary fax machine.
3. Receiving the fax document from another ordinary fax machine.
4. Based on the fax content, the MAC value is regenerated and compared with the one printed on the fax document itself. If they are matched, the fax content is approved. If not, the receiver asks the sender to send the fax document again.

Detail of the fax document creating process in step 1 can be explained as follows. First, a typical text-based fax document is scanned to obtain an image file. Then, a frame line with one-pixel width is inserted to enclose the fax content in that image. The frame detection and cropping algorithm is applied to acquire the image area within the frame, and the result is then put into optical character recognition (OCR) software. Information outputted from the OCR is hashed by the MD5 algorithm and encrypted by DES algorithm with a predefined secret key. The encrypted result known as MAC value is inserted below the fax content outside the frame as a subtitle. Finally, the modified fax image with frame and MAC value is printed out on a white color paper to create a ready-to-send fax document. Figure 1 illustrates the block diagram of fax document creating process performed at the sender side.

After receiving the above fax document at another end, it is verified by our proposed method to validate its content. Detail of the fax content verifying

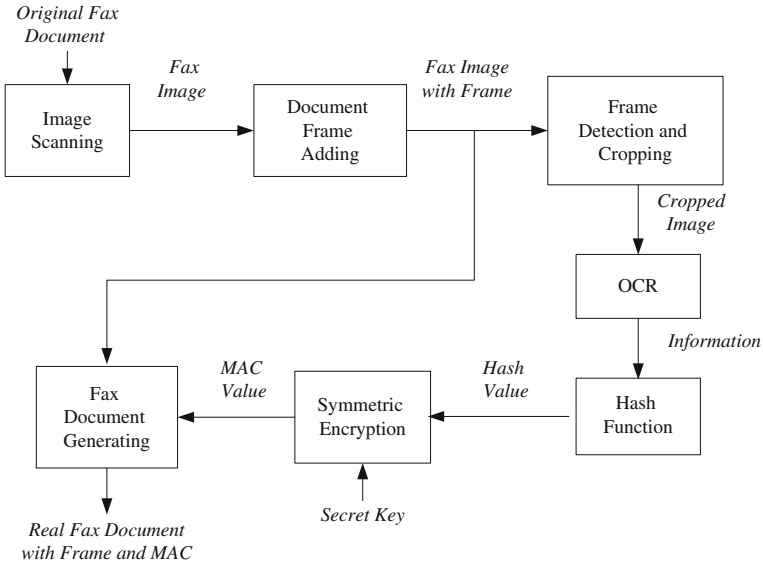


Fig. 1 Block diagram of the proposed fax document creating process

process mentioned in step 4 can be explained as follows. First, the received fax, called faxed document, is scanned back to obtain an image file. With the frame surrounding the fax document, we apply the rotation and scaling correction algorithm described in [9, 10] to fix any incorrect inclination and image resolution caused by the improper scanning setting/process. The result is then divided into two parts by the same frame detection and cropping algorithm, that is, the image area inside the frame representing the fax content and the image area outside the frame representing the MAC value. Both image areas are then input to the same OCR software as used at the sender side independently. The information obtained from the first image area inside the frame is hashed and encrypted with the MD5 and DES algorithms and the same secret key to obtain a MAC value, while the information obtained from the second image area outside the frame is used for verifying purpose. Finally, both MAC values from different processes are compared. If they are matched, the authentication and integrity of the fax content are verified. If not, someone may add/delete/alter the content of the fax. Figure 2 illustrates the block diagram of fax document verifying process performed at the receiver side.

In this research work, we consider any error possibly introduced to the faxed document during the fax transmission via communication channels e.g. telephone line. However, according to the results obtained, such errors were automatically removed by the OCR software because the output from the OCR process contained text-based information only.

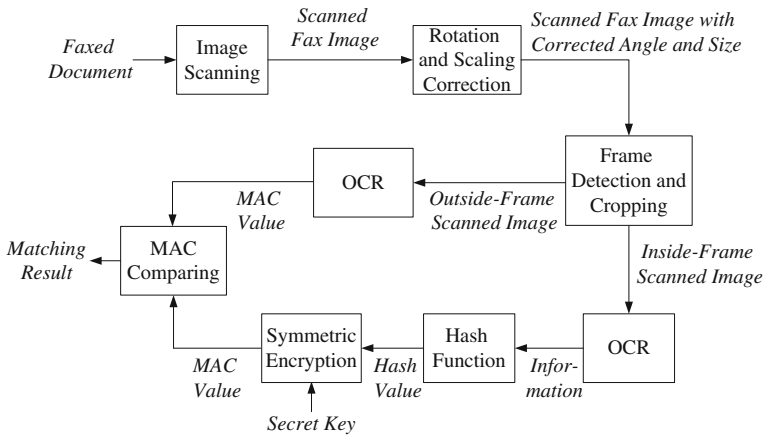


Fig. 2 Block diagram of the proposed fax document verifying process

3 Experimental Setting and Results

3.1 Experimental Setting

Since most types of fax document frequently used nowadays can be divided into two types, depending on the operation of fax machine i.e. ordinary A4 white color paper and thermal white color rolled paper, we thus considered to test both typed of them. In the experiments, we used A4 white color paper with the size of 210 × 297 mm. and thermal white color rolled paper with the size of 210 × 216 mm as the fax document. For the scanning process, the flatbed scanner ‘Lexmark X8350 All-in-One’ was used to scan the ‘Original fax Document’ and ‘faxed Document’ at 72 dpi to obtain a gray scale image stored in bitmap format. Experimentally, each *fax image* file required 1.6 MB storage space approximately. For the printing process, the inkjet printer ‘Canon PIXMA MP145’ with true color image was used to generate the real fax document. For the OCR process, the C# library ‘Asprise OCR v 4.0’ [11] was used to build the OCR part in our proposed method.

Two different types of fax machine were used. The first one i.e. ‘Lexmark X8350 All-in-One’ was used to send/receive fax document with ordinary A4 white color paper, while the second one i.e. ‘Panasonic KX FT903 fax roll machine’ was used to send/receive fax document with ordinary thermal white color rolled paper. In addition, both facsimiles used in the experiments can transmit data across telephone lines in accordance with the International Telephone and Telegraph Consultative Committee (CCITT) standard of digital group 3 fax machines. For example, for a standard resolution with T.24 ITU recommendation [12] of 1,728 pels/line, fax machines support speeds with 2,400 bit/s, and typically operate at 9,600 bit/s.

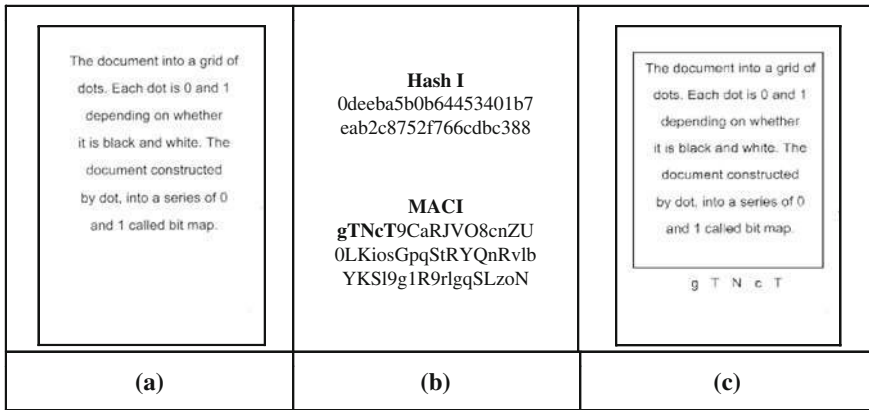


Fig. 3 **a** Original fax document at the sender side; **b** Hash value and MAC value from the computation process and **c** real fax document with frame and MAC

3.2 Experimental Results

Figures 3, 4 and 5 demonstrate some results obtained from the implementation of our proposed method. Figure 3a shows the example of original fax document to be sent from the sender side; Fig. 3b shows the hash and MAC values obtained from the OCR outcome. Finally, the ready-to-send version of fax document was produced and shown in Fig. 3c. Note that we printed only the first five characters of the resulting MAC value as a subtitle on the real fax document. This is because, according to the properties of MAC [8], it still provides enough information for message authentication and integrity purposes. In fact, any part of the MAC value can be used to detect any change on the fax content.

When the faxed document was received at another end, some errors during the transmission stage were also accompanied see Fig. 4a. After it was scanned back, and fixed for any incorrect inclination and image resolution, the area inside the image frame was separated, OCRed, hashed and encrypted to acquire the MAC value, see Fig. 4b. Another MAC value obtained from the OCR process of the area outside the image frame, noted by MAC I' is shown in Fig. 4c. Accordingly, the comparison result reported as "Match".

In case the original fax content was changed, identified by the red circle in Fig. 5a, it is obvious that the resulting MAC value from the computation process was different, compared to the one obtained from the direct OCR process, and the comparison result was hence reported as "No-Match", as shown in Fig. 5c.

We also tested the effectiveness of our proposed method on various font types and font sizes, that is, the font 'Arial' with font sizes of 36, 34, and 36 and 'Calibri' with font sizes of 28, 26, and 18. From the results obtained, any change on the fax content could be successfully detected on both types of fax document. However, the accuracy was sometimes decreased when we tested our proposed

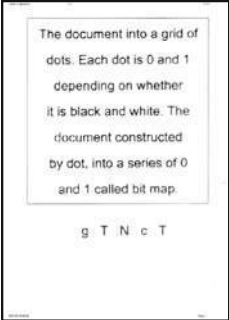
	<p>Hash II 0deeba5b0b64453401b7 eab2c8752f766cdb388</p> <p>MAC II gTNeT9CaRJVO8cnZU 0LKiosGpqStiRYQnRvIb YKSI9g1R9rlgqSLzoN</p>	<p>MACI' = gTNeT</p> <p>MACII = gTNeT</p> <p>Result = Match</p>
(a)	(b)	(c)

Fig. 4 a Faxed document at the receiver side; b Hash value and MAC value from the computation process and c the comparison result of the first five MAC characters between two identical MAC values

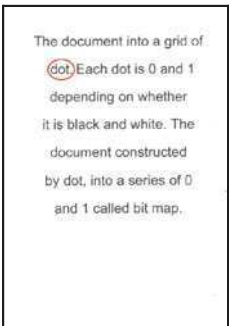
	<p>Hash III 00e3114e020171a52a1ab 44013db38ecf597654c</p> <p>MAC III uJot+Rd2e5cm5qYSKa Mqq8+nh201hP/QD69o BqUvvy2VjTN0niR9IQ</p>	<p>MACI' = gTNeT</p> <p>MACIII = uJot+</p> <p>Result = No Match</p>
(a)	(b)	(c)

Fig. 5 a Example of the modified faxed document; b Hash value and MAC value from the computation process and c the comparison result of the first five MAC characters between two different MAC values

method on ‘*Calibri*’ with the font sizes of 18 on thermal rolled paper several times. This is probably because the performance limitation of the OCR library used.

4 Conclusions

In this paper, we have presented the method of verifying message authentication and integrity for a fax document based on the use of MAC algorithms. The experimental results showed that our proposed method can practically be used to

detect any change on the faxed content sent via ordinary fax machine. Moreover, it was shown that the proposed method can also be used efficiently with different types of fax document paper. In the future, we plan to improve our proposed method to cover other different font types and font sizes. Also, we are studying to find out a higher efficient OCR algorithm to be implemented with our method.

References

1. Williams WJ, Zalubas EJ, Hero AO (2000) Word spotting in bitmapped fax documents. *Inf Retr* 2(2–3):207–226
2. Musmann HG, Preuss D (1977) Comparison of redundancy reducing codes for facsimile transmission of documents. *IEEE Trans Commun* 25(11):1425–1433
3. Garain U, Halder B (2009) Machine authentication of security documents. In: Proceedings of 10th IEEE international symposium on ICDAR, Barcelona, Spain, 26–29 July 2009, pp 718–722
4. Geisselhardt W, Iqbal T (2007) High-capacity invisible background encoding for digital authentication of hardcopy documents. In: Proceedings of IWDW, Guangzhou, China, 3–5 December 2007, pp 203–221
5. Kale S, Naphade S, Valecha V (2008) Application for a secure fax system. In: Proceedings of ICDCIT, New Delhi, India, 10–13 December 2008, pp 83–88
6. Knudsen LR, Preneel B (1998) Mac DES MAC algorithm based on DES. *Electron Lett* 34(9):871–873
7. Rivest RL (1992) The MD5 message digest algorithm. RFC 1321
8. Bellare M, Canetti R, Krawczyk H (1996) Keying hash functions for message authentication. In: Proceedings of CRYPTO, Santa Barbara, California, USA, 18–22 August 1996, pp 417–426
9. Thongkor K, Lhawchaiyapurk R, Mettripun N, Amornraksa T (2010) Enhancing method for printed and scanned watermarked documents. In: Proceedings of ITC-CSCC, Pattaya, Thailand, 4–7 July 2010, pp 977–980
10. Mettripun N, Lhawchaiyapurk R, Amornraksa T (2010) Method of rearranging watermarked pixels for printed and scanned watermarked documents. In: Proceedings of IEEE ISIT, Tokyo, Japan, 26–29 October 2010, pp 492–497
11. Asprise L (2011) Asprise OCR v 4.0: speed. accuracy simplicity portability. <http://asprise.com/home/>
12. Recommendation ITU-T T.24 (1998) Standardized digitized image set

Dynamic Grooming with Capacity aware Routing and Wavelength Assignment for WDM based Wireless Mesh Networks

Neeraj Kumar, Naveen Chilamkurti and Jongsung Kim

Abstract Wavelength division multiplexing (WDM) based wireless mesh networks (WMNs) are emerging as a new technology having enormous resources such as bandwidth and high throughput to satisfy the end users requirements. In this paper, we propose a Dynamic Grooming with Capacity aware Routing and Wavelength (DGCRW) assignment algorithm for such networks. To choose an optimized route, a cost value (CV) metric is proposed as the existing routing metric hop count does not give optimal results in WMNs for some applications (Zhao et al., J sys softw 83:1318–1326, 2010). A Utilization Matrix (UM) having load value (LV) is constructed dynamically as the requests for path/(light path) construction flow through the nodes. Using UM, utilization rating (UR) for each link is calculated. Finally CV is calculated from UR. The minimum value of CV is chosen to construct the path between source and destination. The value of CV

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (No. 2011-0014876).

N. Kumar
School of Computer Science and Engineering,
SMVD University, Katra, J&K, India
e-mail: nehra04@yahoo.co.in

N. Chilamkurti
Department of Computer Engineering,
LaTrobe University, Melbourne, Australia
e-mail: n.chilamkurti@latrobe.edu.au

J. Kim (✉)
Division of e-business, Kyungnam University, Masan, Korea
e-mail: jongsung.k@gmail.com

is compared with existing hop count metric. The proposed algorithm is simulated on WDM based NSFNet network. The results obtained show that the proposed algorithm is quite effective to route the packets to less congested path and has higher throughput and less blocking probability than the other proposed algorithms.

Keywords WDM · WMNs · Routing

1 Introduction

Over the years, wireless mesh networks (WMNs) have emerged as a new technology for providing the low cost, reliable, self healing and self configured network infrastructure for the next generation multi-hop wireless networks [1]. These networks have emerged as a popular choice for providing high speed Internet access, video on demand (VoD), Voice over IP (VoIP), videoconferencing and other high speed data access services to the end users. Typically, WMNs consist of statically positioned mesh routers (MRs) which are assumed to be reliable, scalable, and cost-effective [2]. Generally, these MRs connect with each other using wireless links and providing services to the mesh clients (MCs) which may be mobile or static. Moreover, each MR passes MCs request to Mesh Gateway (MGs). MGs are connected to internet using optical fibre. Wavelength Division Multiplexing (WDM) based networks provide enormous bandwidth, and are promising candidates for information transmission in high-speed networks [3], because fiber bandwidth is partitioned into multiple data channels in which different data can be transmitted simultaneously on different wavelengths. Traffic grooming is widely used to fill the gap between bandwidth required by a connection and available bandwidth for that connection [4]. To satisfy the MCs requests, wavelength is divided into multiple time slots and different MCs requests are mapped onto different timeslots for efficient use of available bandwidth [5, 6].

Motivated by these facts, in this paper, we propose a GD CRW algorithm for WDM based WMNs. Specifically; following are the key contributions in this paper:

- Propose a new routing metric for routing and wavelength assignment.
- Propose a new Dynamic Grooming with Capacity aware Routing and Wavelength assignment (DG CRW) algorithm based upon the defined metric.

The rest of the paper is organized as follows: [Sect. 2](#) discusses the related work, [Sect. 3](#) describes the system model, [Sect. 4](#) describes the proposed approach, [Sect. 5](#) explores on simulation results, and finally [Sect. 6](#) concludes the article.

2 Related Work

Routing with respect to traffic grooming has been studied widely in WDM networks [7]. An Integer Linear Program (ILP) formulation can be used to optimize the network throughput [8]. In this proposal, authors proposed two heuristics, namely maximizing single-hop traffic (MST) that tries to establish the light paths between source–destination pairs with the largest traffic demand, and maximizing resource utilization (MRU) that attempts to construct the light paths according to maximum resource utilization value. The problem of traffic routing with traffic grooming is proposed using Lagrangian-based heuristic [9] and graphical model [10] with an edges in the graph represent network constraints and weights in the network.

In [11], the authors propose a dynamically changing light-tree using a layered graph to solve the traffic grooming problem. In [12], the authors consider the sparse placement of grooming nodes in WDM based mesh networks. In [13–15], the authors described dynamic traffic grooming using graph model to solve the on-line traffic grooming problem. A genetic algorithm based approach is used in [16] to solve the static traffic grooming problem. This GA based approach significantly improved the network throughput as compared to the existing MST algorithm. The algorithm in [17] constructs the light paths according to availability of shortest edge disjoint paths (EDPs) for each source–destination pair and maximum resource utilization.

The multi-objective static traffic grooming and routing problem for WDM optical networks have been considered in [18]. Recently wavelength and routing assignment in optical WDM based mesh network is proposed in [19]. The problem of routing is addressed by the authors using integer linear programming (ILP) and heuristics are proposed to solve the problem of routing and wavelength assignment in such networks. A review of traffic grooming in WDM optical network is presented in [20].

3 System Model

In WMNS, due to the broadcast nature of the network, multiple flows are going over a single link in a particular interval of time. But the links have limited capacity in terms of available bandwidth and hence the aggregated sum of all the flows going over a link should be bounded by the capacity of the link, i.e.,

$$\sum_{l \in L} trans(f)^l \leq l^{cap} \quad (1)$$

Now if n^{req} are the number of requests/flows are going through a particular link l in a particular interval of time, Cap is the capacity of the link then calculate load value (LV) of the link as follows:

$$\text{Define } LV_{i,j} = \begin{cases} 0, & i = j \\ \frac{b^a}{n^{req}}, & i \neq j \end{cases}, \text{ where } i, j \text{ are indices of set } V \text{ elements,} \quad (2)$$

b_i^a is the available bandwidth of link.

$(b_i^a)^t$ is the bandwidth at time interval t , $(b_i^a)^0$ is the bandwidth initially.

Now based upon the values of LV of each link, define a $l \times l$ utilization matrix (UM) as follows:

$$UM = \begin{bmatrix} LV_{11} & LV_{12} & \dots & LV_{1l} \\ LV_{21} & LV_{22} & \dots & LV_{2n} \\ \dots & \dots & \dots & \dots \\ LV_{l1} & LV_{l2} & \dots & LV_{ll} \end{bmatrix}$$

Define utilization rating (UR) from UM as follows.

$$UR = \sum_P (UM \times J), \quad J_i = [t_{ij}^h - t_{ij}^l] \quad (3)$$

$0 \leq t_{ij}^l \leq t_{ij}^h$, $1 \leq i \leq n$ is the variation in delay known as jitter.

Once the link capacity and UM are constructed from the above equations, we formulate and propose a new routing metric called cost value (CV) for selecting a particular link route from the available ones for efficient use of available bandwidth in WDM based WMNs. The metric is defined as follows:

$$CV = \min(UR) \quad (4)$$

Each link in the network has finite CV which is used to evaluate the suitable path from the available ones. Our objective is to maximize the throughput to map the low connection requests to suitable light path with constraints in terms of finite available wavelength with limited capacity. Hence the objective function is defined as:

$$\omega^{obj} = \max \sum_1^E (\text{throughput}) \quad (5)$$

4 Proposed Approach

In this section, we will explain the proposed routing and wavelength assignment strategy. The proposed strategy is divided into two parts as routing and wavelength assignment. The shortest path is calculated dynamically between source and destination node based upon the defined metric CV. As the WMNs serve number of heterogeneous services with varying traffic demands, so static decisions of routing and wavelength assignments would not work in these networks. To satisfy the demands of end users, dynamic routing decisions are required with some defined metric. We have used CV in which capacity of the link is calculated dynamically and links are assigned based upon LV, UM and UR.

4.1 Route Discovery

Starting from source node S , requests are mapped to a light path depending upon the values of LV. Each entry of LV is made into UM that gives LV of each individual link. Based upon the values in UM, UR is calculated by multiplying the values in delay variations (jitter). For each individual link there may be many paths available but the path having minimum UR is chosen among the available ones. As the links are assigned, a routing tree (RT) is constructed. Initially, RT is empty, but as the requests are assigned to a particular link the size of the tree grows.

RT is expanded by adding the nodes in the partial routing tree starting from initial empty tree. The request for a particular light path is started from S towards destination D . If number of channels are not sufficient then the request for light path is dropped otherwise it is passed to the next intermediate node using value of CV. The intermediate nodes receive the request and process it. At each stage of the request, the intermediate nodes calculates the number of available resources in terms of UM (calculated using values of LVs) and number of free channels to establish the light path. The value of CV is calculated using these values.

As the requests are satisfied, the bandwidth consumed is taken into account and is updated accordingly. Moreover, the difference in time taken in assigning and selection of request is also calculated. The variation in time difference is known as jitter. The values of UM and jitter are multiplied to get UR. Finally after processing all the requests, the minimum value of UR is chosen that is the CV for link routing for a particular request.

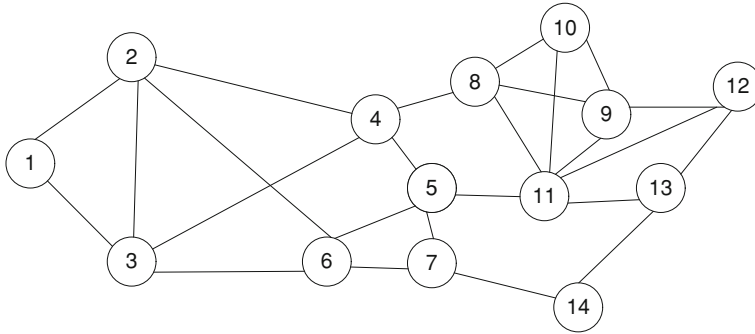


Fig. 1 Network topology used for NSFNet with 14 nodes

4.2 Wavelength Assignment

In the second phase, wavelength is assigned to the path selected in the above phase. We assume that number of wavelengths on each link is same and each node is capable of transmitting and receiving on any wavelength. Moreover, there is one optical fibre per link and all the links are bidirectional.

At any node, requests can come at any time for a particular service from MCs. If there is a direct light path existing between source and destination node, then we route current request over that light path. But if there exists no such path, then the light path is constructed dynamically (described in Sect. 4.1) using the proposed DGCRW algorithm. We start wavelength reservation over the physical route having the entire possible wavelength over a fiber and assign the wavelength using CV value dynamically.

5 Simulation Environment with Results and Discussion

5.1 Simulation Environment

To test the performance of the proposed algorithm simulation is carried out on 14-node NSFNet network as shown in Fig. 1. The performance of the proposed algorithm is measured and compared with FSP, DSP, and AP algorithms [15] with respect to the blocking probability and wavelength utilization with respect to traffic load.

DGCRW assignment algorithm**Input:** Source S , Destination D , Set of nodes V **Output:** Light path establishment with wavelength assignmentInitialize: $RT = \phi$ **repeat** S sends the route request to V_k (say) as (S, CV, D, C^n, RT) **While** $(L \neq \phi)$ Intermediate node V_k receives the request from S **If** $Cap(\lambda \in L) = B$, then

process the incoming request

else if $(\min \leq Cap(\lambda \in L) \leq B)$

Calculate the values of UM and LV as above for path

 $V_k \rightarrow V_k + 1$

Calculate the value of LV as defined in equation (2)

Calculate the value of UM as above for each link

Calculate UR and CV using UM and LV

else

Request cannot be satisfied

End if **End if** **If** $(CV^k < CV^{k-1})$ Set $CV \leftarrow CV^k$ **Else** $CV \leftarrow CV^{k-1}$ Choose a link using CV for the light path from V_k $RT \leftarrow RT \cup V_k$

Update the value of link bandwidth as follows

 $(b^a_i)^t \leftarrow (b^a_i)^0 - (n^{req})$ Propagate the message back to all the $k - 1$ nodes V_k passes the request to node V_{k+1} with updated values of CV and C^n **return** (RT) **until** $(V \neq \phi)$

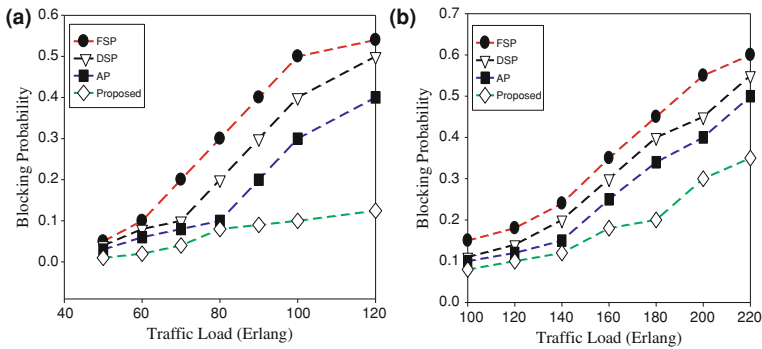


Fig. 2 Blocking probability versus traffic load for **a** $w = 50$; **b** $w = 100$ in NSFNet

5.2 Results and Discussions

5.2.1 Impact on Blocking Probability

Figure 2 show the impact of the proposed DGCRW assignment algorithm on blocking probability on NSFNet network. The results obtained show that the proposed algorithm has smallest blocking probability than all other algorithms. FSP has worst performance in all the algorithms. The possible reason for this may be due to fixed route used. Also with an increase in traffic load, the % increase in the blocking probability of the proposed algorithm is less than the other proposed algorithms. Moreover, with an increase in the wavelength from 50 to 100, there is a considerable decrease in the blocking probability. This is expected as now more wavelengths available for assignment. Again the blocking probability reduces considerably in the proposed algorithm than the other proposed algorithms.

5.2.2 Impact on Wavelength Utilization

Figure 3 show the impact of wavelength utilization in all algorithms with varying traffic load on NSFNet networks. The results obtained show that the proposed DGCRW have higher wavelength utilization with an increase in traffic load than the other proposed algorithms. This is due to that fact that all the other algorithms does not serve the new routing request due to the wavelength continuity constraints, i.e., these algorithms consider the wavelength on available which results in scattered wavelength available for each link. This is not the case with the proposed algorithm, as it selects the wavelength based upon the proposed CV metric and the calculation of available resources in terms of bandwidth is done hop by hop using CV. Hence only the best available wavelength is taken for assignment.

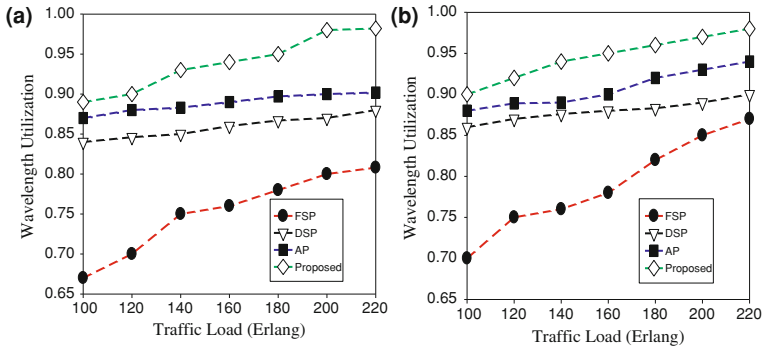


Fig. 3 Wavelength utilization. **a** $w = 50$; **b** $w = 100$ for NSFnet network

6 Conclusions

WDM based WMNs are emerging as a new technology for high speed data transfer for the next generation networks. In these types of networks, Routing and wavelength assignment is a crucial issue for efficient use of available resources. In this paper, we propose a grooming based dynamic grooming with capacity aware routing and wavelength assignment (DGCRW) algorithm for such networks. A new routing metric CV is proposed and compared with the existing hop count metric. To assign the traffic demands to a light path, UM is calculated for each link which uses LV of each link. LV is a measure of load value, i.e., each link capacity in terms of available bandwidth and number of requests serving in a particular interval of time. The requests are mapped on high speed light path based upon the entries in UM.

The performance of the proposed algorithm is compared with other proposed algorithms on NSFNet, network with respect to the metrics such as blocking probability, wavelength utilization. The results obtained show that the proposed algorithm has less blocking probability and higher wavelength utilization with varying traffic load. Hence the proposed algorithm outperforms the other algorithms in this category with respect to these metrics.

References

1. Akyildiz I, Wang X, Wang W (2005) Wireless mesh networks: a survey. *Comput Netw* 47:445–487
2. Subramanian A, Gupta H, Das S (2007) Minimum interference channel assignment in multi-radio wireless mesh networks. *SECON*, pp 481–490
3. Ramaswami R, Sivarajan KN (1998) *Optical networks: a practical perspective*. Morgan Kaufmann Publishers, Los Altos
4. Modiano E, Lin P (2001) Traffic grooming in WDM networks. *IEEE Commun Mag* 39(7):124–129

5. Srinivasan R, Somani AK (2002) Request-specific routing in WDM grooming networks. In: Proceedings of the IEEE ICC'02, vol 5. New York, pp 2876–2880
6. Srinivasan R, Somani AK (2003) Dynamic routing in WDM grooming networks. *Photonic Netw Commun* 5(2):123–135
7. Dutta R, Rouskas GN (2002) Traffic grooming in WDM networks, Past and future. *IEEE Netw* 16(6):46–56
8. Zhu K, Mukherjee B (2002) Online approaches for provisioning connections of different bandwidth granularities in WDM mesh networks. In: Proceedings of the OFC'02, Anaheim, pp 549–551
9. Patrocinio Z, Mateus G (2003) A Lagrangian-based heuristic for traffic grooming in WDM optical networks. *IEE GLOBECOM'03*, San Francisco, pp 2767–2771
10. Zhu H, Zang H, Zhu K, Mukherjee B (2003) A novel generic graph model for traffic grooming in heterogeneous WDM mesh networks. *IEEE/ACM Trans Networking* 11(2):285–299
11. Huang X, Farahmand F, Jue JP (2004) An algorithm for traffic grooming in WDM mesh networks with dynamically changing light-trees. *GLOBECOM'04*, Dallas, pp 1813–1817
12. Zhu H, Zang H, Mukherjee B (2002) Design of WDM mesh networks with sparse grooming capability. *IEEE Globecom'02* 3:2696–2700 November
13. Zhu H, Zang H, Zhu K, Mukherjee B (2002) Dynamic traffic grooming in WDM mesh networks using a novel graph model. *GLOBECOM'02*, Taiwan, pp 2681–2685
14. Zhu H, Zang H, Zhu K, Mukherjee B (2003) Dynamic traffic grooming in WDM mesh networks using a novel graph model. *Opt Netw Mag* 4(3):65–75
15. Zhu H, Zang H, Zhu K, Mukherjee B (2003) A novel generic graph model for traffic grooming in heterogeneous WDM mesh networks. *IEEE Trans Networking* 11(2):285–299
16. De T, Pal P, Sengupta A (2008) A genetic algorithm based approach for traffic grooming, routing and wavelength assignment in optical WDM mesh networks. In: Proceedings of IEEE ICON'08, December 2008
17. Choo MYH, Lee S, Lee TJ (2005) Chung Traffic grooming algorithms in shortest EDP stable in WDM mesh networks. In: Proceedings of ICCS'05, May 2005. *Lecture Notes in Computer Science*, vol 3516. pp 559–567
18. Prathombutr P, Stach J, Park EK (2005) An algorithm for traffic grooming in WDM optical mesh networks with multiple objectives. *J Telecommun Sys* 28:3–4, 369–386
19. Tanmay D, Jain P, Pal A (2010) Distributed dynamic grooming routing and wavelength assignment in wdm optical mesh networks. *Photon Netw Commun* 1–10
20. Zhu K, Mukherjee B (2003) A review of traffic grooming in WDM optical networks: architectures and challenges. *Opt Netw Mag* 4(2):55–64
21. Zhao L, Ahmed Y, Min G (2010) GLBM: a new QoS aware multicast scheme for wireless mesh networks. *J sys softw* 83:1318–1326

Weakness in a User Identification Scheme with Key Distribution Preserving User Anonymity

Taek-Youn Youn and Jongsung Kim

Abstract Recently, Hsu and Chuang proposed a novel user identification scheme with key distribution for distributed computer networks. The Hsu-Chuang scheme permits a user to anonymously log into a system and establish a secret key shared with the system. In this paper, we show that the Hsu-Chuang scheme is not secure against known session key attacks. To show the insecurity, we describe an adversary who can recover the private key of a user by performing known session key attacks. We also provide a countermeasure which can be used for enhancing the security of the Hsu-Chuang scheme.

Keywords Security · User identification · Key distribution · Distributed computer network

1 Introduction

Distributed computer networks permit host computers and user terminals which are connected into the same network to share information and computing power. In these days, it is increasingly important to secure the communications conducted

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (No. 2011-0014876).

T.-Y. Youn
Electronics and Telecommunications Research Institute (ETRI),
dDaejeon, Korea
e-mail: taekyoung@etri.re.kr

J. Kim (✉)
Kyungnam University, Masan, Korea
e-mail: jongsungk@kyungnam.ac.kr

over distributed computer networks, and the following problems are considered as fundamental requirements for secure distributed computer networks:

- User identification: when a user wants to obtain access privilege which is linked to a particular identity, we need some mechanisms for identifying the legitimacy of the requesting user.
- Key distribution: to achieve the confidentiality of communications, we need a tool for establishing a common secret among two or more communicating parties.
- User anonymity: to prevent an adversary from obtaining sensitive personal information from communicating messages, it is desirable to use a protocol which provides user anonymity.

For achieving the above mentioned security goals, several schemes have been proposed [2–4, 6, 7], and most of them have shown to be insecure. In [3], Lee and Chang proposed a user identification scheme with the above mentioned security features. However, Wu and Hsu showed that the Lee-Chang scheme is insecure against impersonation and identity disclosure attacks, and proposed an improved scheme [6]. Immediately, Yang et al. showed that the Wu-Hsu scheme is insecure against a compromising attack, and proposed a modified scheme [7]. Thereafter, Mangipudi and Katti found a vulnerability of the Yang et al. scheme [4]. They also proposed a modification in order to resist their attack. However, recently, Hsu and Chuang demonstrated that the Yang et al. scheme and the Mangipudi-Katti scheme are vulnerable to an identity disclosure attack [2]. Moreover, Hsu and Chuang proposed a user identification scheme which achieves all security goals considered in the literature.

In this paper, we examine the security of the user identification scheme proposed by Hsu and Chuang [2], and show that the identification scheme is not secure. At first, we describe a known session key attack to show the insecurity of the Hsu-Chuang scheme. An adversary can recover the private key of a user by performing the known session key attack. Note that service providers know several session keys of their clients since they share the session keys with their clients. Hence a malicious service provider can recover the private key of its client user by performing the known session key attack using previously shared session keys. To resist against our known session key attack, we propose a simple countermeasure. The proposed countermeasure can enhance the security of the Hsu-Chuang scheme with few additional cost.

2 Review of the Hsu-Chuang Scheme

In the section, we briefly recall the Hsu-Chuang scheme which consists of three phases, the system initialization phase, the registration phase, and the user identification phase.

2.1 Description

System Initialization Phase The trusted authority (TA) chooses two large primes p and q , computes $N = pq$, and determines e and d such that $ed = 1 \pmod{\phi(N)}$, where $\phi(N) = (p - 1)(q - 1)$. The TA randomly chooses an element $g \in \mathbb{Z}_N^*$. The TA publishes (e, N, g) as system parameters, and privately keeps (d, p, q) . Let $E_K(m)$ and $D_K(m)$ be the encryption and decryption of an input message m with a key K , respectively. Let $h(\cdot)$ be a cryptographic hash function.

Registration Phase The user U_i (or the service provider P_i) submits its identity ID_i to the TA. Then, the trusted authority generates the requester's private key as $S_i = ID_i^d \pmod N$. Then, the TA securely sends S_i to the requester U_i (or P_i).

User Identification Phase If the user U_i wants to gain an access privilege from the service provider P_j , the user U_i and the service provider P_j cooperatively perform the following steps:

1. U_i submits the service request to P_j .
2. P_j uses his private key S_j to compute $Z = g^k \cdot S_j \pmod N$ for randomly chosen k , and sends Z to U_i .
3. On receiving Z , U_i chooses a random value t and computes $a = Z^e \cdot ID_j^{-1} \pmod N$, $K_{ij} = a^t \pmod N$, $w = g^{et} \pmod N$, $x = S_i^{h(K_{ij}||Z||w||T)} \pmod N$, and $y = E_{K_{ij}}(ID_i)$, where T is the current timestamp. Then, U_i sends (w, x, y, T) to P_j . Note that, the value K_{ij} is used as a session key.
4. After receiving (w, x, y, T) , P_j verifies the validity of T . If it is invalid, P_j aborts the protocol; otherwise, P_j computes $K_{ij} = w^k \pmod N$ and uses the key K_{ij} to decrypt y as $ID_i = D_{K_{ij}}(y)$. If the equation $ID_i^{h(K_{ij}||Z||w||T)} = x^e \pmod N$ holds, P_j is convinced that U_i is an authorized user.
5. P_j computes $D_i = h(K_{ij}||T'||Z||ID_i||ID_j)$ and sends (D_i, T') to U_i , where T' is the current timestamp.
6. On receiving (D_i, T') , U_i checks the validity of T' . If it is valid, U_i computes $D'_i = h(K_{ij}||T'||Z||ID_i||ID_j)$ and tests if $D'_i = D_i$. If it holds, U_i is convinced that P_j is the valid service provider.

3 Weakness of Hsu-Chuang Scheme

Though the security of the Hsu-Chuang scheme against known session key attacks is not considered in [2], it is obvious that known session key attacks are serious threat against key distribution and key exchange schemes which are used for establishing session keys. However, unfortunately, the Hsu-Chuang scheme is insecure against a known session key attack. To show the insecurity, we describe

an adversary who can recover the private key of a user or disguise a user using known session keys.

3.1 A Known Session Key Attack

Note that, we need at least two session keys to mount our attack. Hence, we consider the scenario where an adversary obtains two session keys K_1 and K_2 . For $\ell \in \{1, 2\}$, let $\{Z_\ell, w_\ell, x_\ell, y_\ell, T_\ell, D_\ell, T'_\ell\}$ be the set of communicating messages generated for establishing the session key K_ℓ . Note that the messages are transmitted through open networks, and so they are visible to the adversary. Using known values, the adversary can compute two hash values

$$H_1 = h(K_1 || Z_1 || w_1 || T_1) \text{ and } H_2 = h(K_2 || Z_2 || w_2 || T_2).$$

Then, the adversary can obtain the following relations:

$$x_1 = S_i^{H_1} \text{ mod } N \text{ and } x_2 = S_i^{H_2} \text{ mod } N.$$

We can consider two cases according to the greatest common divisor of H_1 and H_2 .

Case 1: $\text{gcd}(H_1, H_2) = 1$ In this case, the adversary can recover the private key of the user U_i . Since $\text{gcd}(H_1, H_2) = 1$, the adversary can find two integers α and β such that $\alpha H_1 + \beta H_2 = 1$ by using the extended Euclidean algorithm (EEA). Then, the private key of the user U_i can be computed by

$$S_i = S_i^{\alpha H_1 + \beta H_2} = (S_i^{H_1})^\alpha \cdot (S_i^{H_2})^\beta = x_1^\alpha \cdot x_2^\beta \text{ mod } N.$$

Note that, in the literature [5], it is well-known that the probability that two random numbers are relatively prime is $6/\pi^2 \approx 0.6$. We can apply this fact to our case and obtain $\Pr[\text{gcd}(H_1, H_2) = 1] \approx 0.6$ since the outputs of the hash function $h(\cdot)$ are random. Hence, an adversary can recover the private key of a user with high probability using two session keys of the user.

Case 2: $\text{gcd}(H_1, H_2) \neq 1$ In this case, the adversary can find two integers α and β such that $\alpha H_1 + \beta H_2 = d$ using the EEA where $d = \text{gcd}(H_1, H_2)$, and the values can be used for computing

$$S_i^d = S_i^{\alpha H_1 + \beta H_2} = (S_i^{H_1})^\alpha \cdot (S_i^{H_2})^\beta = x_1^\alpha \cdot x_2^\beta \text{ mod } N.$$

Note that the adversary cannot recover the private key of the user U_i since it is hard to compute the d -th root of S_i^d , but he can disguise the user U_i by generating a set of valid communicating messages $\{w, x, y, T\}$. For generating such messages, the adversary searches a random t such that $d | h(K || Z || w || T)$ where $K = (Z^e \cdot ID_j^{-1})^t \text{ mod } N$, $w = g^{et} \text{ mod } N$, $y = E_K(ID_i)$, and T is the current timestamp.

Since the outputs of hash function are random, the adversary can find desired random value t within d trials. Then, the adversary can compute the authenticating message x as following:

$$x = (S_i^d)^{h(K||Z||w||T)/d} = S_i^{h(K||Z||w||T)} \bmod N.$$

If the greatest common divisor d is large integer, it is not easy to find a set of valid communicating messages. For any integer d , we have the following relation:

$$\Pr[\gcd(H_1, H_2) = d] \leq \Pr[d|H_1] \cdot \Pr[d|H_2] = \frac{1}{d^2}.$$

Hence, the probability $\Pr[\gcd(H_1, H_2) = d]$ is very small for large d . In other words, $\gcd(H_1, H_2)$ is not large integer with high probability. As a result, if $\gcd(H_1, H_2) \neq 1$, an adversary can succeed in disguising the target user with high probability.

3.2 Security Against Malicious Service Providers

Undoubtedly, the proposed known session key attack is serious threat against the Hsu-Chuang scheme. However, in practice, it seems to be hard to mount known session key attacks because it is not easy to obtain session keys of a user. However, in the Hsu-Chuang scheme, service providers can easily collect session keys of their clients. Hence, a malicious service provider can easily mount our known session key attack by collecting session keys of a user. A malicious service provider P_j can collect session keys of a target user U_i by performing one of the following two attack strategies.

Strategy 1 Note that, any legitimate service provider can share a session key with the U_i by performing legitimate user identification phase with the user. Hence, a malicious service provider can easily mount our known session key attack by storing two or more session keys when he performs legitimate user identification phases with the user. As a result, a service provider can easily recover the private key of a target user if the service provider bears ill will.

Strategy 2 If U_i logs only once into the system controlled by a malicious service provider P_j , the service provider cannot collect two or more session keys of the user by executing the first strategy. However, a malicious service provider still can obtain two or more session keys if the user trusts the service provider P_j . In this case, we assume that the user restarts user identification phase when the protocol execution is aborted by some reasons, and the user performs user identification phase with P_j until the protocol is finished successfully. For obtaining several different session keys of the user, the service provider performs user identification phase with the user as following:

1. U_i submits the service request to P_j .
2. P_j computes $Z = g^k \cdot S_j \bmod N$ for randomly chosen k , and sends Z to U_i .
3. On receiving Z , U_i chooses a random value t and computes $a = Z^e \cdot ID_j^{-1} \bmod N$, $K_{ij} = a^t \bmod N$, $w = g^{et} \bmod N$, $x = S_i^{h(K_{ij}||Z||w||T)} \bmod N$, and $y = E_{K_{ij}}(ID_i)$, where T is the current timestamp. Then, U_i sends (w, x, y, T) to P_j . Note that, the value K_{ij} is used as the session key.
4. After receiving (w, x, y, T) , P_j aborts the protocol without verifying given values, computes $K_{ij} = w^k \bmod N$ and $H = h(K_{ij}||Z||w||T)$, and stores (x, H) .

Since the user trusts the malicious service provider, he will initiate the user identification phase again until the protocol is successfully finished. Then, the malicious service provider can obtain sufficient information by performing the above procedure iteratively.

4 Countermeasure

In this section, we provide a simple countermeasure that can be used for enhancing the security of the Hsu-Chuang scheme with few additional cost.

4.1 Basic Idea

Let E be an adversary who performs the known session key attack described in Sect. 3. Then we can assume that E can obtain (two or more) session keys.

Before to introduce our countermeasure, we briefly review the weakness of the Hsu-Chuang scheme. As described in Sect. 3, the private key S_j of the user U_j can be extracted from $x = S_j^{h(K_{ij}||Z||w||T)} \bmod N$ only if we can evaluate the hash value $h(K_{ij}||Z||w||T)$. Note that the session key K_{ij} is the only secret information among four input messages K_{ij} , Z , w , and T . Therefore we need the session key to recover the private key S_j . Since we assumed that E knows the session key, the Hsu-Chuang scheme is insecure against the adversary E who performs the known session key attack.

To achieve the security against known session key attacks, the Hsu-Chuang scheme should be modified so that the adversary cannot extract the private key even though several session keys are revealed to the adversary. In the Hsu-Chuang scheme, the insecurity is caused by the use of the message x which is used only for authenticating the message $K_{ij}||Z||w||T$. Hence the security of the Hsu-Chuang scheme can be improved if we modify the authenticating message x so that the modified message securely authenticates the message $K_{ij}||Z||w||T$ even though all input messages (including the session key K_{ij}) are revealed to the adversary. Since signature schemes securely authenticate a message even though input messages are

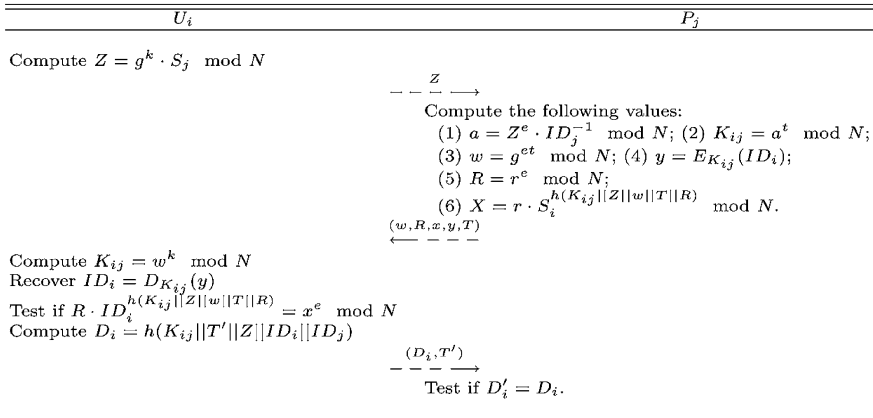


Fig. 1 User identification phase of improved scheme

public information, we can enhance the security of the Hsu-Chuang scheme by replacing x with a digital signature generated by a secure signature scheme Fig. 1.

4.2 Description of Improved Scheme

Note that, in [2], x is used as a kind of identity-based signature on the message $K_{ij}||Z||w||T$, and thus we will use a secure identity-based signature scheme for generating an authenticating message. To improve the security of the Hsu-Chuang scheme, we use the identity-based signature proposed by Guillou and Quisquater (GQ-IBS) [1].

System Initialization Phase The trusted authority (TA) chooses two large primes p and q , computes $N = pq$, and determines e and d such that $ed = 1 \pmod{\phi(N)}$, where $\phi(N) = (p - 1)(q - 1)$. The above mentioned parameters N , e , and d are chosen as described in [1]. The TA randomly chooses an element $g \in \mathbb{Z}_N^*$. The TA publishes (e, N, g) as system parameters, and privately keeps (d, p, q) . Let $E_K(m)$ and $D_K(m)$ be the encryption and decryption of an input message m with a key K , respectively. Let $h(\cdot)$ be a cryptographic hash function.

Registration Phase The user U_i (or the service provider P_i) submits its identity information id_i to the TA. Then, the trusted authority generates the requester’s private key as $S_i = ID_i^d \pmod N$ where ID_i is the hashed value of id_i . Then, the TA securely sends S_i to the requester U_i (or P_i).

User Identification Phase If the user U_i wants to gain an access privilege from the service provider P_j , the user U_i and the service provider P_j cooperatively perform the following steps:

1. U_i submits the service request to P_j .
2. P_j uses his private key S_j to compute $Z = g^k \cdot S_j \bmod N$ for randomly chosen k , and sends Z to U_i .
3. On receiving Z , U_i chooses two random values t and r . Then the user computes $a = Z^e \cdot ID_j^{-1} \bmod N$, $K_{ij} = a^t \bmod N$, $w = g^{et} \bmod N$, $R = r^e \bmod N$, $X = r \cdot S_i^{h(K_{ij}||Z||w||T||R)} \bmod N$, and $y = E_{K_{ij}}(ID_i)$, where T is the current timestamp. Then, U_i sends (w, R, X, y, T) to P_j . Note that, the value K_{ij} is used as the session key.
4. After receiving (w, R, X, y, T) , P_j verifies the validity of T . If it is invalid, P_j aborts the protocol; otherwise, P_j computes $K_{ij} = w^k \bmod N$ and uses the key K_{ij} to decrypt y as $ID_i = D_{K_{ij}}(y)$. If the equation $R \cdot ID_i^{h(K_{ij}||Z||w||T||R)} = X^e \bmod N$ holds, P_j is convinced that U_i is an authorized user.
5. P_j computes $D_i = h(K_{ij}||T'||Z||ID_i||ID_j)$ and sends (D_i, T') to U_i , where T' is the current timestamp.
6. On receiving (D_i, T') , U_i checks the validity of T' . If it is valid, U_i computes $D'_i = h(K_{ij}||T'||Z||ID_i||ID_j)$ and tests if $D'_i = D_i$. If it holds, U_i is convinced that P_j is the valid service provider.

4.3 Security

The only difference between the improved scheme and the Hsu-Chuang scheme is the authenticating message, and thus the improved scheme has all security properties of the Hsu-Chuang scheme for the same reason. Hence, detailed discussions for the security features are not included in this paper.

Here we analyze the security of the improved identification scheme against known session key attacks. In the improved scheme, x is replaced by $\{R, X\}$ which is a signature generated by the GQ-IBS scheme. In the original description of the GQ-IBS scheme, the signed messages $K_{ij}||Z||w||T||R$ is also included as a signature. However, in the improved scheme, the signed message is not included as a part of authenticating messages. Though the adversary who performs known session key attacks can obtain the signed message, it is hard to extract S_i from R , X and $K_{ij}||Z||w||T||R$ since the GQ-IBS scheme is a secure signature scheme. Therefore, the improved scheme is secure against known session key attacks.

5 Conclusion

In this paper, we showed the insecurity of the Hsu-Chuang scheme by describing a known session key attack, in which an adversary can recover the private key of a user or disguise a user. We also showed that a malicious service provider can easily recover the private key of a user by executing a number of legitimate runs of

the scheme. Moreover, we provide a simple countermeasure which can enhance the security of the Hsu-Chuang scheme.

References

1. Guillou LC, Quisquater J-J (1988) A paradoxical indentity-based signature scheme resulting from zero-knowledge. In: Proceedings of Crypto'88, LNCS 403. Springer, Berlin, pp 216–231
2. Hsu C-L, Chuang Y-H (2009) A novel user identification scheme with key distribution preserving user anonymity for distributed computer networks. *Inf Sci* 179:422–429
3. Lee WB, Chang CC (1999) User identification and key distribution maintaining anonymity for distributed computer network. *Comput Syst Sci Eng* 15(4):113–116
4. Mangipudi K, Katti R (2006) A secure identification and key agreement protocol with user anonymity (SIKA). *Comput Secur* 25(6):420–425
5. Nymann JE (1972) On the probability that positive integers are relatively prime. *J Number Theory* 4:469–473
6. Wu TS, Hsu CL (2004) Efficient user identification scheme with key distribution preserving anonymity for distributed computer networks. *Comput Secur* 23(2):120–125
7. Yang Y, Wang S, Bao F, Wang J, Deng RH (2004) New efficient user identification and key distribution scheme providing enhanced security. *Comput Secur* 23(8):697–704

A Compact S-Box Design for SMS4 Block Cipher

Imran Abbasi and Mehreen Afzal

Abstract This paper proposes a compact design of SMS4 S-box using combinational logic which is suitable for the implementation in area constraint environments like smart cards. The inversion algorithm of the proposed S-box is based on composite field $GF(((2^2)^2)^2)$ using normal basis at all levels. In our approach, we examined all possible normal basis combinations having trace equal to one at each subfield level. There are 16 such possible combinations with normal basis and we have compared the S-box designs based on each case in terms of logic gates it uses for implementation. The isomorphism mapping and inverse mapping bit matrices are fully optimized using greedy algorithm. We prove that our best case reduces the complexity upon the SMS4 S-box design with existing inversion algorithm based on polynomial basis by 15% XOR and 42% AND gates.

Keywords Composite field arithmetic · SMS4 · Normal basis · S-box

1 Introduction

SMS4 is the mandatory block cipher standard for securing Wireless Local Area Network (WLAN) devices in China. The Office of State Commercial Cipher Administration of China (OSCCA) released the cipher description in January, 2006 [1] and the English version of the document is published by Diffie and Ledin [2].

I. Abbasi (✉) · M. Afzal
College of Telecommunication (MCS),
National University of Sciences and Technology, Islamabad, Pakistan
e-mail: imranabbasi@mcs.edu.pk

M. Afzal
e-mail: mehreenafzal@mcs.edu.pk

SMS4 is used in WLAN Authentication and Privacy Infrastructure (WAPI) standard in order to provide data confidentiality. The Chinese WLAN industry widely uses WAPI, and it is supported by many international corporations like SONY in the relevant products.

The efficiency of SMS4 hardware implementation in terms of power consumption, area and throughput mainly depends upon the implementation of its S-box. It is the most computationally intensive operational structure of SMS4 as it comprises of non-linear multiplicative inversion. The designers of the SMS4 had chosen its S-box design similar to Rijndael which employs inversion base mapping [3]. Implementing a circuit to find the multiplicative inverse in the $GF(2^8)$ using Extended Euclidean algorithm or Fermat theorem is very complex and costly. Several architectures of $GF(2^8)$ inverter have been proposed by researchers over the period of time for area efficient implementation of S-boxes that comprises of inversion in their algebraic expressions. An efficient way to implement S-box is to use combinational logic because it requires small area for implementation. Rijmen [4] proposed the first hardware implementation of AES S-box using composite field representation. The proposed design suggested the use of Optimal Normal Basis for efficient inversion in $GF(2^8)$. Wolkerstorfer [5] and Rudra [6] implemented the AES S-box by representing $GF(2^8)$ as a quadratic extension of the $GF(2^4)$ using polynomial basis. In this approach a byte in $GF(2^8)$ is first decomposed into linear polynomial with coefficients in $GF(2^4)$ and different arithmetic operations in $GF(2^4)$ are computed using combinational logic. The inversion in hardware is then implemented with the simple logic gates by further decomposing $GF(2^4)$ into $GF(2^2)$ operations. Satoh [7] and Mentens [8] further optimized the hardware implementation of AES S-box by applying a composite field with multiple extensions of smaller degrees. The tower field $GF(2^8) \rightarrow GF(((2^2)^2)^2)$ is constructed with repeated degree 2 extensions using polynomial basis. Canright in [9] analyzed all possible combinations of normal and polynomial basis at subfield levels of $GF(((2^2)^2)^2)$ and proved that use of normal bases at all levels of composite field decomposition further reduces the area of the AES S-box implementation. Bai [10] proposed a $GF(2^8)$ inversion algorithm for SMS4 S-box based on slight modification of design in [5].

In this paper, a new combinational structure of SMS4 S-box with the inversion algorithm in tower field representation $GF(2^8) \rightarrow GF(((2^2)^2)^2)$ based on normal basis, has been proposed. We have analyzed all possible combinations of normal basis at each level with trace one from the field generated by irreducible primitive polynomial of SMS4 cipher. The comparison of our resulting best case architecture with the S-box design based on proposed $GF(2^8)$ inverter of [10] is also given.

The organization of the rest of paper is as follows. In subsequent section, structure of SMS4 block cipher is briefly described with the focus on its S-box. In Sect. 3, the design of S-box using the composite field representation with normal basis is explicated. Section 4 gives the comparison of combinatorial S-box designs of SMS4 with different normal basis combinations at subfield level. In Sect. 5, a comparative analysis is given between our proposed design of S-box with the one based on the inversion algorithm presented in [10]. Conclusions and work in progress are stated in Sect. 6.

2 The SMS4

SMS4 block cipher is based on the iterative feistel structure with input, output, and key size of 128 bits each. The data input is divided into four 32 bit words. The algorithm comprises of 32 rounds, and in each round one word is modified by adding it to other three words with a keyed function. Encryption and decryption processes have the similar structure and only the key schedule is reversed. For the detailed description of cipher one may refer to [2]. The official depiction of SMS4 S-box is given as a lookup table (LUT) with 256 entries. The S-box is commonly implemented with the ROM lookup table where the pre-computed values are stored. However, significant hardware resources are required if lookup table is implemented with 16×16 entries. SMS4 S-box is bijective and it substitutes byte input for byte output using arithmetic computations over $GF(2^8)$. A method suitable for hardware implementation of S-box is to first perform affine transformation on $GF(2)$, then carry out inversion in $GF(2^8)$, followed by second affine transformation over $GF(2)$ [3, 11]. The S-box algebraic structure is given as the following expression [11].

$$s(x) = A_2(A_1 \cdot x + C_1)^{-1} + C_2. \tag{1}$$

The row vectors are $C_1 = 0xCB = (11001011)_2$ and $C_2 = 0xD3 = (11010011)_2$. The cyclic matrices A_1 and A_2 in the algebraic expression are as below:

$$A_1 = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \end{bmatrix} \quad A_2 = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix} \tag{2}$$

The irreducible primitive polynomial in $GF(2^8)$ is

$$f(x) = (x^8 + x^7 + x^6 + x^5 + x^4 + x^2 + 1). \tag{3}$$

3 SMS4 S-box Design in Composite Field

In this section we describe the proposed SMS4 combinatorial structure based on composite field $GF(((F^2)^2)^2)$ in normal basis with the logical equations for inversion, multiplications, squaring and addition. SMS4 S-box design in composite field arithmetic is more efficient than using ROM/RAM for lookup tables (LUT) in

area constrained environments [10]. All finite fields of same cardinality are isomorphic but their arithmetic efficiency depends significantly on the choice of basis that is used for the field element representation. For the hardware implementation, normal basis has significant advantage over polynomial basis as mathematical operations in normal basis representation generally comprises of rotation, shifting and XORing [12, 13].

3.1 $GF(2^8)$ Inversion Algorithm using Normal Basis

For input byte x to SMS4 S-box, inverse is computed for the expression $(A_1.x + C_1)$. The complexity of basis conversion is dependent on the selected irreducible polynomial and if the polynomial is adequately chosen, the basis conversion is simple [8]. Following are the irreducible polynomials and their corresponding normal basis representation.

$$\begin{aligned} GF(2^2) & : z^2 + z + 1 \rightarrow (z + Z)(z + Z^2) \quad \text{Normal basis}(Z^2, Z) \\ GF\left((2^2)^2\right) & : y^2 + Ty + N \rightarrow (y + Y)(y + Y^4) \quad \text{Normal basis}(Y^4, Y) \\ GF\left(\left((2^2)^2\right)^2\right) & : x^2 + \tau x + \eta \rightarrow (x + X)(x + X^{16}) \quad \text{Normal basis}(X^{16}, X) \end{aligned} \quad (4)$$

where $T = Y^4 + Y$ is the trace and $N = Y^4.Y$ is the norm in $GF(2^4)/GF(2^2)$, $\tau = X^{16} + X$ is the trace and $\eta = X^{16}.X$ is the norm in $GF(2^8)/GF(2^4)$. To minimize the operations and simplify inversion circuit in composite field we consider only those basis combinations which have $\tau = T = 1$. The nested structure of $GF(2^8)$ inverter comprises of different subfield operations. In the following sections logical structures for inversion, multiplication and scaling in composite field are given.

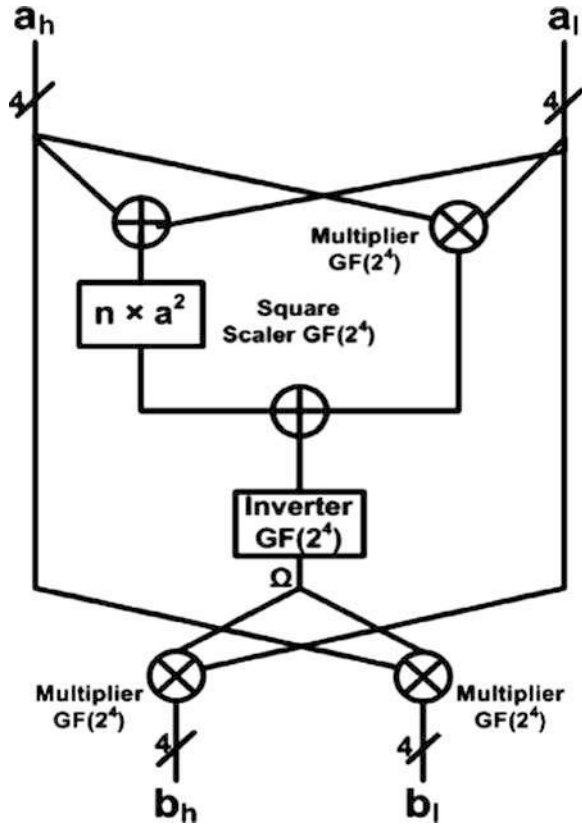
Inversion in $GF(2^8)$, $GF(2^4)$ and $GF(2^2)$. Let the pair $(a_h, a_l) \in GF(2^4)$ represents $a \in GF(2^8)$ in terms of Normal basis (X^{16}, X) . If $b \in GF(2^8)$ is inverse of a , then product of a and b is 1.

$$\begin{aligned} a & = a_h X^{16} + a_l X \\ b & = b_h X^{16} + b_l X \\ a \times b & = (a_h X^{16} + a_l X)(b_h X^{16} + b_l X) = 1. \end{aligned} \quad (5)$$

Substituting $X + X^{16} = 1$, $(X^{16})^2 = X^{16} + n$ and $(X)^2 = X + n$ and solving for b_h and b_l .

$$\begin{aligned} b_h & = \left[(a_h \otimes a_l) \otimes \left((a_h \otimes a_l)^2 \otimes n \right) \right]^{-1} \otimes a_l. \\ b_l & = \left[(a_h \otimes a_l) \otimes \left((a_h \otimes a_l)^2 \otimes n \right) \right]^{-1} \otimes a_h. \end{aligned} \quad (6)$$

Fig. 1 $GF(2^8)$ inverter



where \otimes is multiplication and \oplus is addition in $GF(2^4)$. If $\Omega = [(a_h \otimes a_l) \oplus ((a_h \oplus a_l)^2 \otimes n)]^{-1}$, then inversion in $GF(2^8)$ is expressed by following relation.

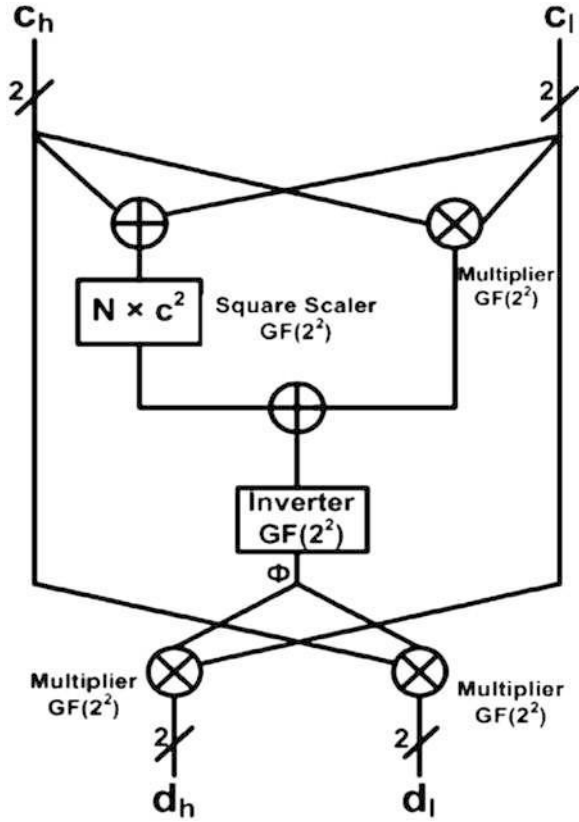
$$b = a^{-1} = (\Omega \otimes a_l)X^{16} + (\Omega \otimes a_h)X. \tag{7}$$

The logical structure of $GF(2^8)$ inverter is shown in Fig. 1. Similarly, if $c \in GF(2^4)$ and it has an inverse $d \in GF(2^4)$ using normal basis (Y^4, Y) , then $c = c_h Y^4 + c_l Y$, $c_h, c_l \in GF(2^2)$ and $d = d_h Y^4 + d_l Y$, $d_h, d_l \in GF(2^2)$. If \otimes is multiplication and \oplus is bitwise addition in $GF(2^2)$ and $\Phi = [(c_h \otimes c_l) \oplus ((c_h \oplus c_l)^2 \otimes N)]^{-1}$, then equation for $GF(2^4)$ inversion is given as below:

$$d = c^{-1} = (\Phi \otimes c_l)Y^4 + (\Phi \otimes c_h)Y. \tag{8}$$

The $GF(2^4)$ inverter is depicted in Fig. 2. The inversion in $GF(2^2)$ is same as squaring and implemented without gates by swapping of bits. If $e \in GF(2^2)$ is represented in normal basis (Z^2, Z) as $e = e_h Z^2 + e_l Z$, $e_h, e_l \in GF(2)$ and f is the inverse of e in $GF(2^2)$ then inversion in $GF(2^2)$ is:

Fig. 2 $GF(2^4)$ inverter



$$f = e^{-1} = (e_l)Z^2 + (e_h)Z. \tag{9}$$

Multiplication in $GF(2^4)$ and $GF(2^2)$. The structures of multipliers in $GF(2^4)$ and $GF(2^2)$ in normal basis are derived as below.

$$(c_h Y^4 + c_l Y)(d_h Y^4 + d_l Y) = c_h d_h (Y^4)^2 + c_h d_l Y^4 Y + c_l d_h Y^4 Y + c_l d_l Y^2 \tag{10}$$

Substituting $Y + Y^4 = 1$, $(Y^4)^2 = Y^4 + N$ and $(Y)^2 = Y + N$.

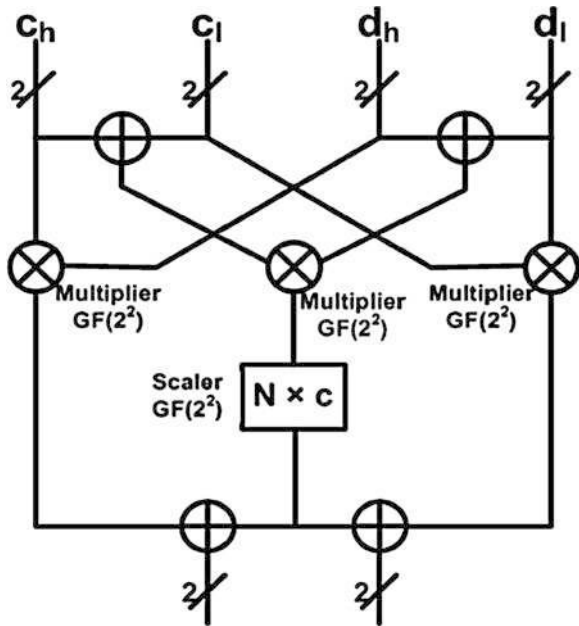
$$(c_h Y^4 + c_l Y)(d_h Y^4 + d_l Y) = (c_h d_h \oplus \epsilon) Y^4 + (c_h d_l \oplus \epsilon) Y. \tag{11}$$

Where \oplus is bit wise addition, \otimes is multiplication in $GF(2^2)$ and $\epsilon = (c_h \oplus c_l) \otimes (d_h \oplus d_l) \otimes N$. Similarly $GF(2^2)$ multiplier in normal basis is represented as:

$$(e_h Z^2 + e_l Z)(f_h Z^2 + f_l Z) = (e_h f_h \oplus \wedge) Z^2 + (e_l f_l \oplus \wedge) Z. \tag{12}$$

\oplus represents the bit addition, \otimes is AND operation and $\wedge = (e_h \oplus e_l) \otimes (f_h \oplus f_l)$. The above mentioned structures are illustrated in Figs. 3 and 4 respectively.

Fig. 3 $GF(2^4)$ multiplier



Scaling and Squaring in $GF(2^4)$ and $GF(2^2)$. In $GF(2^8)$ and $GF(2^4)$ inverters there are constant multiplication operations ($n \times a^2$) and ($N \times c^2$) and in $GF(2^4)$ multiplier there is constant multiplication term ($N \times c$). The combination of squaring and scaling operation results in further optimization [9]. The computation of these terms depends on the values of n in $GF(2^4)$ and N in $GF(2^2)$ for the chosen normal basis. $N \in GF(2^2)$ and N is not equal to zero or one, therefore N and $N + 1$ are the roots of $z^2 + z + 1$. So depending on the choice of basis, scalars for N and N^2 implies to scalars for z or z^2 . The two bit factor ($N \times c$) is given in two ways.

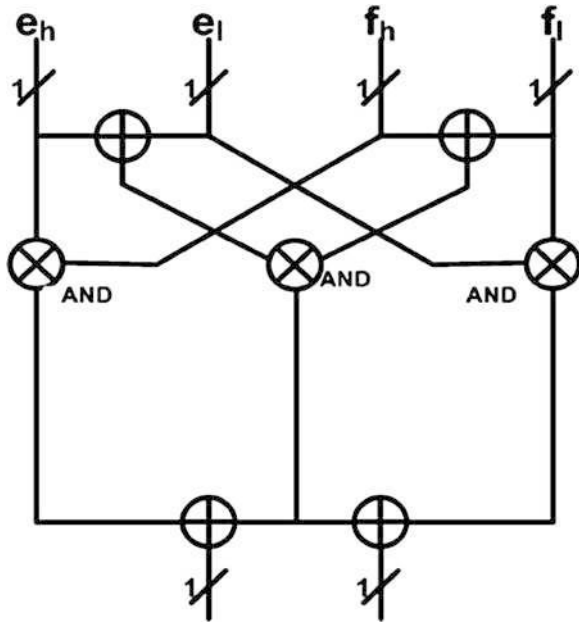
$$\begin{aligned} Z \times (e_h Z^2 + e_l Z) &= (e_h \oplus e_l) Z^2 + e_h Z. \\ Z^2 \times (e_h Z^2 + e_l Z) &= e_l Z^2 + (e_h \oplus e_l) Z. \end{aligned} \tag{13}$$

Similarly the square scaling two bit factor ($N \times c^2$) is represented in following two ways depending upon choice of conjugate basis pair.

$$\begin{aligned} Z \times (e_h Z^2 + e_l Z)^2 &= (e_h \oplus e_l) Z^2 + e_h Z. \\ Z^2 \times (e_h Z^2 + e_l Z)^2 &= e_h Z^2 + (e_h \oplus e_l) Z. \end{aligned} \tag{14}$$

The scaling operation ($n \times a^2$) is a four bit factor in $GF(2^8)$ inverter and its computation in $GF(2^2)$ depends on the normal basis types and the relation between norm n and N as in [9]. For computations in $GF(2^4)$, tables in appendix ‘B’ are used.

Fig. 4 $GF(2^2)$ multiplier



3.2 Generating Isomorphic and Inverse Mapping Functions

The standard SMS4 form is defined by 8 bit vector as coefficients of powers of x which is root of irreducible primitive polynomial in (3). Multiplicative inversion in composite field is computed after a byte in $GF(2^8)$ is mapped to its composite field representation using isomorphism function δ [7]. After the multiplicative inverse is computed in the composite field, the 8 bit result is mapped back to standard equivalent representation in $GF(2^8)$ using inverse isomorphic function δ^{-1} . The isomorphic and its inverse mapping is one to one and onto mapping and is represented as 8×8 matrix [14]. If byte s is in standard polynomial basis then it can be represented as a quadratic extension as $s = a_h X^{16} + a_l X$, $a_h, a_l \in GF(2^4)$, where each 4 bit coefficient is represented as $c = c_h Y^4 + c_l Y$, $c_h, c_l \in GF(2^2)$, each of which is then further represented as pair of bits $e = e_h Z^2 + e_l Z$ in $GF(2^2)/GF(2)$. If the new byte is given as $t_7 t_6 t_5 t_4 t_3 t_2 t_1 t_0$ then we have the following expression [9].

$$\begin{aligned}
 & s_7 S^7 + s_6 S^6 + s_5 S^5 + s_4 S^4 + s_3 S^3 + s_2 S^2 + s_1 S^1 + s_0 S^0 \\
 &= \{ (t_7 Z^2 + t_6 Z) Y^4 + (t_5 Z^2 + t_4 Z) Y \} X^{16} + (t_3 Z^2 + t_2 Z) Y^4 + (t_1 Z^2 + t_0 Z) X, \\
 &= t_7 Z^2 Y^4 X^{16} + t_6 Z Y^4 X^{16} + t_5 Z^2 Y X^{16} + t_4 Z Y X^{16} + t_3 Z^2 Y^4 X \\
 &\quad + t_2 Z Y^4 X + t_1 Z^2 Y X + t_0 Z Y X.
 \end{aligned}
 \tag{15}$$

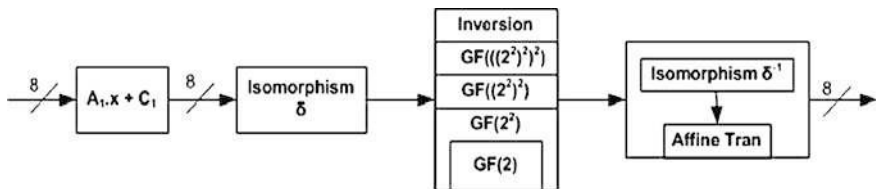


Fig. 5 SMS4 S-box block diagram

The values of X, Y and Z are substituted from the conjugate basis chosen and these 8 hexadecimal values with coefficient t_i represents the columns of 8×8 reverse base transformation matrix δ^{-1} . The inverse matrix δ is used for changing standard basis to corresponding composite field representation [9]. The inverse mapping matrix δ^{-1} is combined with affine transformation matrix A_2 for further optimization as in [7]. The block diagram of SMS4 S-box is given in the Fig. 5.

4 Results

For the possible choices of norms in $GF(2^4)$ and $GF(2^2)$ along with the normal basis at each subfield level satisfying $\tau = T = 1$, we have 16 possible cases as shown in appendix ‘A’. SMS4 S-box design based on each case is fully tested and simulated. The most compact case is the one which gives the least number of XOR gates for implementation. It can be observed from the results in Table 1 that choosing different normal basis combination results in small difference in number of XOR gates. These small differences exist due to different mapping matrices and slight differences in the inverter architectures. The matrices operations for mapping, inverse mapping and affine transformation are fully optimized using greedy algorithm [14]. The greedy algorithm operates iteratively on the mentioned matrices determining the occurrences of all possible repeating pairs in the output. The repeating pairs are pre-computed to reduce the number of XOR gates. Our best case S-box design (case 5, Table 1) saves 35 XOR gates by application of greedy algorithm.

The $GF(2^8)$ inverter in normal basis comprises of one $GF(2^4)$ inverter, three $GF(2^4)$ multipliers, one square scaling and two additions in $GF(2^4)$ as shown in Fig. 1. One $GF(2^4)$ inversion is computed using three multipliers, one inversion, one square scaling and two additions in $GF(2^2)$ as depicted in Fig. 2, where one $GF(2^4)$ multiplier comprises of three multipliers, four additions and a scaling operation in $GF(2^2)$ as in Fig. 3. Thus total number of logic gates computed in hierarchical structure of inverter for our best case S-box is 91 XOR and 36 AND. The structures of multipliers in Figs. 3 and 4 depicts that it requires summation of high and low halves of each input factor. If the same factor is shared by two different multipliers then share factor can save one subfield addition [9].

Table 1 All cases of SMS4 S-box design using Normal basis in $GF(((2^2)^2)^2)$

No	Conjugate ordered pair basis			Logic Gates S-box	
	(X^{16}, X)	(Y^4, Y)	(Z^2, Z)	XOR	AND
1	$(0 \times 98, 0 \times 99)$	$(0 \times 51, 0 \times 50)$	$(0 \times 5C, 0 \times 5D)$	137	36
2	$(0 \times 98, 0 \times 99)$	$(0 \times 0C, 0 \times 0D)$	$(0 \times 5C, 0 \times 5D)$	135	36
3	$(0 \times BF, 0 \times BE)$	$(0 \times 51, 0 \times 50)$	$(0 \times 5C, 0 \times 5D)$	135	36
4	$(0 \times BF, 0 \times BE)$	$(0 \times 0C, 0 \times 0D)$	$(0 \times 5C, 0 \times 5D)$	139	36
5	$(0 \times 94, 0 \times 95)$	$(0 \times 51, 0 \times 50)$	$(0 \times 5C, 0 \times 5D)$	134	36
6	$(0 \times 94, 0 \times 95)$	$(0 \times 0C, 0 \times 0D)$	$(0 \times 5C, 0 \times 5D)$	136	36
7	$(0 \times EF, 0 \times EE)$	$(0 \times 51, 0 \times 50)$	$(0 \times 5C, 0 \times 5D)$	138	36
8	$(0 \times EF, 0 \times EE)$	$(0 \times 0C, 0 \times 0D)$	$(0 \times 5C, 0 \times 5D)$	136	36
9	$(0 \times C5, 0 \times C4)$	$(0 \times 51, 0 \times 50)$	$(0 \times 5C, 0 \times 5D)$	136	36
10	$(0 \times C5, 0 \times C4)$	$(0 \times 0C, 0 \times 0D)$	$(0 \times 5C, 0 \times 5D)$	136	36
11	$(0 \times E3, 0 \times E2)$	$(0 \times 51, 0 \times 50)$	$(0 \times 5C, 0 \times 5D)$	139	36
12	$(0 \times E3, 0 \times E2)$	$(0 \times 0C, 0 \times 0D)$	$(0 \times 5C, 0 \times 5D)$	136	36
13	$(0 \times C9, 0 \times C8)$	$(0 \times 51, 0 \times 50)$	$(0 \times 5C, 0 \times 5D)$	138	36
14	$(0 \times C9, 0 \times C8)$	$(0 \times 0C, 0 \times 0D)$	$(0 \times 5C, 0 \times 5D)$	139	36
15	$(0 \times B3, 0 \times B2)$	$(0 \times 51, 0 \times 50)$	$(0 \times 5C, 0 \times 5D)$	137	36
16	$(0 \times B3, 0 \times B2)$	$(0 \times 0C, 0 \times 0D)$	$(0 \times 5C, 0 \times 5D)$	137	36

Table 2 Logic Gates for our Best case SMS4 S-box

Mathematical operation	XOR	AND
Affine Trans 1 ($x.A_1 + C_1$)	29	–
Map $GF(2^8) \rightarrow GF((2^2)^2)^2$	15	–
Map inv + Affine Trans 2	17	–
$GF(2^8)$ Inversion	73	36
Total	134	36

Thus, a four bit common factor in one $GF(2^4)$ multiplier can save five XOR gates and a two bit common factor in $GF(2^2)$ multiplier can save one XOR gate. In $GF(2^8)$ inverter in Fig. 1, all three $GF(2^4)$ multipliers have share factors i.e. Ω , a_h , a_l are all shared between respective two $GF(2^4)$ multipliers thus saving 15 XOR gates. Similarly in $GF(2^4)$ normal inverter we have Φ , c_h , c_l shared between respective two $GF(2^4)$ multipliers thus saving 3 XOR gates. In total $15 + 3 = 18$ XOR gates can be saved by the share factors in $GF(2^8)$ and $GF(2^4)$ normal inverters in hardware implementation. Thus total number of gates required for case 5 SMS4 S-box are 73 XOR and 36 AND gates (Table 2).

5 Comparative Analysis

Our most compact SMS4 S-box comprises of 134 XOR and 36 AND gates with conjugate pair basis $(0 \times 94, 0 \times 95)$, $(0 \times 51, 0 \times 50)$ and $(0 \times 5C, 0 \times 5D)$ respectively. We provide comparison of our most compact case 5 S-box design

Table 3 Logic Gates SMS4 S-box based on Polynomial Basis Inverter of [10]

Mathematical Operation	Instances	XOR	AND
Affine Trans 1 ($x.A_1 + C_1$)	1	29	–
Map $GF(2^8) \rightarrow GF(2^4)^2$	1	12	–
Map inv $GF(2^4)^2 \rightarrow GF(2^8)$	1	10	–
Map $GF(2^4) \rightarrow GF(2^2)^2$	1	3	–
Map inv $GF(2^2)^2 \rightarrow GF(2^4)$	1	2	–
Affine Trans 2 ($y.A_2 + C_2$).	1	29	–
$GF(2^4)$ Multiplier	3	45	48
$GF(2^4)$ Squaring	1	2	–
$GF(2^4)$ Scaling	1	1	–
$GF(2^4)$ Addition	2	8	–
$GF(2^2)$ Multiplier	3	9	15
$GF(2^2)$ Squaring	1	1	–
$GF(2^2)$ Scaling	1	1	–
$GF(2^2)$ Addition	2	4	–
$GF(2^2)$ Inverter	1	1	–
<i>Total</i>		157	63

with the one based on $GF(2^8)$ inversion algorithm proposed in [10] that uses polynomial basis. The operations in the subfield and the number of XOR and AND logic gates required to design SMS4 S-box based on [10] is given in Table 3. The matrices computations are optimized using greedy algorithm as in [5].

6 Conclusion and Future Work

In this paper we have proposed an improved design for SMS4 S-box based on the combinational logic with a low gate count. The proposed algorithm for computing SMS4 S-box function is based on composite field $GF(((2^2)^2)^2)$ and we have simulated all the possible cases of subfield combination depending upon the choice of normal basis, from which we have determined the best case. All the transformation matrices are optimized using greedy algorithm. We have proved that our best case S-box design results in much lower gate count and reduces the complexity by 15% XOR gates and 42% AND gates over the S-box based on the inversion algorithm of [10]. Our compact architecture of SMS4 S-box can save a significant amount of chip area in the hardware implementation of SMS4 in ASICs and it can be used for area constrained and demanding throughput SMS4 integrated circuits for applications ranging from smart cards to high speed processing units. The future work will concentrate on the ASIC implementation of the S-box, where our design can be further improved using the logic gate optimizations depending on specific CMOS standard library.

Appendix A: $GF(2^8)$ Representation for Sms4 S-Box

The Table A.1 gives the decimal, hexadecimal and binary values of the $GF(2^8)$ generated modulo irreducible primitive polynomial $f(x) = x^8 + x^7 + x^6 + x^5 + x^4 + x^2 + 1$. Let A be the root of $f(x)$ then the field generated with respective names of elements is as below.

Dec	Hex	Binary	θ^i	Name	Dec	Hex	Binary	θ^i	Name
0	00	00000000	-	0	39	27	00100111	θ^{187}	β^4
1	01	00000001	θ^0	1	40	28	00101000	θ^{16}	A^{16}
2	02	00000010	θ^1	A	41	29	00101001	θ^{104}	G^8
3	03	00000011	θ^{134}	G^{128}	42	2A	00101010	θ^{153}	γ^8
4	04	00000100	θ^2	A^2	43	2B	00101011	θ^{119}	β^8
5	05	00000101	θ^{13}	G	44	2C	00101100	θ^{176}	F^{16}
6	06	00000110	θ^{135}	H^{128}	45	2D	00101101	θ^{223}	q^{32}
7	07	00000111	θ^{76}	J^4	46	2E	00101110	θ^{169}	b^2
8	08	00001000	θ^3	B	47	2F	00101111	θ^{114}	d^{128}
9	09	00001001	θ^{210}	a^{16}	48	30	00110000	θ^{138}	K^{128}
10	0A	00001010	θ^{14}	D^2	49	31	00110001	θ^{250}	n
11	0B	00001011	θ^{174}	g^{16}	50	32	00110010	θ^{241}	m^2
12	0C	00001100	θ^{136}	α^8	51	33	00110011	θ^{160}	C^{32}
13	0D	00001101	θ^{34}	α^2	52	34	00110100	θ^{36}	E^4
14	0E	00001110	θ^{77}	b^{16}	53	35	00110101	θ^{82}	P^{16}
15	0F	00001111	θ^{147}	d^4	54	36	00110110	θ^{90}	a^2
16	10	00010000	θ^4	A^4	55	37	00110111	θ^{96}	B^{32}
17	11	00010001	θ^{26}	G^2	56	38	00111000	θ^{79}	k^{16}
18	12	00010010	θ^{211}	k^4	57	39	00111001	θ^{47}	j^{16}
19	13	00010011	θ^{203}	J^4	58	3A	00111010	θ^{54}	N^2
20	14	00010100	θ^{15}	H	59	3B	00111011	θ^{220}	e^{32}
21	15	00010101	θ^{152}	J^8	60	3C	00111100	θ^{149}	Q^{128}
22	16	00010110	θ^{175}	n^{16}	61	3D	00111101	θ^{50}	M^2
23	17	00010111	θ^{168}	K^8	62	3E	00111110	θ^{10}	C^2
24	18	00011000	θ^{137}	J^{128}	63	3F	00111111	θ^{31}	m^{32}
25	19	00011001	θ^{240}	H^{16}	64	40	01000000	θ^6	B^2
26	1A	00011010	θ^{35}	M^{32}	65	41	01000001	θ^{165}	a^{32}
27	1B	00011011	θ^{89}	Q^8	66	42	01000010	θ^{144}	E^{16}
28	1C	00011100	θ^{78}	d^{16}	67	43	01000011	θ^{73}	P^{64}
29	1D	00011101	θ^{53}	b^{64}	68	44	01000100	θ^{28}	D^4
30	1E	00011110	θ^{148}	P^4	69	45	01000101	θ^{93}	g^{32}
31	1F	00011111	θ^9	E	70	46	01000110	θ^{111}	l^{16}
32	20	00100000	θ^5	C	71	47	01000111	θ^{184}	L^8
33	21	00100001	θ^{143}	m^{16}	72	48	01001000	θ^{213}	g^2
34	22	00100010	θ^{27}	N	73	49	01001001	θ^{193}	D^{64}
35	23	00100011	θ^{110}	e^{16}	74	4A	01001010	θ^{58}	f^{64}
36	24	00100100	θ^{212}	b	75	4B	01001011	θ^{181}	c^2
37	25	00100101	θ^{57}	d^{64}	76	4C	01001100	θ^{205}	e^2

(continued)

(continued)

Dec	Hex	Binary	θ^i	Name	Dec	Hex	Binary	θ^i	Name
38	26	00100110	θ^{204}	γ^4	77	4D	01001101	θ^{99}	N^{32}
78	4E	01001110	θ^{188}	J^{64}	123	7B	01111011	θ^{238}	β
79	4F	01001111	θ^{61}	k^{64}	124	7C	01111100	θ^{11}	F
80	50	01010000	θ^{17}	α	125	7D	01111101	θ^{253}	q^2
81	51	01010001	θ^{68}	α^4	126	7E	01111110	θ^{32}	A^{32}
82	52	01010010	θ^{105}	a^8	127	7F	01111111	θ^{208}	G^{16}
83	53	01010011	θ^{129}	B^{128}	128	80	10000000	θ^7	D
84	54	01010100	θ^{154}	b^{32}	129	81	10000001	θ^{87}	g^8
85	55	01010101	θ^{39}	d^8	130	82	10000010	θ^{166}	b^8
86	56	01010110	θ^{120}	H^8	131	83	10000011	θ^{201}	d^2
87	57	01010111	θ^{196}	J^{64}	132	84	10000100	θ^{145}	M^{16}
88	58	01011000	θ^{177}	N^{16}	133	85	10000101	θ^{172}	Q^4
89	59	01011001	θ^{230}	e	134	86	10000110	θ^{74}	P^2
90	5A	01011010	θ^{224}	D^{32}	135	87	10000111	θ^{132}	E^{128}
91	5B	01011011	θ^{234}	g	136	88	10001000	θ^{29}	f^{32}
92	5C	01011100	θ^{170}	λ^2	137	89	10001001	θ^{218}	c
93	5D	01011101	θ^{85}	λ	138	8A	10001010	θ^{94}	j^{32}
94	5E	01011110	θ^{115}	e^{128}	139	8B	10001011	θ^{158}	k^{32}
95	5F	01011111	θ^{216}	N^8	140	8C	10001100	θ^{112}	D^{16}
96	60	01100000	θ^{139}	L^{128}	141	8D	10001101	θ^{117}	g^{128}
97	61	01100001	θ^{246}	l	142	8E	10001110	θ^{185}	e^{64}
98	62	01100010	θ^{251}	q^4	143	8F	10001111	θ^{108}	N^4
99	63	01100011	θ^{22}	F^2	144	90	10010000	θ^{214}	c^8
100	64	01100100	θ^{242}	j	145	91	10010001	θ^{232}	f
101	65	01100101	θ^{244}	k	146	92	10010010	θ^{194}	F^{64}
102	66	01100110	θ^{161}	G^{32}	147	93	10010011	θ^{127}	q^{128}
103	67	01100111	θ^{64}	A^{64}	148	94	10010100	θ^{59}	h^{64}
104	68	01101000	θ^{37}	P	149	95	10010101	θ^{179}	h^4
105	69	01101001	θ^{66}	E^{64}	150	96	10010110	θ^{182}	c^{64}
106	6A	01101010	θ^{83}	b^4	151	97	10010111	θ^{71}	f^8
107	6B	01101011	θ^{228}	d	152	98	10011000	θ^{206}	h^{16}
108	6C	01101100	θ^{91}	c^{32}	153	99	10011001	θ^{236}	h
109	6D	01101101	θ^{163}	f^4	154	9A	10011010	θ^{100}	M^4
110	6E	01101110	θ^{97}	F^{32}	155	9B	10011011	θ^{43}	Q
111	6F	01101111	θ^{191}	q^{64}	156	9C	10011100	θ^{189}	l^{64}
112	70	01110000	θ^{80}	C^{16}	157	9D	10011101	θ^{226}	L^{32}
113	71	01110001	θ^{248}	m	158	9E	10011110	θ^{62}	n^{64}
114	72	01110010	θ^{48}	B^{16}	159	9F	10011111	θ^{20}	C^4
115	73	01110011	θ^{45}	a	160	A0	10100000	θ^{18}	E^2
116	74	01110100	θ^{55}	e^8	161	A1	10100001	θ^{41}	P^8
117	75	01110101	θ^{141}	N^{128}	162	A2	10100010	θ^{69}	K^{64}
118	76	01110110	θ^{221}	β^2	163	A3	10100011	θ^{125}	n^{128}
119	77	01110111	θ^{102}	γ^2	164	A4	10100100	θ^{106}	b^{128}
120	78	01111000	θ^{150}	a^{128}	165	A5	10100101	θ^{156}	d^{32}
121	79	01111001	θ^{24}	B^8	166	A6	10100110	θ^{130}	C^{128}

(continued)

(continued)

Dec	Hex	Binary	θ^i	Name	Dec	Hex	Binary	θ^i	Name
122	7A	01111010	θ^{51}	γ	167	A7	10100111	θ^{199}	m^8
168	A8	10101000	θ^{155}	e^4	212	D4	11010100	θ^{84}	K^4
169	A9	10101001	θ^{198}	N^{64}	213	D5	11010101	θ^{215}	n^8
170	AA	10101010	θ^{40}	C^8	214	D6	11010110	θ^{229}	j^2
171	AB	10101011	θ^{124}	m^{128}	215	D7	11010111	θ^{233}	k^2
172	AC	10101100	θ^{121}	j^{128}	216	D8	11011000	θ^{92}	L^4
173	AD	10101101	θ^{122}	k^{128}	217	D9	11011001	θ^{183}	l^8
174	AE	10101110	θ^{197}	L^{64}	218	DA	11011010	θ^{164}	P^{32}
175	AF	10101111	θ^{123}	l^{128}	219	DB	11011011	θ^{72}	E^8
176	B0	10110000	θ^{178}	Q^{16}	220	DC	11011100	θ^{98}	J^{32}
177	B1	10110001	θ^{70}	M^{64}	221	DD	11011101	θ^{60}	H^4
178	B2	10110010	θ^{231}	p^8	222	DE	11011110	θ^{192}	B^{64}
179	B3	10110011	θ^{126}	p^{128}	223	DF	11011111	θ^{180}	a^4
180	B4	10110100	θ^{225}	H^{32}	224	E0	11100000	θ^{81}	K^{16}
181	B5	10110101	θ^{19}	J	225	E1	11100001	θ^{95}	n^{32}
182	B6	10110110	θ^{235}	n^4	226	E2	11100010	θ^{249}	p^2
183	B7	10110111	θ^{42}	K^2	227	E3	11100011	θ^{159}	p^{32}
184	B8	10111000	θ^{171}	g^4	228	E4	11100100	θ^{49}	J^{16}
185	B9	10111001	θ^{131}	D^{128}	229	E5	11100101	θ^{30}	H^2
186	BA	10111010	θ^{86}	Q^2	230	E6	11100110	θ^{46}	L^2
187	BB	10111011	θ^{200}	M^8	231	E7	11100111	θ^{219}	l^4
188	BC	10111100	θ^{116}	f^{128}	232	E8	11101000	θ^{56}	D^8
189	BD	10111101	θ^{107}	c^4	233	E9	11101001	θ^{186}	g^{64}
190	BE	10111110	θ^{217}	h^2	234	EA	11101010	θ^{142}	f^{16}
191	BF	10111111	θ^{157}	h^{32}	235	EB	11101011	θ^{109}	c^{128}
192	C0	11000000	θ^{140}	M^{128}	236	EC	11101100	θ^{222}	l^{32}
193	C1	11000001	θ^{101}	Q^{32}	237	ED	11101101	θ^{113}	L^{16}
194	C2	11000010	θ^{247}	q^8	238	EE	11101110	θ^{103}	h^8
195	C3	11000011	θ^{44}	F^4	239	EF	11101111	θ^{118}	h^{128}
196	C4	11000100	θ^{252}	p	240	F0	11110000	θ^{151}	j^8
197	C5	11000101	θ^{207}	p^{16}	241	F1	11110001	θ^{167}	k^8
198	C6	11000110	θ^{23}	L	242	F2	11110010	θ^{25}	M
199	C7	11000111	θ^{237}	l^2	243	F3	11110011	θ^{202}	Q^{64}
200	C8	11001000	θ^{243}	p^4	244	F4	11110100	θ^{52}	G^4
201	C9	11001001	θ^{63}	p^{64}	245	F5	11110101	θ^8	A^8
202	CA	11001010	θ^{245}	n^2	246	F6	11110110	θ^{239}	q^{16}
203	CB	11001011	θ^{21}	K	247	F7	11110111	θ^{88}	F^8
204	CC	11001100	θ^{162}	K^{32}	248	F8	11111000	θ^{12}	B^4
205	CD	11001101	θ^{190}	n^{64}	249	F9	11111001	θ^{75}	a^{64}
206	CE	11001110	θ^{65}	C^{64}	250	FA	11111010	θ^{254}	q
207	CF	11001111	θ^{227}	m^4	251	FB	11111011	θ^{133}	F^{128}
208	D0	11010000	θ^{38}	J^2	252	FC	11111100	θ^{33}	E^{32}
209	D1	11010001	θ^{195}	H^{64}	253	FD	11111101	θ^{146}	P^{128}
210	D2	11010010	θ^{67}	G^{64}	254	FE	11111110	θ^{209}	f^2
211	D3	11010011	θ^{128}	A^{128}	255	FF	11111111	θ^{173}	c^{16}

The minimal polynomials over GF(2) and their respective conjugate roots in terms of θ^i are presented in the following Table A.2.

Name	Minimal polynomial	Conjugate roots (θ^i)
l	$x + 1$	θ^0
λ	$x^2 + x + 1$	$\theta^{85}, \theta^{170}$
α	$x^4 + x + 1$	$\theta^{17}, \theta^{34}, \theta^{68}, \theta^{136}$
β	$x^4 + x^3 + 1$	$\theta^{238}, \theta^{221}, \theta^{187}, \theta^{119}$
γ	$x^4 + x^3 + x^2 + x + 1$	$\theta^{51}, \theta^{102}, \theta^{204}, \theta^{153}$
A	$x^8 + x^7 + x^6 + x^5 + x^4 + x^2 + 1$	$\theta^1, \theta^2, \theta^4, \theta^8, \theta^{16}, \theta^{32}, \theta^{64}, \theta^{128}$
B	$x^8 + x^7 + x^5 + x^4 + x^3 + x^2 + 1$	$\theta^3, \theta^6, \theta^{12}, \theta^{24}, \theta^{48}, \theta^{96}, \theta^{192}, \theta^{129}$
C	$x^8 + x^4 + x^3 + x + 1$	$\theta^5, \theta^{10}, \theta^{20}, \theta^{40}, \theta^{80}, \theta^{160}, \theta^{65}, \theta^{130}$
D	$x^8 + x^6 + x^5 + x^4 + 1$	$\theta^7, \theta^{14}, \theta^{28}, \theta^{56}, \theta^{112}, \theta^{224}, \theta^{193}, \theta^{131}$
E	$x^8 + x^5 + x^4 + x^3 + x^2 + x + 1$	$\theta^9, \theta^{18}, \theta^{36}, \theta^{72}, \theta^{144}, \theta^{33}, \theta^{66}, \theta^{132}$
F	$x^8 + x^6 + x^3 + x^2 + 1$	$\theta^{11}, \theta^{22}, \theta^{44}, \theta^{88}, \theta^{176}, \theta^{97}, \theta^{194}, \theta^{133}$
G	$x^8 + x^7 + x^3 + x^2 + 1$	$\theta^{13}, \theta^{26}, \theta^{52}, \theta^{104}, \theta^{208}, \theta^{161}, \theta^{67}, \theta^{134}$
H	$x^8 + x^5 + x^4 + x^3 + 1$	$\theta^{15}, \theta^{30}, \theta^{60}, \theta^{120}, \theta^{240}, \theta^{225}, \theta^{195}, \theta^{135}$
J	$x^8 + x^5 + x^3 + x^2 + 1$	$\theta^{19}, \theta^{38}, \theta^{76}, \theta^{152}, \theta^{49}, \theta^{98}, \theta^{196}, \theta^{137}$
K	$x^8 + x^7 + x^6 + x^4 + x^3 + x^2 + 1$	$\theta^{21}, \theta^{42}, \theta^{84}, \theta^{168}, \theta^{81}, \theta^{162}, \theta^{69}, \theta^{138}$
L	$x^8 + x^7 + x^2 + x + 1$	$\theta^{23}, \theta^{46}, \theta^{92}, \theta^{184}, \theta^{113}, \theta^{226}, \theta^{197}, \theta^{139}$
M	$x^8 + x^7 + x^4 + x^3 + x^2 + 1$	$\theta^{25}, \theta^{50}, \theta^{100}, \theta^{200}, \theta^{145}, \theta^{35}, \theta^{70}, \theta^{140}$
N	$x^8 + x^7 + x^3 + x + 1$	$\theta^{27}, \theta^{54}, \theta^{108}, \theta^{216}, \theta^{177}, \theta^{99}, \theta^{198}, \theta^{141}$
P	$x^8 + x^5 + x^3 + x + 1$	$\theta^{37}, \theta^{74}, \theta^{148}, \theta^{41}, \theta^{82}, \theta^{164}, \theta^{73}, \theta^{146}$
Q	$x^8 + x^7 + x^6 + x^5 + x^2 + x + 1$	$\theta^{43}, \theta^{86}, \theta^{172}, \theta^{89}, \theta^{178}, \theta^{101}, \theta^{202}, \theta^{149}$
a	$x^8 + x^7 + x^6 + x^4 + x^2 + x + 1$	$\theta^{45}, \theta^{90}, \theta^{180}, \theta^{105}, \theta^{210}, \theta^{165}, \theta^{75}, \theta^{150}$
b	$x^8 + x^7 + x^6 + x^3 + x^2 + x + 1$	$\theta^{212}, \theta^{169}, \theta^{83}, \theta^{166}, \theta^{77}, \theta^{154}, \theta^{53}, \theta^{106}$
c	$x^8 + x^7 + x^5 + x^3 + 1$	$\theta^{218}, \theta^{181}, \theta^{107}, \theta^{214}, \theta^{173}, \theta^{91}, \theta^{182}, \theta^{109}$
d	$x^8 + x^7 + x^5 + x + 1$	$\theta^{228}, \theta^{201}, \theta^{147}, \theta^{39}, \theta^{78}, \theta^{156}, \theta^{57}, \theta^{114}$
e	$x^8 + x^7 + x^6 + x^5 + x^4 + x + 1$	$\theta^{230}, \theta^{205}, \theta^{155}, \theta^{55}, \theta^{110}, \theta^{220}, \theta^{185}, \theta^{115}$
f	$x^8 + x^7 + x^6 + x + 1$	$\theta^{232}, \theta^{209}, \theta^{163}, \theta^{71}, \theta^{142}, \theta^{29}, \theta^{58}, \theta^{116}$
g	$x^8 + x^6 + x^5 + x^4 + x^2 + x + 1$	$\theta^{234}, \theta^{213}, \theta^{171}, \theta^{87}, \theta^{174}, \theta^{93}, \theta^{186}, \theta^{117}$
h	$x^8 + x^6 + x^5 + x^3 + 1$	$\theta^{236}, \theta^{217}, \theta^{179}, \theta^{103}, \theta^{206}, \theta^{157}, \theta^{59}, \theta^{118}$
j	$x^8 + x^6 + x^5 + x + 1$	$\theta^{242}, \theta^{229}, \theta^{203}, \theta^{151}, \theta^{47}, \theta^{94}, \theta^{188}, \theta^{121}$
k	$x^8 + x^6 + x^5 + x^2 + 1$	$\theta^{244}, \theta^{233}, \theta^{211}, \theta^{167}, \theta^{79}, \theta^{158}, \theta^{61}, \theta^{122}$
l	$x^8 + x^7 + x^6 + x^5 + x^4 + x^3 + 1$	$\theta^{246}, \theta^{237}, \theta^{219}, \theta^{183}, \theta^{111}, \theta^{222}, \theta^{189}, \theta^{123}$
m	$x^8 + x^4 + x^3 + x^2 + 1$	$\theta^{248}, \theta^{241}, \theta^{227}, \theta^{199}, \theta^{143}, \theta^{31}, \theta^{62}, \theta^{124}$
n	$x^8 + x^7 + x^5 + x^4 + 1$	$\theta^{250}, \theta^{245}, \theta^{235}, \theta^{215}, \theta^{175}, \theta^{95}, \theta^{190}, \theta^{125}$
p	$x^8 + x^6 + x^5 + x^4 + x^3 + x + 1$	$\theta^{252}, \theta^{249}, \theta^{243}, \theta^{231}, \theta^{207}, \theta^{159}, \theta^{63}, \theta^{126}$
q	$x^8 + x^6 + x^4 + x^3 + x^2 + x + 1$	$\theta^{254}, \theta^{253}, \theta^{251}, \theta^{247}, \theta^{239}, \theta^{223}, \theta^{191}, \theta^{127}$

Appendix B: Tables for $GF(2^4)$ Computations

The Table B.1 gives the decimal, hexadecimal and binary values of the $GF(2^4)$ generated modulo irreducible primitive polynomial $g(x) = x^4 + x + 1$. Let α be the root of $g(x)$ then the field generated with respective names of elements is as below:

Dec	Hex	ANF Ω^i	Bin Ω^i	Ω^i	Name
0	00	0	0000	–	0
1	01	x	0001	Ω^0	1
2	02	x^2	0010	Ω^1	α
3	03	$x + 1$	0011	Ω^4	α^4
4	04	x^2	0100	Ω^2	α^2
5	05	$x^2 + 1$	0101	Ω^8	α^8
6	06	$x^2 + x$	0110	Ω^5	λ
7	07	$x^2 + x + 1$	0111	Ω^{10}	λ^2
8	08	x^3	1000	Ω^3	γ
9	09	$x^3 + 1$	1001	Ω^{14}	β
10	0A	$x^3 + x$	1010	Ω^9	γ^8
11	0B	$x^3 + x + 1$	1011	Ω^7	β^8
12	0C	$x^3 + x^2$	1100	Ω^6	γ^2
13	0D	$x^3 + x^2 + 1$	1101	Ω^{13}	β^2
14	0E	$x^3 + x^2 + x$	1110	Ω^{11}	β^4
15	0F	$x^3 + x^2 + x + 1$	1111	Ω^{12}	γ^4

The Table B.2 below gives the minimal polynomials over $GF(2)$ and their respective conjugate roots in terms of Ω^i are presented using irreducible primitive polynomial $g(x) = x^4 + x + 1$.

Name	Minimal polynomial	Conjugate roots (θ^i)
1	$x + 1$	Ω^0
λ	$x^2 + x + 1$	Ω^5, Ω^{10}
α	$x^4 + x + 1$	$\Omega, \Omega^2, \Omega^4, \Omega^8$
β	$x^4 + x^3 + 1$	$\Omega^{14}, \Omega^{13}, \Omega^{11}, \Omega^7$
γ	$x^4 + x^3 + x^2 + x + 1$	$\Omega^3, \Omega^6, \Omega^{12}, \Omega^9$

The addition Table B.3 in $GF(16)$ using the naming convention in Table A.1 is given below.

\oplus	0	1	α	α^2	γ	α^4	λ	γ^2	β^8	α^8	γ^8	λ^2	β^4	γ^4	β^2	β
0	0	1	α	α^2	γ	α^4	λ	γ^2	β^8	α^8	γ^8	λ^2	β^4	γ^4	β^2	β
1	1	0	α^4	α^8	β	α	λ^2	β^2	γ^8	α^2	β^8	λ	γ^4	β^4	γ^2	γ
α	α	α^4	0	λ	γ^8	1	α^2	β^4	β	λ^2	γ	α^8	γ^2	β^2	γ^4	β^8
α^2	α^2	α^8	λ	0	γ^2	λ^2	α	γ	γ^4	1	β^4	α^4	γ^8	β^8	β	β^2
γ	γ	β	γ^8	γ^2	0	β^8	β^4	α^2	α^4	β^2	α	γ^4	λ	λ^2	α^8	1
α^4	α^4	α	1	λ^2	β^8	0	α^8	γ^4	γ	λ	β	α^2	β^2	γ^2	β^4	γ^8
λ	λ	λ^2	α^2	α	β^4	α^8	0	γ^8	β^2	α^4	γ^2	1	γ	β	β^8	γ^4
γ^2	γ^2	β^2	β^4	γ	α^2	γ^4	γ^8	0	λ^2	β	λ	β^8	α	α^4	1	α^8
β^8	β^8	γ^8	β	γ^4	α^4	γ	β^2	λ^2	0	β^4	1	γ^2	α^8	α^2	λ	α
α^8	α^8	α^2	λ^2	1	β^2	λ	α^4	β	β^4	0	γ^4	α	β^8	γ^8	γ	γ^2
γ^8	γ^8	β^8	γ	β^4	α	β	γ^2	λ	1	γ^4	0	β^2	α^2	α^8	λ^2	α^4
λ^2	λ^2	λ	α^8	α^4	γ^4	α^2	1	β^8	γ^2	α	β^2	0	β	γ	γ^8	β^4
β^4	β^4	γ^4	γ^2	γ^8	λ	β^2	γ	α	α^8	β^8	α^2	β	0	1	α^4	λ^2
γ^4	γ^4	β^4	β^2	β^8	λ^2	γ^2	β	α^4	α^2	γ^8	α^8	γ	1	0	α	λ
β^2	β^2	γ^2	γ^4	β	α^8	β^4	β^8	1	λ	γ	λ^2	γ^8	α^4	α	0	α^2
β	β	γ	β^8	β^2	1	γ^8	γ^4	α^8	α	γ^2	α^4	β^4	λ^2	λ	α^2	0

The multiplication Table B.4 in $GF(16)$ is given as below.

\otimes	0	1	α	α^2	γ	α^4	λ	γ^2	β^8	α^8	γ^8	λ^2	β^4	γ^4	β^2	β
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	1	α	α^2	γ	α^4	λ	γ^2	β^8	α^8	γ^8	λ^2	β^4	γ^4	β^2	β
α	0	α	α^2	γ	α^4	λ	γ^2	β^8	α^8	γ^8	λ^2	β^4	γ^4	β^2	β	1
α^2	0	α^2	γ	α^4	λ	γ^2	β^8	α^8	γ^8	λ^2	β^4	γ^4	β^2	β	1	α
γ	0	γ	α^4	λ	γ^2	β^8	α^8	γ^8	λ^2	β^4	γ^4	β^2	β	1	α	α^2
α^4	0	α^4	λ	γ^2	β^8	α^8	γ^8	λ^2	β^4	γ^4	β^2	β	1	α	α^2	γ
λ	0	λ	γ^2	β^8	α^8	γ^8	λ^2	β^4	γ^4	β^2	β	1	α	α^2	γ	α^4
γ^2	0	γ^2	β^8	α^8	γ^8	λ^2	β^4	γ^4	β^2	β	1	α	α^2	γ	α^4	λ
β^8	0	β^8	α^8	γ^8	λ^2	β^4	γ^4	β^2	β	1	α	α^2	γ	α^4	λ	γ^2
α^8	0	α^8	γ^8	λ^2	β^4	γ^4	β^2	β	1	α	α^2	γ	α^4	λ	γ^2	β^8
γ^8	0	γ^8	λ^2	β^4	γ^4	β^2	β	1	α	α^2	γ	α^4	λ	γ^2	β^8	α^8
λ^2	0	λ^2	β^4	γ^4	β^2	β	1	α	α^2	γ	α^4	λ	γ^2	β^8	α^8	γ^8
β^4	0	β^4	γ^4	β^2	β	1	α	α^2	γ	α^4	λ	γ^2	β^8	α^8	γ^8	λ^2
γ^4	0	γ^4	β^2	β	1	α	α^2	γ	α^4	λ	γ^2	β^8	α^8	γ^8	λ^2	β^4
β^2	0	β^2	β	1	α	α^2	γ	α^4	λ	γ^2	β^8	α^8	γ^8	λ^2	β^4	γ^4
β	0	β	1	α	α^2	γ	α^4	λ	γ^2	β^8	α^8	γ^8	λ^2	β^4	γ^4	β^2

References

1. Office of State Commercial Cipher Administration of China (2006) SMS4 cipher for WLAN products. <http://www.oscca.gov.cn/UpFile/200621016423197990.pdf>
2. Diffie W, Ledin G (2008) SMS4 encryption algorithm for wireless networks. Cryptology ePrint Archive, Report 2008/329 <http://eprint.iacr.org/>
3. Liu F, Ji W, Hu L, Ding J, Shuwang L, Pyshkin A, Weinmann RP (2007) Analysis of the SMS4 Block Cipher. In: ACISP, LNCS, vol 4586. Springer, Heidelberg, pp 158–170
4. Rijmen V (2000) Efficient implementation of the Rijndael S-box www.iaik.tugraz.at/RESEARCH/krypto/AES/old/~rijmen/rijndael/sbox.pdf
5. Wolkerstorfer J, Oswald E, Lamberger M (2002) An ASIC implementation of the AES Sboxes. In: CT-RSA, LNCS, vol 2271. Springer, Heidelberg, pp 67–78
6. Rudra A, Dubey P, Jutla C, Kumar V, Rao J, Rohatgi P (2001) Efficient Rijndael encryption implementation with composite field arithmetic. In: CHES 2001, LNCS, Springer, Heidelberg, pp 171–184
7. Satoh A, Morioka S, Takano K, Munetoh S (2001) A compact Rijndael hardware architecture with S-box optimization. In: ASIACRYPT 2001, LNCS, vol 2248. Springer, Heidelberg, pp 239–254
8. Mentens N, Batina L, Preneel B, Verbauwhede I (2005) A systematic evaluation of compact hardware implementations for the Rijndael S-box. In: CT-RSA, LNCS, vol 3376. Springer, Heidelberg, pp 323–333
9. Canright D (2004) A very compact Rijndael S-box. Technical Report NPS-MA-04-001. Naval Postgraduate School (September) <http://web.nps.navy.mil/~dcanrig/pub/NPS-MA-05-001.pdf>
10. Bai X, Xu Y, Guo L (2008) Securing SMS4 Cipher against differential power analysis and its VLSI implementation. In: ICCS
11. Erickson J, Ding J, Christensen C (2009) Algebraic cryptanalysis of SMS4: Grobner basis attack and SAT attack compared. In: ICISC
12. Lidl R, Niederreiter H (1986) Introduction to finite fields and their applications. Cambridge University Press, New York
13. Deschamps J, Sutter G, Imana J (2009) Hardware Implementation of Finite Field Arithmetic. McGraw-Hill Professional. ISBN: 978-0-07-154582-2
14. Paar C (1994) Efficient VLSI architectures for bit parallel computation in Galois fields. Ph.D thesis, Institute for Experimental Mathematics, University of Essen

Part VIII
Smartphone Applications and Services

iTextMM: Intelligent Text Input System for Myanmar Language on Android Smartphone

Nandar Pwint Oo and Ni Lar Thein

Abstract In recent years, there are huge developments in mobile phone communication technology to 3G. 3G mobile phone in present form has offered to use internet and interact with the computing system in users own language. However, the efficient input method for Myanmar Language that can be used in 3G mobile phones is still a biggest issue. Moreover, the service of character or word prediction text input system is provided for English and other languages. This paper tried to figure out the development of an innovative Myanmar syllable prediction text input system for Android touch screen mobile phones that leverages structural information of Myanmar characters formation and statistical properties of lexicon resources. In iTextMM, by using position aware rule based matching algorithm and bigram model that achieved a desirable inputting performance compared with currently used prevalent mobile Myanmar input method on Android touch phone (MyanDroid). iTextMM has been released to public via Android Market and is currently in use by hundreds of native Myanmar Android smart phone users. An evaluation results show that the proposed method outperforms the conventional Myanmar text inputting method, approximately 50% in inputting performance.

Keywords Touch screen • Virtual keyboard • Smart phone

N. P. Oo (✉) · N. L. Thein
University of Computer Studies, Yangon, Myanmar
e-mail: nandarpwintoo@gmail.com

N. L. Thein
e-mail: nilarthein@gmail.com

1 Introduction

Today, mobile phone trends have changed to smart phone dramatically. As a result, the early day multi-tap's difficulty of Multi Key Stroke per Character (KSPC) can reduce to one key per character [1]. Meanwhile, touch screen keyboards utilize an on-screen virtual keyboard that is software-based. So, they can adapt easily for effective input. There are other ways to reduce KSPC efficiently such as key mapping and keypad Layout [2]. However, to have more efficient KSPC, this virtual keyboard needs to be embedded with some character prediction, word prediction, Part of Speech (POS) prediction, multimodal feedback, word completion and auto-correction techniques. There are many word prediction IME for different languages such English (LatinIME), Japanese, Greek (GreekIME), Chinese (PinyinIME), etc.

In spite of the advancement in ICT, there is no word prediction or syllable prediction input method for Myanmar language yet on today smart phone. The attempt of this paper is to outcome a syllable prediction input method for Myanmar language on Android smart phones. This system leverages the structural information of Myanmar character formation and syllables prediction mechanism, with which mobile phone users can input Myanmar text easier and faster. The writing order of Myanmar language (phonetic based scripts) is left to right and space do not use between words. In addition, Myanmar language has more alphabet than English language. Currently, Myanmar input methods on Multi-tap keypad phone are romanized input [11] and positional mapping [3]. Each method tries to tackle the input issue in totally different aspects. For example, romanized input requires to type equivalent pronunciation in English [4]. Romanization requires more keystrokes or taps because of significant differences of Myanmar language writing style from Latin based scripts. Key mapping (positional mapping), another input method, accepts keys in positional mapping according to hand writing order of the Myanmar language.

This paper proposes a new input method by predicting Myanmar syllables in candidate view using both character level prediction and syllable level prediction mechanism. The organization of the proposed system is as follows: Related work of the proposed system is described in Sect. 2. After that, Sect. 3 is the place of the architecture of the proposed system. Experimental results are discussed in Sect. 4. The last section is devoted to conclusion and further extension.

2 Related Work

Yao Xia-xia [2] proposed a realization of Chinese input method on Android with C language (also known as native language) rather than Java language that can reduce the usage of resources and power consumption, and accelerated response time. Shtinji Suematsu [1] introduced the idea of changing the predicted candidate

word by estimating the context of users according to the position information from Global Positioning System (GPS). The idea of predicting words as a context aware is acceptable. But, the context of predicted word is not only demanded on location of users. It also depends on the mobile phone users typing usage pattern over time.

Ye Kyaw Thu [5] proposed Myanmar Language SMS text entry system for Multi-tap keypad phone with the idea of consonant clustering prediction. Moreover, for Myanmar Language, Positional Mapping [3] idea is proposed for small computing devices such as mobile phones. However, all of these methods are relevant for Multi-tap keypad phone and the innovative ways to input Myanmar character on smart phone is still a challenging issue. Jianwei [6] presented hybrid Chinese input method for touch screen mobile phones that leverages hieroglyphic properties of Chinese characters to enable faster and easier input of Chinese character on mobile phone. Ahmet Cuneyd Tantug [7] used n-gram probabilistic and K best Viterbi decoding to generate a list of predictions for Multi-tap keypad phone. According to the lecture review, prediction text entry system for Myanmar Language hasn't tried out. The proposed system is the innovative input method on touch screen mobile phone that combined the Position Aware Matching Algorithm and Statistical Probabilistic Language Model (Bigram model).

3 Architecture of the Proposed System

The design of the soft keyboard on mobile phone is not so easy because it depends on user experiences, ways to less control to user, size of screen and resources (memory, complexity), etc. Myanmar characters are more than English character. Thus, it is pressing problem to set the proper key arrangement with less user control. According to the experiment, the proper key arrangement according to the Language Model can also enhance Key Stroke. However, some prediction soft keyboard can reduce not only key searching time but also Key Stroke per Character. This paper proposed innovative Myanmar syllable prediction soft keyboard. The proposed method (iTextMM) runs with two prediction engines: character level prediction and syllable level prediction as depicted in Fig. 1. Whenever the user input key to form one syllable in Myanmar language, character level prediction engine predicts the user's desire syllable by using Position Aware Matching (PAM) Model. This model uses Myanmar syllable dictionary as a linguistic resource. After the user commits the preferred syllable from the candidate list, it can be assumed as one syllable in Myanmar language and the syllable level prediction engine predicts next syllable in the candidate list with the use of training corpus from linguistic resources.

In Myanmar language, to be one syllable, E-vowel [ေ], Medial [ေ့ ချ ငှ ဝ်] Upper vowel [ိ ဝိ ဝဲ], lower vowel [ု ဝူ ဝွ ဝ်], Anusvara [ံ], Lower Dot [ံ့] and Visarga [း] are written with a consonant.

For example, to enter a word “friend” in Myanmar language, the character composition is in the following order.

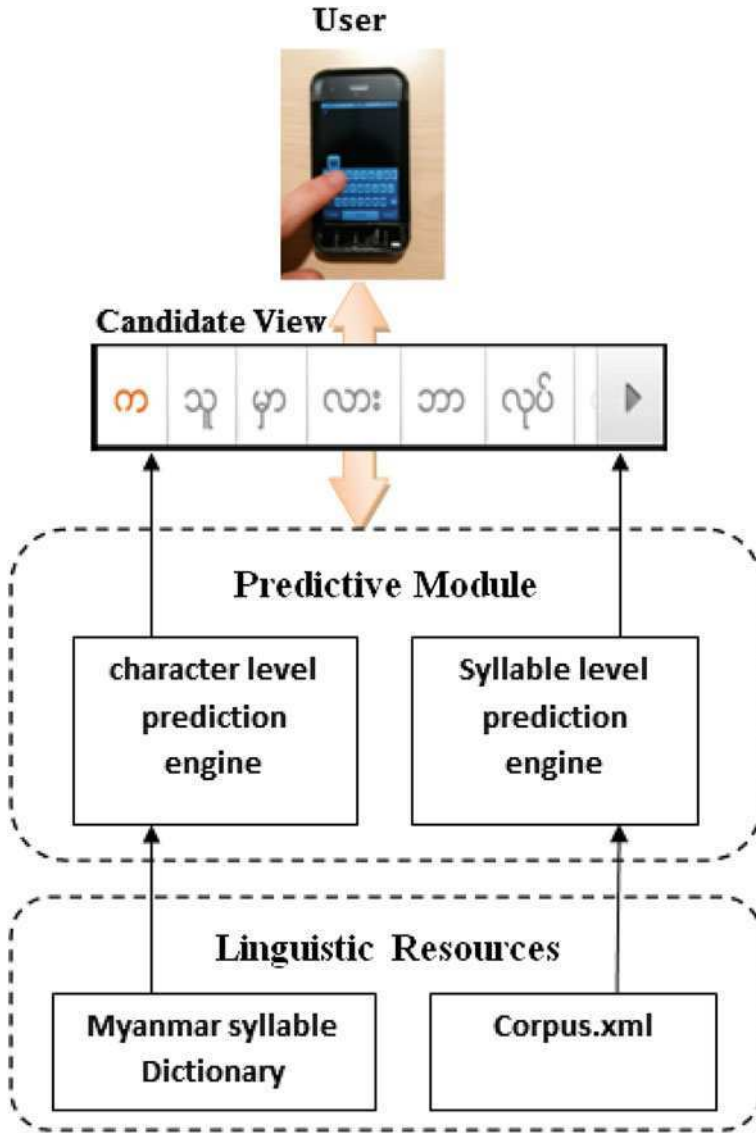


Fig. 1 Architecture of the Proposed System

သူ ငယ် ချင်း (tu nge chin; friend)

သ (consonant) + ဝ (lower vowel),

င (consonant) + ဝ (consonant) + င (medial) and

ခ (consonant) + င (medial) + င (consonant) + င (medial) + ဝ (Visarga)

Table 1 Myanmar character usage with position

1st position	2nd position	3rd position	4th position	5th position	6th position	7th position	8th position
Consonant	F1-F13	F1-F13	F1-F13	F1-F13	-	-	-
E-vowel	RM1-RM4	RM1-RM4	RM1	V	-	-	-
Left_Medial	U1-U3	U1-U3	U3	-	-	-	-
	D1-D4	D1-D4	D3-D2	-	-	-	-
	V	V	V	-	-	-	-
	A	A	A	A	A	A	A
	LM	-	-	-	-	-	-
	-	L	L	L	L	L	L

F = Final, RM = Right_Medial, U = Upper_Vowel, D = Down_Vowel, V = Visarga, A = Anusvara, LM = Left_Medial, L = Lower_Dot

As a result, typing Myanmar text to be one word is time consuming nature and to figure out the most effective way to input Myanmar text on today smartphone is the necessary task. After making an analysis of the Myanmar syllable formation, it can be seen that there are at most 8 positions in length to be one syllable. It is impossible to save all of Myanmar syllable with tree structure in mobile phone due to load excessive traversing time and memory usage. The proposed system figured out innovative Position Aware Rule Based Matching Algorithm to save memory consumption and to get quick responsiveness. According to Unicode 6.0, there are 72 characters for Myanmar script and the remaining are for other national dialect such as Mon, Sgaw Karen, etc. In the proposed system, to predict the next character only 49 characters are used and tagged with different names according to the position category.

When all Myanmar syllables are traversed according to positional level, their possible positions can be seen as shown in Table 1. According to the Table 1, the characters of the 2nd, 3rd and 4th are unpredictable.

E_vowel (ေ) is not considered in vowel combination because E_vowel is always at the first position. Similarly, the Myanmar dependent vowel such as က္က(e), က္ကိ(ee), ဥ(u), ဥိ(uu), ဓ(a), ဓြ(aw), ဓြိ(aww), ဓ်(hnite), ဓ်ိ(ei) and သ(tha gyi) are not concerned with the proposed algorithm because the dependent vowel are not combined with consonant to be one syllable in Myanmar language.

3.1 Position Aware Matching Algorithm

For character level prediction, instead of saving Myanmar syllable in tree structure or indexing with dictionary like LatinIME, to get quick responsiveness, the proposed system utilized character combination at run time according to algorithm 1. As shown in algorithm, the system accepts the current user touch key and combines it with corresponding Finals, Visarga and Annusvara, etc. according to the Position

information as shown in Table 1. Finally, the illegal combined syllables are reduced by analyzing that the combined syllables are contained in the syllable dictionary and give candidate list to user.

Algorithm-1 Position Aware Rule Based

Matching (Candidate_List)

```
{Assuming the inputs are current_key,buffer,
needed_vowel_group, syllable_dictionary}
begin
  buffer += current_user_pressed_key;
  repeat
    pos: = Check_Position(current_key);
    needed_vowel_group: = Load_vowel_group(pos);
    combined_word : = matching(current
_key,needed_vowel_group);
  if (syllable_dictionary.iscontained(combined_word))
    Candidate_List.add(combined_word);
  else
    Discard(combined_word);
  until (buffer.length() < 8 && !user_commit)
end.
```

3.2 Statistical Language Model

After the user has committed one syllable from candidate view, bi-gram prediction model takes the responsibility of the next syllable prediction. Assume $S = s_1, s_2, \dots, s_n$, where s_i denotes a word. Also, let s_{ij} denote $s_i, s_{i+1}, \dots, s_{j-1}, s_j$ and $P(X = x)$ is denoted as $P(x)$. In language model, a unit s occurs based on the sequence occurring just before. In k^{th} - order Markov models, given a sequence s_1, s_2, \dots, s_{i-1} the next word s_i is assumed to occur based on the limited length K of the previous history. If $K = 0$, the occurrence of the next syllable will not influence on the previous syllable. In this case, $P(S)$ is modeled as unigram model by Eq. (1).

$$P(S) \approx \prod_{i=1}^n P(s_i) \quad (1)$$

For syllable prediction, two-syllable possibilities of s_i and s_{i+1} is needed to consider. Therefore, the prediction model for next syllable is modeled as bigram model by Eq. (2).

$$P(s_n | s_{n-1}) = \frac{\text{count}(S_{n-1}, S_n)}{\text{count}(S_{n-1})} \quad (2)$$

In the proposed system, bigram predictor takes one previous syllable into account. It first looks for all lexicon syllables that match the existing syllable prefix, and then retrieves the bigram probability for each candidate syllable. In $P(s_i, s_j)$, s_i is the predicted syllable given syllable s_j . To save the response time, smoothing technique is not used to avoid the data sparseness problem. Moreover, to get quick responsiveness priority queue is used to sort the candidate syllable according to the probability ranking. In bigram calculation, linguistic resources (also known as corpus) play a very important role. In corpus building, real world messages sample are collected from six different age groups (10–20, 20–25, 25–30, 30–40, 40–50, Over 50) to avoid bias of sample sentences in one domain. Finally, the system can list the candidate words based on the input sequence. Otherwise, the users can also manually press a dedicated button on the layout to generate Myanmar sentences.

4 Performance Evaluation

The proposed system is developed with Java on Android platform (2.2 Froyo). Also, experiments with six native users (four males and two females) each at different age groups. Before making analysis, all learners are given 5 min demonstration time and 15 min practice time. Finally, the users were asked to type six sentences used in daily conversation between friends, and composed of most of the consonants, vowels, medials, anusvara, lower dot and visarga. Moreover, the users were requested to type 10 trials and the average time taken and key stroke are collected separately. The proposed system is evaluated regarding the number of Key Stroke per Character (KSPC), the time taken to type one Syllable per Minutes (SPM) and Error Rate (ER) including delete key and extra KeyStroke due to errors. These parameters are calculated as shown in the following Equations.

$$KSPC = \frac{c + ic + F}{N} \quad (3)$$

$$SPM = \frac{N}{T} \quad (4)$$

$$ER = \frac{E}{N} \quad (5)$$

C = Correct Key Stroke

IC = Incorrect Key Stroke

F = Key Stroke to fix typed errors

E = Total Numbers of Errors

T = Total Numbers of Time

N = Total Numbers of Characters

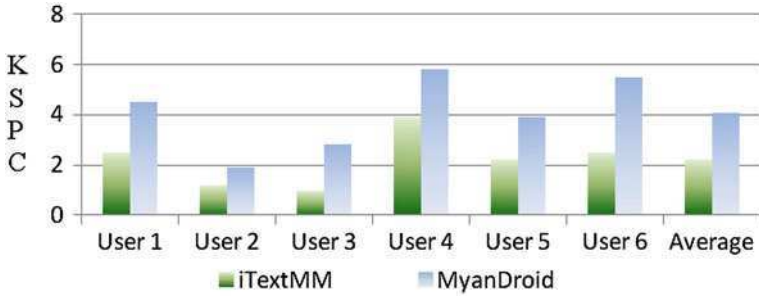


Fig. 2 KSPC comparison of iTextMM and MyanDroid

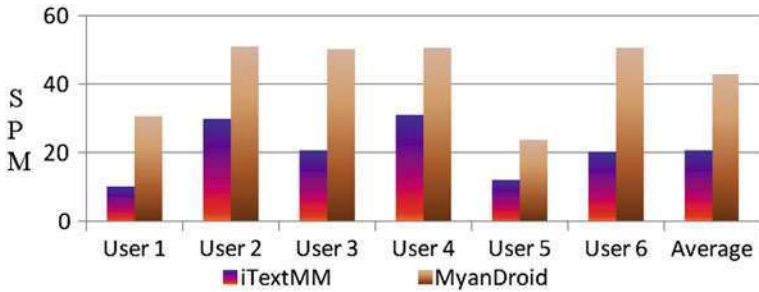


Fig. 3 SPM comparison of iTextMM and MyanDroid

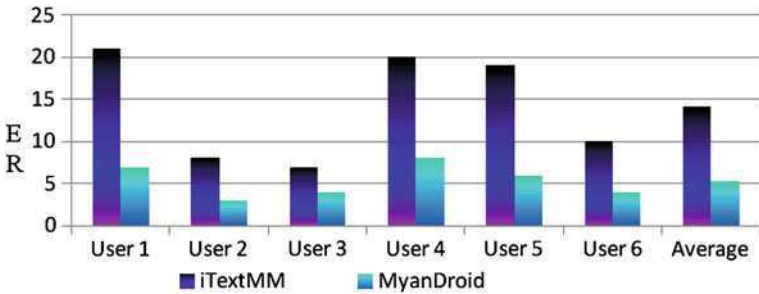


Fig. 4 ER comparison of iTextMM and MyanDroid

The analyzing impact of the proposed iTextMM related with the KSPC, ER and SPM are depicted in Figs. 2, 3 and 4 respectively. These results indicated that in all age level, the experienced iTextMM users can speed up inputting performance in speed and reduce Keystroke average 50%. Moreover, the participant were prefer to select their desire syllable in candidate view instead of typing all the character to input one syllable for Myanmar text and they all confirmed that the proposed iTextMM can reduce their key searching time. However, according to the merit of



Fig. 5 iTextMM allows a user to quickly input Myanmar text by syllable prediction

Error Rate (ER) the users cannot be benefited with the iTextMM method, because to fix the undesired syllable choice, it need more key press than MyanDroid.

5 Conclusion

iTextMM tried to figure out a syllable prediction text input method for Myanmar language on Android touch phone. This method leverages the structural characteristic of Myanmar syllable formation and statistical language model to help users input much faster and easier. Experiment result shows that iTextMM outperforms 50% in inputting performance than the currently used Myanmar text input methods on Android touch phone (MyanDroid). These input methods can be furnished with some personalization mechanism to enhance performance more and can applied across a variety of devices such as tables, virtual screen (Fig. 5).

References

1. Suematsu S, Arakawa Y, Tagashira S, Fukuda A (2010) Network-based context-aware input method editor. Sixth international conference on networking and services
2. Xia-sia Y, Yan-hui W, He-Jin (2010) An innovation of Chinese input based on android multimedia mobile device. First international conference on networking and distributed computing
3. Kyaw TY, Urano Y (2007) Positional mapping Myanmar text input scheme for mobile devices

4. Paek T, Chang K, Almog I, Badger E, Sengupta T (2010) A practical examination of multimodal feedback and guidance signals for mobile touchscreen keyboards. Methods for touch screen mobile phone, MobileHTC'10
5. Kyaw TY, Urano Y (2008) Positional prediction: consonant cluster prediction text entry method for Burmese (Myanmar Language), CHI Proceedings student research competition
6. Niu J, Zhu L, Yan Q, Liu Y, Wang K (2010) Stroke ++: a hybrid Chinese input method for touch screen mobile phones. MobileHTC'10
7. Tantug AC (2010) A probabilistic mobile text entry system for agglutinative languages. IEEE Trans Consumer Electron 56(2)
8. Aliparandi C, Carrnignani N, Mancarella P (2007) An inflected-sensitive letter and word prediction system. Int J Comput Inf Sci 5(2):79–85
9. Sharma MK, Dev S, Shah PK, Samanta D (2010) Parameters effecting the predictive virtual keyboard. In: Proceedings of the 2010 IEEE student's technology symposium

A Novel Technique for Composing Device Drivers for Sensors on Smart Devices

Deok hwan Gim, Seng hun Min and Chan gun Lee

Abstract Recently the techniques for reading accurate sensor values have emerged as an important issue. Identifying the context information of a mobile device by utilizing sensors enables new types of sensor-based applications ranging from games to scientific exploration software. In this article, we propose a new technique for composing device drivers for sensors on smart devices. We show how to increase the accuracy of a sensor by utilizing the correlation of other sensors. A systematic scheme for composing device drivers is defined.

Keywords Device driver · Sensor · Kalman filter · Smart device

1 Introduction

Recently the hardware capabilities of sensors have been much sophisticated. They are adopted in many areas including medical, military, and safety related applications. Even modern smart phones are equipped with various advanced sensors such as accelerometers and gyroscopes. Unfortunately, typical software of the

D. h. Gim · S. h. Min · C. g. Lee (✉)
Department of Computer Science, Chung-Ang University,
HeukSeok 221, Dongjak, Seoul 156-756, Korea
e-mail: cglee@cau.ac.kr
URL: <http://sites.google.com/site/rtselab>

D. h. Gim
e-mail: gimdeokhwan@gmail.com

S. h. Min
e-mail: sogomaster@gmail.com

smart phones controlling and reading such sensors are not near to the state of the art; hence the applications fail to fully exploit the advances of the sensors.

There have been many efforts for improving the performance of the sensors by correlating their outputs [1, 2]. However, most of them failed to consider the characteristics and software architecture of smart phones.

In this study, we propose a novel approach for the device driver for the sensors in smart phones. To put it simple, we present a systematic solution for realizing the sensor fusion method in the form of an integrated device-driver. We are mainly focused on the fusion of accelerometer and gyroscope sensors here, however, our proposed approach can be extended to other sensors easily.

Our proposal has the following advantages.

- Improving the accuracy of the sensors without modifying the applications using them
- Reducing the overhead for accessing sensors

The rest of our paper is composed as shown in the following. [Section 2](#) introduces related work on sensor fusion and device-driver development approaches. [Section 3](#) describes our approach, the challenges, and the solutions in detail. In [Sect. 4](#) we show the result of simulation and analysis. Finally, [Sect. 5](#) summarizes our approach and suggests future work.

2 Related Work

There have been active research toward improving the accuracy of the sensors by using multiple heterogeneous sensors. Kalman filter is one of the most well-known approaches in this field. In [1–3] they presented techniques for sensor fusion utilizing accelerometer and gyroscope sensors.

However, the utility of those work is limited to specific hardware platforms. For example, the solutions proposed in [2–4] were not general in that any changes of the hardware required re-design of the coefficients of the filter.

Moreover, we argue that it is not trivial to apply Kalman filter in the device-driver level. In order for the application use the filtered output from Kalman filter, we need to translate the output to one(s) that the application can understand. This will be further discussed in [Sect. 3](#).

3 Our Approach

In this paper, we propose a new technique for composing device drivers for sensors on smart devices. It should be noted that our approach applies Kalman filter to the device-driver level. In addition, the approach does not depend on a specific hardware platform, but can be applied to general cases.

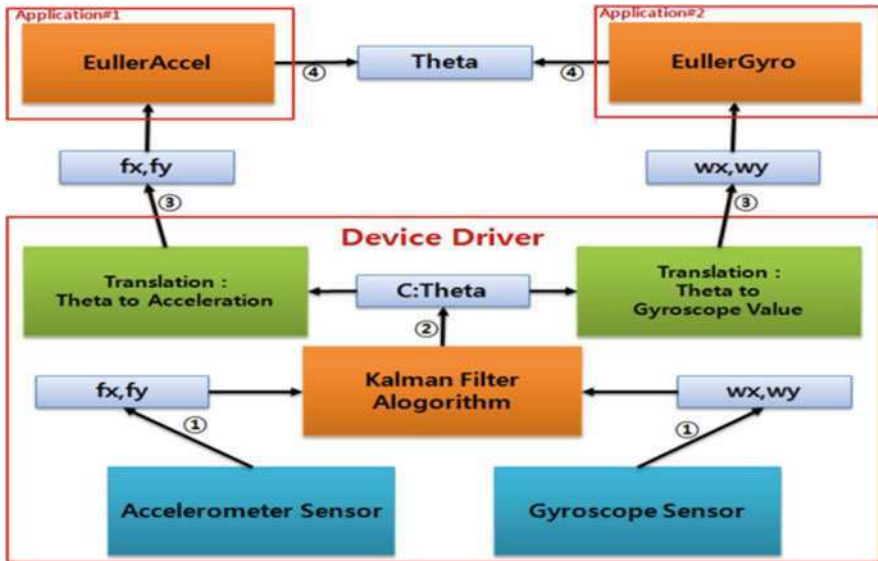


Fig. 1 Software architecture of integrated device driver

We propose an integrated device driver for multiple sensors instead of having separate independent device driver for each sensor. For the clarification, we shall consider a case where the system has two devices, accelerometer and gyroscope sensors, and present how our approach can be applied.

Figure 1 shows the internal software architecture of an integrated device-driver for the case. As shown in the figure, the integrated device driver consists of sensor controllers, a Kalman filter, and sensor-specific translators. Kalman filter takes the outputs from the two sensors and it calculates the orientation. Then each translator converts this orientation to value(s) in a format appropriate for the sensor.

In conventional schemes, an application can receive only one sensor data from a device driver at once; in case an application wants to improve the accuracy of received sensor data in the schemes, it should issue two requests to the device drivers. In our approach, the application issues a request, and a value with improved accuracy will be returned to the application.

Our approach provides the following advantages:

Firstly, legacy applications can receive the benefit of improved accuracy of the sensors without modifying them in the case of hardware upgrade. This is due to the fact that, the sensor fusion algorithm is built in the proposed device driver; hence the application can be ignorant of the other sensors required for the sensor fusion. For example, an App originally written for iPhone 3 GS would not attempt to use gyroscope sensor to improve the accuracy of accelerometer. When we install this App on iPhone 4, the App would get the benefit of the newly equipped gyroscope sensor without any modification in case our integrated driver is installed on the phone.

Secondly, we can expect the reduced run-time overhead for accessing multiple sensors. Typical operating systems have two operating modes, user-mode and kernel mode, and CPUs have corresponding modes. Most applications are operated in user mode, but when the system needs to control the hardware device, its operating mode is changed to kernel mode. Changing between user-mode and kernel mode incurs run-time overhead [5, 6]. Hence, by reducing the number of mode changes, the less run-time overhead can be expected. In addition, our approach can enjoy the faster execution times of IRQ and/or FIQ modes than user modes, which is a property of ARM processors adopted in many smart phones. Our approach runs the sensor fusion algorithm in IRQ and/or FIQ modes.¹

Our proposal needs the modeling of mathematical interdependency between sensors, hence if there is a configuration change of the sensors (i.e., addition or deletion of a sensor), then the another modelling should be done. Another limitation is that our integrated device driver should have a prior knowledge that how the application will use the sensors. For example, the device driver should know that which formats of the value, distance or orientation, is needed by the application for a given sensor output, such as accelerometer.

Fortunately, the first problem does not occur (or extremely rare) in consumer smart devices, where the hardware sensor changes are almost impossible after the purchase by the customers. The second problem can be overcome by analysing the usage pattern of sensor devices in applications. By identifying those patterns, we should be able to model the integrated device drivers.

4 Challenges and Solutions

While implementing the proposed idea mentioned in the previous section, we encountered the following technical difficulties.

1. Kalman filter should be applied to the level where the actual data transition occurs.
2. Accelerometers in typical smart phones do not have high performance, hence we cannot calculate angular value with only accelerometers.
3. At kernel mode, transcendental functions such as trigonometrical are not supported.

In order to solve the first problem, we adopted sensor-specific translator modules in our integrated device driver as shown in Fig. 1. In the figure, Accelerometer Sensor and Gyroscope Sensor indicate the sensor devices. Kalman filter algorithm finds the orientation (θ) from the output of the sensors. Translation modules translate the orientation into acceleration or gyroscope values. These values are f_x , f_y (acceleration values with direction of x, y-axis) or w_x , w_y

¹ Specifically we use WFI to enable the faster execution in IRQ and FIQ modes.

(gyroscope value with direction of x, y-axis). Hence, application modules such as EulerAccel and EulerGyro can receive these values. The application EulerAccel calculates the orientation value from the acceleration values. Similarly, EulerGyro calculates the orientation value from the gyroscope values.

We found that the accelerometers in typical smart phones are not capable of finding the orientation due to their low performance. In order to solve this issue, we exploited the characteristic of smart devices; the range of the movement of smart devices controlled by humans are quite limited compared to the cases of the air plane bodies. We modified the formula for calculating the orientation by using accelerometer as follows:

$$\begin{pmatrix} f_x \\ f_y \\ f_z \end{pmatrix} = \begin{pmatrix} u \\ v \\ w \end{pmatrix} + \begin{pmatrix} 0 & w & -v \\ -w & 0 & u \\ v & -u & 0 \end{pmatrix} \begin{pmatrix} \dot{u} \\ \dot{v} \\ \dot{w} \end{pmatrix} + g \begin{pmatrix} \sin \theta \\ \cos \theta \sin \varphi \\ \cos \theta \cos \varphi \end{pmatrix} \tag{1}$$

Where f_x, f_y, f_z are acceleration with direction of x-axis, y-axis and z-axis, respectively obtained from sensor. u, v, w are velocity of movement. $\dot{u}, \dot{v}, \dot{w}$ are acceleration of movement. θ, φ are euler angular values. g is the acceleration of gravity. In the formula, we need a navigation sensor with high performance, which are very expensive thus are not found in typical phones, in order to determine u, v, w and $\dot{u}, \dot{v}, \dot{w}$. As mentioned above, we observe that the range of the movement of smart devices controlled by humans are quite limited compared to the cases of the air plane bodies. It means that the velocity and direction of movement will not be greatly changed under normal circumstances.

Therefore we can assume that the velocity of device is nearly uniform. Firstly, we consider the case where the device is not moving at all. In this case, movement velocity and acceleration are zero.

$$\begin{cases} u = v = w = 0 \\ \dot{u} = \dot{v} = \dot{w} = 0 \end{cases} \tag{2}$$

Secondly, we address the case where the velocity of the movement is uniform.

$$\begin{cases} \dot{u} = \dot{v} = \dot{w} = 0 \\ p = q = r = 0 \end{cases} \tag{3}$$

In this condition, the movement accelerations are zero and angular velocity is also zero because the changes of movement are almost nothing. Now we can derive the Eq. 4 from the Eqs. 1, 2 and 3.

$$\varphi = \sin^{-1}(f_x/(g\cos\theta)), \theta = \sin^{-1}((-f_x)/g) \tag{4}$$

Thus, we solved the second challenge.

We solved the third challenge by applying the Maclaurin series. By doing so, we have the following advantages. Our system can handle the trade-off between the accuracy of sensor and run-time cost by modifying the number of terms in the

Table 1 Experiment results

Accuracy improvement techniques	sensor type	Average error	Maximum error
Proposed approach	Accelerometer	2.0479×10^{-16}	1.0658×10^{-14}
None	Accelerometer	5.1342	21.0584
Proposed approach	Gyroscope	0.6300	3.4252
None	Gyroscope	22.6693	48.7651

Maclaurin series. As mentioned earlier, because the movement of smart devices is practically limited we can reasonably reduce the number of terms in the series.

5 Simulation Results and Analysis

We implemented a simulation model realizing the case of two sensors, accelerometer and gyroscope sensors in Matlab. The number of sample sensor data² was 41500. Our simulation was done for the case where the errors of accelerometer are maximized. The application programs (EulerAccel and EulerGyro) did not have to be re written because our proposed integrated device driver was running.

We compared the results for the original case and our proposed scheme in Table 1. The error was defined by the difference between the ideal value and the output from the device driver. As shown in the result, the errors are greatly reduced by adoption of our approach.

For the case of accelerometer, our approach showed dramatic improvement of error rates both in average and maximum. Similarly, in the case of gyroscope, our approach overwhelmed the conventional solution. Most importantly, it should be noted that we did not have to modify anything of the application.

6 Conclusion and Future work

In this paper, we proposed a novel approach for the device driver for the sensors in smart phones. Our study presents a systematic solution for realizing the sensor fusion method in the form of an integrated device-driver. Our approach brings the advantages of improving the accuracy of the sensors without modifying legacy applications and reducing the run-time overhead for accessing the sensors. We showed the cases for sensor fusion of accelerometer and gyroscope sensors; however, our proposed approach is generic and it can be extended to other sensors easily.

For the future work, we are planning to apply the complementary separate bias kalman filter [1] to our integrated device driver for reducing the gyro bias error.

² Crossbow Corporation. Model: Nav420.

We also hope to find the sensor usage patterns in the mobile applications. The power consumption issue for sensor fusion can be another interesting future work.

Acknowledgments This work was supported by the National Research Foundation of Korea Grant funded by the Korean government (No. 20110013924) and a grant (CR070019M093174) from Seoul R&BD Program.

References

1. Eric F (1996) Inertial head-tracker sensor fusion by a Complimentary separate-bias kalman filter. In: Proceeding of the virtual reality annual international symposium
2. Moravec HP (1989) Sensor fusion in certainty grids for mobile robots, sensor devices and systems for robotics. Springer, New York, pp 253–276
3. Lee H-J, Jung S (2009) Gyro sensor drift compensation by Kalman filter to control a mobile inverted pendulum robot system. In: Proceeding of IEEE international conference on industrial technology
4. Chen X (2003) Modeling random gyro drift by time series neural networks and by traditional method. In: Proceeding of the international conference on neural networks and signal processing
5. David FM et al (2007) Context switch overheads for Linux on ARM platforms. In: Proceeding of the ACM workshop on experimental computer science
6. Liedtke J (1995) On micro-kernel construction. In: Proceeding of ACM symposium on operating systems principles

Various Artistic Effect Generation From Reference Image

Hochang Lee, Sang-Hyun Seo, Seung-Taek Ryoo
and Kyung-Hyun Yoon

Abstract Nowadays smart phone is widely converge and the growth of mobile and small device market is rapidly increasing. And many image based applications are developed for generating contents. In this paper, we propose a system for human-friendly image generation from user chosen reference images. In mobile application research, performance and waiting time is main problems. Because mobile OS is not fast as PC. So we modify previous texture transfer techniques considering waiting time and performance. For this, we modified from scan-line approach to random-access approach. Our proposed system can make various artistic results automatically. From this framework, we develop android operating system based smart phone application. This system can be extended to various imaging devices (IPTV, camera, stylish photo).

Keywords Mobile application · Non-photorealistic rendering; smart phone

H. Lee · K.-H. Yoon (✉)
School of Computer Science and Engineering,
Chung-Ang University, Heukseok-dong, Dongjak-gu, Seoul, Korea
e-mail: khyoon@cau.ac.kr

H. Lee
e-mail: Hclee0126@cglab.cau.ac.kr

S.-H. Seo
Bâtiment Nautibus, Université Claude Bernard Lyon 1,
43 bd du 11 novembre 1918, Villeurbanne Cedex, France
e-mail: shseo75@gmail.com

S.-T. Ryoo
School of Computer Engineering, Han-Shin University,
Yongsan-dong, Osan-si, Gyeonggi-do, Korea
e-mail: sryoo@hs.ac.kr

1 Introduction

In recent, various portable devices have been launched and improving their performance. Especially, the smart phone market is growing rapidly in 2010, and applications for various user styles have been developed. Also more active consumers who make a UCC in their own rights and upload it on their blog or twitter are increasing. Therefore, human-friendly image generation techniques can utilize the various portable devices which use images, such as camera or smart phone as shown Fig. 1.

Non-photorealistic rendering (NPR) [1, 2] is a part of the computer graphics area, which is a technique to generate human-friendly images. There are some attempts applying NPR techniques in a portable device application. However, it has not been tried widely because of limitation of portable device's performance.

In this paper, we propose an image generation technique which expresses many artistic effects based on various reference images (Fig. 2). Various techniques are studied about painting effects [3–5]. There are many approach for expressing artistic effect, such as stroke, kernel, per pixel. From this, we select pixel based approach [6, 7], because it is effective in time cost, so good in mobile environment. For this technique, we use extension techniques of Lee's directional texture transfer algorithm [8]. These techniques are effective in time and memory cost, and it is possible that they can express unlimited artistic styles based on a single framework. We improved Lee's techniques and make them to express interactively drawing effects. From this, we can overcome the limitation of waiting time. Also our architecture is constructed in parallel process, so it will be possible to use GPU in future.

Contributions of our paper are as follows. First, this technique is possible to express various artistic effects based on a user chosen reference image. It can be used as funny and stylish effect in a photo, and it will create synergy effects by applying applications using images. Second, we tested various mobile OS and analyzed performance results. This data will utilize many image processing researchers who use mobiles.

The rest of this paper is organized as follows. Section 2 presents the architecture of the proposed artistic image generation system. Section 3 describes the employed recommendation mechanism based on the free network and method to overcome waiting time. In Sect. 4, we show the experimental results and performances. Finally, conclusion and future research directions are given in Sect. 5.

2 Artistic Effect Generation from Reference Image

Lee's directional texture transfer techniques [2] can generate various artistic effects on the results based on a single framework. This method uses texture transfer techniques [3] and expresses more natural texture effects than previous



Fig. 1 Smart phone App related NPR (I-Phone): (Left ToonPaint, Right PaintMe)



Fig. 2 Overview of our system

texture transfer techniques by preserving object shape information. These techniques can express artistic effects such as Oil (Van Gogh style), pastel, pen watercolor and so on. It is a simple algorithm as other NPR techniques and easy to extend to other portable devices, because it is effective in time cost and use only small memory size. It is an optimal algorithm to use in portable client devices.

It generates the result by updating each pixel based on a reference image, so each pixel is selected based on scan-line order (from up-left). It use distance factors of average color and deviation between the current pixel and the candidate pixel to find the best texture from the candidate set. For deviation distance, they define L-shape neighborhood and consider this neighbor. L-shape neighbor means the pixels already done (Fig. 3).

However, scan line order approach has limitation of interactive visualization. Therefore, we reconstruct the algorithm for operating random-location order approach. We use pixels which are already done in square-shape neighborhood instead of L-shape neighbor (selected neighbor kernel). Also, we find the position of reference image which has similar pattern with current result area. Figure 4 show the algorithm for finding best suitable candidate pixel from random access approach (without L-shape neighbor).

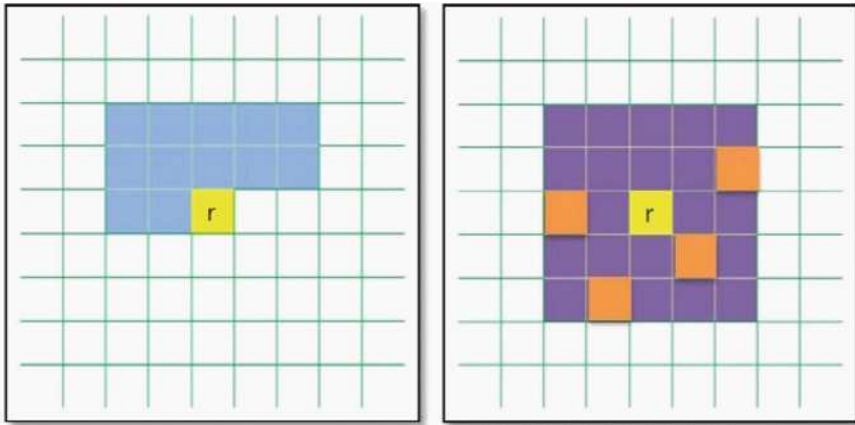


Fig. 3 left L-shape neighborhood and (right) selected neighbor kernel for random access position, blue and orange color pixel is already processed pixel

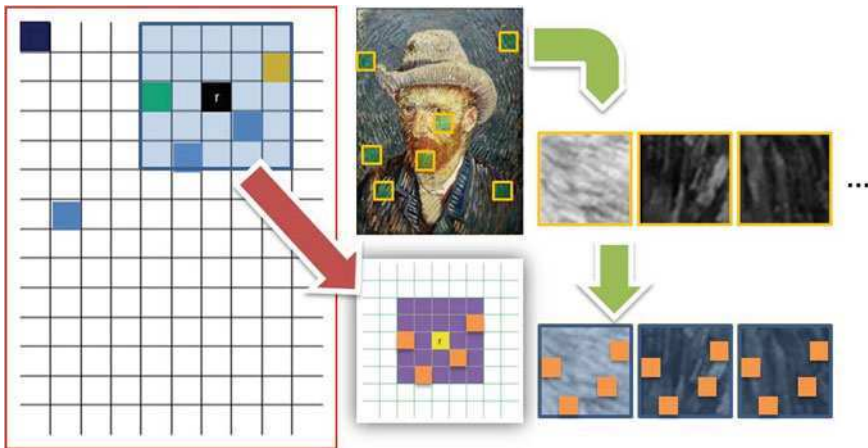


Fig. 4 Random access approach method. Current result (left) and kernel generation from current neighbor. We calculate deviation distance from selected neighbor kernel

3 Experimental Results and Application

Figure 5 shows the comparing results between the scan-line order and random access approach. From our new texture transfer techniques, we can express our result more interactively (Fig. 6). It has more visual effects than the scan-line order approach and it will be helpful to reduce the waiting time. Figure 7 shows the result applied in a smart phone. We implement in Android Operation system.

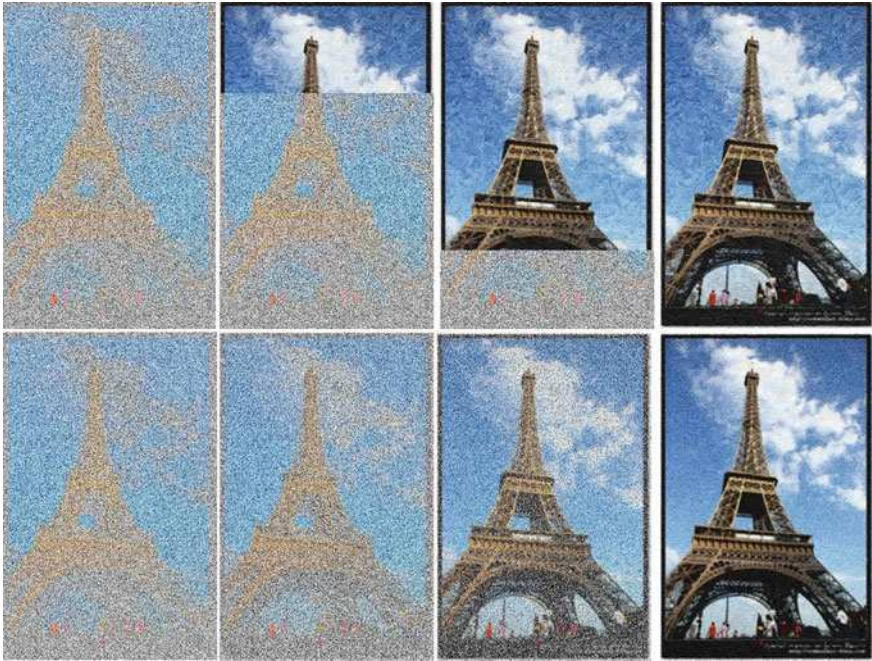


Fig. 5 Random access approach method. Current result (*left*) and kernel generation from current neighbor. We calculate deviation distance from selected neighbor kernel

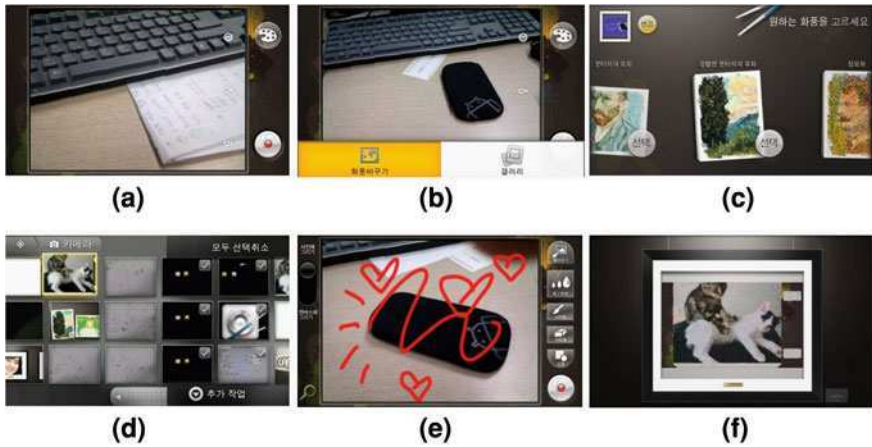


Fig. 6 Our application image (from *left top*) (a) camera view, (b) mode selection, (c) various style selection, (d) image selection from gallery, (e) user drawing mode, (f) result view



Fig. 7 Various result from our system

We construct DB from various real artist images. Currently, operating proposed techniques to the server, it takes 3–4 s in VGA image size. If we operate in the client device, it takes 1 min and half (1 GHz mobile CPU). Thus, we use QVGA image size and simpler algorithm than the server operator. In this case, it takes about 20–25 s. Because of using input photos and reference images, it spends only small memory. Finally, we select four representative style, such as tough oil, soft oil, pastel, pointillism.

Also we add another techniques using UI. Our application offer user other drawing mode by finger movement. We tested our system in other mobile OS such as I-phoneOS (apple) and Bada (Samsung).

4 Conclusion

In this paper, we propose a system for generating human-friendly image based on user chosen reference images. To use in a mobile device, we reconstruct Lee's texture transfer techniques and it is possible to show drawing steps interactively. To reduce the waiting time, we use a server if it is possible to use free network. Our techniques can express various artistic effects according to consumer's style.

Currently, we adapt our system to other smart phone operation and we expect that it will be used in diverse image devices.

Acknowledgments This research was supported by by the Korea Science and Engineering Foun-dation (KOSEF) grant funded by the Korea government (MEST) (20110018616). This work was also supported by Seoul R&BD Program (PA090920M093173)

References

1. Hertzmann A (2003) A survey of stroke-based rendering. *IEEE Comput Graph Appl* 23:70–81
2. Lee H, Lee CH, Yoon K (2009) Motion based painterly rendering. *Comput Graph Forum* 28(4):1207–1215
3. Hertzmann A (1998) Painterly rendering with curved brush strokes of multiple sizes. In: *Proceedings of ACM SIGGRAPH*, pp 453–460
4. Hays J, Essa I (2004) Image and video based painterly animation. In: *Proceedings of NPAR2004*, pp 113–120
5. Haeberli P (1990) Paint by numbers: abstract image representations. In: *Proceedings of SIGGRAPH*, pp 207–214
6. Ashikhmin M (2003) Fast texture transfer. *IEEE Comput Graph Appl* 23(4):38–43
7. Ashikhmin M (2001) Synthesizing natural textures. In: *ACM symposium on interactive 3D graphics*, pp 217–226
8. Lee H, Seo S, Yoon K (2011) Directional texture transfer with edge enhancement. *Comput Graph* 35(1):81–91

A Photomosaic Image Generation on Smartphone

Dongwann Kang, Sang-Hyun Seo, Seung-Taek Ryoo
and Kyung-Hyun Yoon

Abstract As the usage of mobile device, such as smartphone is becoming common, persons' interests in user created contents (UCC) are increasing gradually. Especially, the mobile devices combined with camera make that anyone can create UCC easily. In this paper, we introduce the implementation of smartphone application for converting an image which is taken by the phone into a non-photorealistic photomosaic image. Generally, photomosaic requires large database in order to create high quality result. Because the resource of mobile device is restricted, it is hard to store the large database of photomosaic in mobile device. We obtained the effect which increases the database by using the database which consists of rotatable images. We also offer a solution for the performance issue of best match search.

Keywords Photomosaic · Smartphone · Database · Best match search

D. Kang · K.-H. Yoon (✉)
School of Computer Science and Engineering, Chung-Ang University,
Heukseok-dong, Seoul, Dongjak-gu, Korea
e-mail: khyoon@cau.ac.kr

D. Kang
e-mail: dongwann@cglab.cau.ac.kr

S.-H. Seo
Bâtiment Nautibus, Université Claude Bernard, Lyon 1,
43, boulevard du 11 Novembre 1918, Villeurbanne Cedex, France
e-mail: shseo75@gmail.com

S.-T. Ryoo
School of Computer Engineering, Han-Shin University, Yangsan-dong,
Osan-si, Gyeonggi-do, Korea
e-mail: sryoo@hs.ac.kr

1 Introduction

As the mobile devices such as smartphone, PDA, PMP and etc. are developed, an environment that users can create contents directly is formed. The social network services such as YouTube (<http://www.youtube.com>), Flickr (<http://www.flickr.com>), and Facebook (<http://www.facebook.com>) which share the user-created contents (UCC) are popular. The commercial advertisements using the UCC are being given a great deal of weight on Internet. Currently, the UCC is just made by editing photo or video manually. Therefore, the method which assists the user to make more high level UCC is required in future. Non-photorealistic rendering techniques can be used for it.

The photomosaic which generates a target image using a lot of photographs as tiles is one of the most popular NPR techniques. As the smartphones with a built-in camera are being popular, snapshots are frequently taken by the user, so that snapshots are good material for making the UCC and are appropriate for the reference image of photomosaic. If the photomosaic is achieved on smartphone, then it will be enabled to make the UCC using photomosaic image which is generated using snapshots in the smartphone and share it on the social network. However, photomosaic requires large database, so that it is not desirable to store database into the smartphone with limited resource. In this paper, we propose a method that efficiently implements the photomosaic on smartphone. Using rotating photographs in database, we obtain an effect which is similar to use large database. In order to rotate photographs, we use rotatable object images.

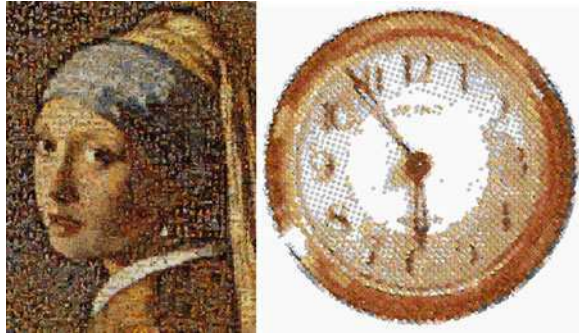
2 Related Work

General A photomosaic is a mosaic method that represents a source image using photo tiles. This technique proposed by Silvers [1] is composed of following simple steps: first, divide a source image into the blocks that tile will be attaches on; next, search the nearest neighbor image of each block from image database; and finally, replace the blocks with searched nearest neighbor images.

After proposed the photomosaic by Silvers, Finkelstein and Range [2] suggested a method that employ a hexagonal block division and apply a color correction. Kim and Pellacini [3] made a applied photomosaic method that represents a source image using the packing of arbitrary shaped image tiles. Besides the above, various photomosaic extensions such as video mosaics [4] or 3D mosaic [5] are developed.

Nicholas Tran [6] studied the performance of suggested the effectiveness for a measure of the performance of photomosaic. The effectiveness means the similarity between a source image and result image of photomosaic. Generally, in order to enhance the effectiveness of photomosaic a huge image-database is required. Because the existence probability of image more near to the block of

Fig. 1 Photomosaic (*left*) and stack mosaic (*right*)



source image becomes high as the size of database grows. As mentioned before, large database of photomosaic is not adequate for the smartphone which has limited resource. Therefore, to utilize the database efficiently stands out as the one of the most important issue of photomosaic algorithms (Fig. 1).

The stack mosaic [7] which is a variant of the photomosaic is the technique that represent target image by placing several tile layers. This technique is similar to the photomosaic in the fact that both of them make an image by composing several images. However stack mosaic places arbitrary shaped images which are rotated by various angles.

In general, the quality of the photomosaic result is in proportion to the size of image database. This means that the photomosaic requires a large database. On the other hand, stack mosaic obtains the effect that the size of database is dramatically decreased by using the rotated images in database. Because stack mosaic uses images of rotatable objects as tiles, there are holes between tiles. Therefore, stack mosaic places several layers of tiles in order to cover the hole. This is visually new style of photomosaic.

The Stack mosaic has a limitation that the cost of rotating tile layer and each tile is very high. This is not adequate for the environment of smart phone which has limited resource such as the size of memory, the clock speed of processor, and etc. In this paper, we present the method which overcomes above limitation on smartphone.

3 Algorithm

Our algorithm is divided into two steps: optimization of the database of object images and generation of result image on smartphone. In this section, we present the detail of each step.



Fig. 2 Transforms of image in database: original image, intensity transform, color transform, saturation transform, contrast transform

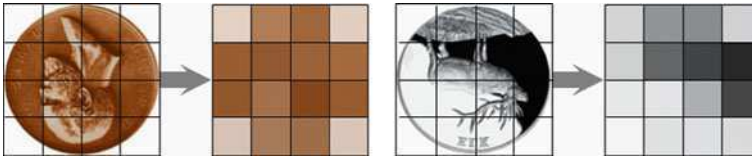
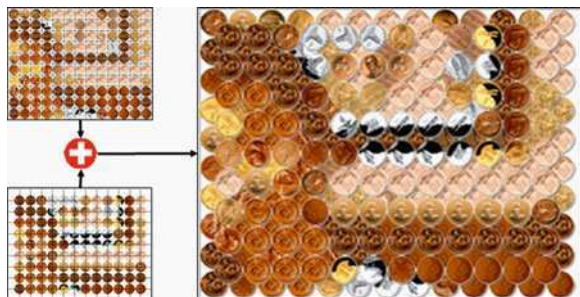


Fig. 3 The feature of each image in database: average color of each grid is stored in index

Fig. 4 Placement of tile layers



3.1 Optimization of Image Database

There are two considerations about image database of the stack mosaic. The first consideration is uneven distribution of images in database. Because our method uses limited size of database, the color that our database can express is very limited. We solve this problem by diversifying the color of image in database.

By applying the transform of the intensity, color, saturation, and contrast on each image, we generate additional images which slightly differ from original one. If the degree of the transform is excessively high, artificial images can be generated. Therefore we adjust the degree of transform appropriately (Fig. 2).

Second consideration is the size of image database. In general, the size of available memory of smartphone is relatively smaller than the PC. Therefore, the database should be compact in order to load into the smartphone.



Fig. 5 The results of our algorithm

In this paper, we employ the k -means clustering [8] in order to reconstruct the image database. This method aims to images in database into k clusters which consist of similar images. We take a representative image per each cluster, and construct the database using these images. Therefore, it is possible to control the size of database by adjusting the k .

For searching process of next step, we store the feature of each image of database into the index. We divide each image into several grids, and employ the average color of each grid as the feature (Fig. 3).

3.2 Generation of Stack Mosaic

In order to generate stack mosaic, each part of target image should be replaced by an image which is similar to the part. First of all, we divide target image into grids,

then search the best match image in the database. This process is similar to the photomosaic algorithm. We obtain the feature of each grid in the same way we present in Sect. 3.1, and calculate the L_2 distance between a feature of grid and each feature in index. Finally, we select the shortest distant image, and replace the part of target image with the selected image.

Unlike original stack mosaic [7] which has several tile layers, we employ only two layers. Although, to place several layers makes more visually pleasing result than ours, our method decreases the execution time. Moreover, to align grid with arbitrary angle is relatively expensive operation which causes low performance. We eliminate the process which aligns grids by placing upper layer to cover the hole of lower layer, so that obtain high performance (Fig. 4).

4 Results

We implemented our application on Android Gingerbread on Samsung Galaxy S. We used 800×600 target images and database which consists of 16×16 coin images. We set the value of k to 300. In these setting, the execution time of our application was 1–2 s. Figure 5 shows the results of our algorithm.

5 Conclusion

In this paper we proposed the stack mosaic method on smartphone. Due to the limited size of memory, we optimized the database using k -means clustering. We also suggested the method that places two tile layers cost-effectively. In conclusion, we implemented efficient stack mosaic application for smartphone.

Acknowledgments This work was supported by the Seoul R&BD Program (No. PA090920M093173) and by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MEST) (No. 20110018616).

References

1. Silvers R, Hawley M (1997) Photomosaics. Henry Holt, New York
2. Finkelstein A, Range M (1998) Image mosaics. In: RIDT1998, pp 11–22
3. Kim J, Pellacini F (2002) Jigsaw image mosaics. In: SIGGRAPH 2002, pp 657–664
4. Klein AW, Grant T, Finkelstein A, Cohen MF (2008) Video mosaics. In: NPAR2002, pp 21–28
5. Dos Passos V, Walter M (2008) 3D mosaics with variable-sized tiles. The visual computer, vol 24. Springer, New York, pp 617–623
6. Tran N (1999) Generating photomosaics: an empirical study. In: Symposium on applied computing 1999, pp 105–109

7. Park J, Yoon K, Ryoo S (2006) Multi-layered stack mosaic with rotatable objects. In Proceedings of computer graphics international 2006, pp 12–23
8. MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Fifth Berkeley symposium on mathematical statistics and probability 1967, pp 281–297

Author Index

A

A. A. Shahidan, 387
A. C. C. Lo, 407
Angel P. del Pobil, 421
Apichat Taweesiriwate, 13
Arief Marwanto, 407
Arnon Rungsawang, 13
Asokan Thondiyath, 455

B

Bakhta Meroufel, 43
Bhanu Shrestha, 127
Bonghwa Hong, 127
Bundit Manaskasemsak, 13
Byung-Chul Kim, 275
Byung-Jae Choi, 429

C

Chai-Jong Song, 359
Chan gun Lee, 519
Chang Seok Bae, 529, 537, 547, 557, 575,
583, 591, 597
Chang Sun Shin, 265, 295
Changbin Lee, 211
Chang-Heon Oh, 23
ChangSun Shin, 305
Chan-Gun Lee, 519
Chavinee Chaisri, 613
Cheol Sig Pho, 265
Chiang Lee, 87
Chul-Gyu Kang, 23
Chul-Sun Park, 23
ChulYoung Park, 285, 305

D

DaeHeon Park, 285, 305, 315
Daeyoung Kim, 565
Daisuke Kurabayashi, 377
Deok hwan Gim, 671
Dongho Won, 203
Dong-oh Kang, 603
Dongwann Kang, 687
Dongwon Han, 557

E

Edward Mattison, 367
Eunjeong Choi, 529
Eunil Park, 421

F

Farizah Yunus, 387

G

Ghalem Belalem, 43

H

Hae Young Lee, 77
Hakhyun Kim, 251
Hangbae Chang, 69
Heau-Jo Kang, 141
Hoang Huu Viet, 433
Hochang Lee, 679
Hochong Park, 351
HoSeong Cho, 285
Hwa-Young Jeong, 127

H (*cont.*)

Hyewon Song, 529
 Hyukjun Kwon, 119
 Hyung-Gik Lee, 547
 Hyungjik Lee, 603
 Hyun Ju Hwang, 433
 Haong Huu Viet, 433

I

I-Fang Su, 87
 Imran Abbasi, 641
 Ingeol Chun, 77
 In-Gon Park, 295
 Ishmael Makitla, 55

J

Jae Yong Lee, 157
 Jaeho An, 241
 Jaehwan Lim, 119
 Jaesung You, 251
 Jaewook Jung, 251
 Jae-Yong Lee, 265
 JangWoo Park, 285
 Jeeyeon Kim, 221
 Jeong Heon Kim, 485
 Jeun Woo Lee, 529
 Jeunwoo Lee, 157
 Ji Soo Park, 157
 Jin Myoung Kim, 77
 Jinho Yoo, 537
 Jin-Young Moon, 547
 Jong Hyuk Park, 149, 157, 165, 495
 Jonggu Kang, 69
 Jong-Jin Jung, 343
 Jongsung Kim, 621
 Joonyoung Jung, 565
 Jung-hoon Lee, 477
 Junghyun Nam, 203, 221
 Jung-Sik cho, 503
 Jung-Wan Ko, 503

K

K. M. Khairul Rashid, 407
 Kanad Ghose, 367
 Kang Ryoung Park, 359
 Keonsoo Lee, 173
 Ki Hong Kim, 3
 Ki Jung Yi, 157
 Kilhung Lee, 99, 109
 Ki-Seong Lee, 519
 Kwang Nam Choi, 485

Kwangwoo Lee, 211
 Kyo-Hoon Son, 275
 Kyoungyong Cho, 315
 Kyuchang Kang, 557, 575
 Kyung-Hack Seo, 351
 Kyung-Hyun Yoon, 679, 687
 Kyusuk Hann, 135

L

Li Yu, 189
 Long Chen, 333

M

M. Adib Sarijari, 407
 Mashhur Sattorov, 141
 Md Hasanuzzaman, 443
 Mehreen Afzal, 641
 Minh Nguyen, 189
 Minkoo Kim, 173
 Minwoo Cheon, 181
 Modar Safir Shbat, 397
 Moonsik Kang, 109
 Mucheol Kim, 495

N

N. Faisal, 387, 407
 N. Hija Mahalin, 407
 Namje Park, 211, 231
 Nandar Pwint Oo, 661
 Narong Mettripun, 630
 Naveen Chilamkurti, 621
 Neeraj Kumar, 621
 Nha Nguyen, 189
 Ni Lar Thein, 661
 Nor-Syahidatul N. Ismail, 387

P

Phuong Pham, 189
 Phyto Htet Kyaw, 467
 Piljae Kim, 377

R

Rozeha A. Rashid, 407

S

S. K. S. Yusof, 407
 Saeyoung Ahn, 333
 Sang Oh Park, 495

Sang-Hyun Seo, 679, 687
 Sang-Soo Yeo, 141, 511
 Se-Han Kim, 265
 Seng hun Min, 671
 Seok-Pil Lee, 343, 351, 359
 Seongsoo Cho, 127
 Seonguk Heo, 557, 575
 Seungjoo Kim, 251
 Seung-Min Park, 77
 Seungmin Rho, 173
 Seung-Taek Ryoo, 679
 Soo-Cheol Kim, 503, 511
 Soon Suck Jarng, 31
 Sung Kwon Kim, 503
 Sunshin An, 333

T

TaeChoong Chung, 433, 467
 Taegon Kim, 181
 Taek-Youn Youn, 631
 Taeshik Shon, 135
 Tetsunari Inamura, 443
 Thareswari Nagarajan, 455
 Thomas Fogwill, 55
 Thumrongrat Amornraksa, 613

V

Vyacheslav Tuzlukov, 397

W

Wei-Chang Yeh, 583
 Won-Tae Kim, 77
 Woongryul Jeon, 203

X

Xing Xiong, 429
 Xuanyou Lin, 87

Y

Yangsun Lee, 181
 Yong Yun Cho, 305
 Yongpil Park, 181
 Yongyun Cho, 315
 Young-Sik Jeong, 495
 Youngsook Lee, 203, 221
 Youngsub Na, 69
 Yu-Chi Chung, 87
 Yuk Ying Chung, 583