

Lecture Notes in Electrical Engineering 215

Kuinam J. Kim
Kyung-Yong Chung
Editors

IT Convergence and Security 2012

 Springer

Lecture Notes in Electrical Engineering

Volume 215

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

Kuinam J. Kim · Kyung-Yong Chung
Editors

IT Convergence and Security 2012

 Springer

Editors

Kuinam J. Kim
Convergence Security
Kyoung-gi University
Suwon, Gyeonggi-do
Republic of South Korea

Kyung-Yong Chung
Department of Computer Information
Engineering
Sangji University
Wonju-si, Gangwon-do
Republic of South Korea

ISSN 1876-1100

ISSN 1876-1119 (electronic)

ISBN 978-94-007-5859-9

ISBN 978-94-007-5860-5 (eBook)

DOI 10.1007/978-94-007-5860-5

Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2012953016

© Springer Science+Business Media Dordrecht 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

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

Contents

Part I Security Fundamentals

Development of an Automatic Document Malware Analysis System.	3
Hong-Koo Kang, Ji-Sang Kim, Byung-Ik Kim and Hyun-Cheol Jeong	
Study of Behavior-Based High Speed Visit/Inspection Technology to Detect Malicious Websites	13
Ji-Sang Kim, Hong-Koo Kang and Hyun-Cheol Jeong	
One-Way Hash Function Based on Cellular Automata	21
Jun-Cheol Jeon	
A Novel Malware Detection Framework Based on Innate Immunity and Danger Theory	29
Mohamed Ahmed Mohamed Ali and Mohd Aizaini Maarof	
A Forensic Evidence Collection Procedures of Smartphone in Crime Scene	35
Jeong-Hyun Lim, Chang-Woo Song, Kyung-Yong Chung, Ki-Wook Rim and Jung-Hyun Lee	
NAC System Analysis and Design for Improvement Model	43
Seung-Jae Yoo, Jeong-Mo Yang and Hwan-Seok Yang	

Part II Industrial and Business Information Security

Analysis of Methods for Detecting Compromised Nodes and Its Countermeasures	53
Fangming Zhao, Takashi Nishide, Yoshiaki Hori and Kouichi Sakurai	
Protecting Advertisers Against Click Frauds	61
Rattikorn Hewett and Abhishek Agarwal	
A Novel Intrusion Tolerant System Based on Adaptive Recovery Scheme (ARS)	71
Seondong Heo, Jungmin Lim, Minsoo Lee, Soojin Lee and Hyunsoo Yoon	
Design and Implementation of Linked Network Security System Based on Virtualization in the Separate Network Environment.	79
Dong-Hwi Lee and Kyong-Ho Choi	
A Study About Security Awareness Program Based on RFID Access Control System	87
Kyong-Ho Choi and DongHwi Lee	
A Framework for Anomaly Pattern Recognition in Electronic Financial Transaction Using Moving Average Method	93
Ae Chan Kim, Won Hyung Park and Dong Hoon Lee	
An Investigation on the Research Topics in Relation to Information Systems in Supply Chain Management	101
Younjung Kim, Young-Ho Lee, Kyung-Yong Chung and Kang-Dae Lee	
Designing a Model for Context Awareness Based Reliable Auction-Recommendng System (CARARS) by Utilizing Advanced Information	111
Hwa-Jin Park and Sang-Beom Kim	
 Part III Security Protocols and Applications	
A Secure Self-Encryption Scheme for Resource Limited Mobile Devices	121
Yongjoo Shin, Seungjae Shin, Minsoo Kim and Hyunsoo Yoon	

A Study on the Network Stability Measurement Based on an Articulation Nodes with Weight Value	131
Myung-Ki Jung and Seongjin Ahn	
An Improved Forward Secure Elliptic Curve Signcryption Key Management Scheme for Wireless Sensor Networks	141
Suman Bala, Gaurav Sharma and Anil K. Verma	
An Identity-Based Ring Signcryption Scheme	151
Gaurav Sharma, Suman Bala and Anil K. Verma	
A Secure DS-CDMA Technique with Capacity Enhancement for Ad Hoc Wireless Networks	159
Muhammad Zeeshan, Shoab Ahmed Khan and Muhammad Yasir Malik	
A Study on Traceback of Illegal Users Using Anonymity Technology in BitTorrent	169
Geon Il Heo, Nam Hun Kim, Ah Ra Jo, Sae In Choi and Won Hyung Park	
Internet Anonymity in Syria, Challenges and Solution	177
T. Eissa and Gi-hwan Cho	
 Part IV Convergence Security	
Comparison of Attacks and Security Analyses for Different RFID Protocols	189
Jung Tae Kim	
Honeypot Using Dynamic Allocation Technique with IP Scan	197
Hwan-Seok Yang	
Attribute-Based Encryption for Commercial Content Distribution	205
Hyoseung Kim, Seunghwan Park, Jong Hwan Park and Dong Hoon Lee	
 Part V IT Convergence Applications	
Sensibility Extraction for Bicycle Design Using RFID Tag-Attached Crayons	217
Ho-Il Jung, Seung-Jin Lee, Jeong-Hoon Kang, Min-Hyun Kim, Jong-Wan Kim, Bo-Hyun Lee, Eun-Young Cho and Kyung-Yong Chung	

Efficient Handover Schemes for Railroad Communications in 4G Mobile Networks	225
Ronny Yongho Kim and Baik Kim	
Cyber Threat Prediction Model Using Security Monitoring System Event	233
Neo Park and Won Hyung Park	
A Study for Car Insurance Service Using Vehicle Real Time Information	241
Yong-Yoon Shin and Byung-Yun Lee	
An MMU Virtualization for Embedded Systems.	247
Sung-Hoon Son	
The Design and Implementation of Start Stop System with Multi-Protocol in Automotive Smart Key System	253
Kyeong-Seob Kim, Yun-Sub Lee, In-Seong Song and Sang-Bang Choi	
The Design and Implementation of Improved Anti-Collision Algorithm for Vehicle User Authentication System.	261
Kyeong-Seob Kim, Yun-Sub Lee and Sang-Bang Choi	
Multi-Port Register File Design and Implementation for the SIMD Programmable Shader	267
Kyeong-Seob Kim, Yun-Sub Lee and Sang-Bang Choi	
Design Exploration Technique for Software Component Mapping of AUTOSAR Development Methodology	273
Kabsu Han and Jeonghun Cho	
Interoperability and Control Systems for Medical Cyber Physical Systems	283
Min-Woo Jung and Jeonghun Cho	
Ubiquitous Logistics Management in the Steel Industry	293
Sang-Young Lee and Yoon-Seok Lee	
Performance Evaluation of Train Propulsion Control on Ethernet Network Using TrueTime	301
Hyeon-Chyeol Hwang, Yong-Kuk Oh and Ronny Yongho Kim	

Intelligent Control System for Railway Level Crossing Safety 309
 Bong-Kwan Cho, Sang-Hwan Ryu, Hyeon-Chyeol Hwang,
 Seoung-Chon Koh and Do-Hyeon Lee

**Moving Average Estimator Least Mean Square Using
 Echo Cancellation Algorithm** 319
 Sang-Yeob Oh and Chan-Shik Ahn

**An Abrupt Signal Detection as Accident Detection by Hamiltonian
 Eigenvalue on Highway CCTV Traffic Signal** 325
 In Jeong Lee

**Implementation of Improved DPD Algorithm Using
 the Approximation Hessian Technique and an Adaptive Filter** 335
 Jeong-Seok Jang and Gyong-Hak Lee

**An InGaP HBT MMIC High Efficient Dual Path Power
 Amplifier for CDMA Handset Application** 343
 Song-Gang Kim, Hwan-Seok Yang and Seung-Jae Yoo

**U-Health Platform for Health Management Service
 Based on Home Health Gateway** 351
 Jong-Hun Kim, Si-Hoon Ahn, Jae-Young Soh and Kyung-Yong Chung

**A 2-D Visual Model for Sasang Constitution Classification
 Based on a Fuzzy Neural Network** 357
 Zhen-Xing Zhang, Xue-Wei Tian and Joon S. Lim

**Graph Coloring Algorithms and Applications to the Channel
 Assignment Problems** 363
 Surgwon Sohn

Thermal Performance in Ecological Korean House 371
 Jaewon Kim, Kyungeun Cho and Eunyoung Ahn

**An Automatic Roof Frame Design Method for Korean
 Traditional Wooden Architecture** 377
 Eunyoung Ahn and Noyoon Kwak

**Adaptive Load Partitioning Algorithm for Massively
 Multiplayer Online Games** 383
 Tae-Hyung Kim

Game-Based Learning System Using Graduated-Interval Recall Method	393
Ming Jin and Yoon Sang Kim	
An Efficient DRAM Converter for Non-Volatile Based Main Memory	401
Sung-In Jang, Cheong-Ghil Kim and Shin-Dug Kim	
An M2M-Based Interface Management Framework for Vehicles with Multiple Network Interfaces	409
Hong-Jong Jeong, Sungwon Lee, Dongkyun Kim and Yong-Geun Hong	
A Framework of the Wireless Sensor Based Railway Signal System.	417
Tarun Kumar, Ajay Chaudhary, Ganesh Singh and Richa Sharma	
Evolutionary Bio-Interaction Knowledge Accumulation for Smart Healthcare	425
Sung-Kwan Kang, Jong-Hun Kim, Kyung-Yong Chung, Joong-Kyung Ryu, Kee-Wook Rim and Jung-Hyun Lee	
A Non-Volatile Buffered Main Memory Using Phase-Change RAM	433
Do-Heon Lee, Chung-Pyo Hong and Shin-Dug Kim	
A Study on the Real-Time Location Tracking Systems Using Passive RFID.	441
Min-Su Kim, Dong-Hwi Lee and Kui-Nam J Kim	
SmartKeyboard for the Disabled in Smartwork	449
Juyoung Park, Seunghan Choi and Do-Young Kim	
Addressing the Out-of-date Problem for Efficient Load Balancing Algorithm in P2P Systems	459
Khaled Ragab and Moawia Elfaki Yahia	
Unified Performance Authoring Tool for Heterogeneous Devices and Protocols	481
Youngjae Kim, Seungwon Oh, Dongwook Lee, Haewook Choi, Jinsul Kim and Minsoo Hahn	
Discrimination System of Surround Illumination Using Photo Sensor's Output Voltage Ratio	489
Eun Su Kim, Hee Dong Park and Kyung Nam Park	

The Development of Korea: Computer Access Assessment System (K-CAAS) for Persons with Physical Disabilities	499
Jinsul Kim and Juhye Yook	
Color Coding for Massive Bicycle Trajectories	509
Dongwook Lee, Jinsul Kim, Haewook Choi and Minsoo Hahn	
User-Oriented Load Balancing Scheme for MMORPG	515
Hye-Young Kim	
Smart Desk: A Single Interface Based File Sharing System.	521
Naveed Ejaz, Sung Wook Baik and Ran Baik	
Feature Reduction and Noise Removal in SURF Framework for Efficient Object Recognition in Images	529
Naveed Ejaz, Ran Baik and Sung Wook Baik	
Automatic Segmentation of Region of Interests in MR Images Using Saliency Information and Active Contours	537
Ifan Mehmood, Ran Baik and Sung Wook Baik	
Digital Image Magnification Using Gaussian-Edge Directed Interpolation	545
Muhammad Sajjad, Ran Baik and Sung Wook Baik	
 Part VI Mobile Computing and Future Networks	
An Examination of Psychological Factors Affecting Drivers' Perceptions and Attitudes Toward Car Navigation Systems	555
Eunil Park, Ki Joon Kim and Angel P. del Pobil	
Handover Performance Evaluation of 4G Mobile Networks	563
Baik Kim and Ronny Yongho Kim	
An Emergency Message Broadcasting Using Traffic Hazard Prediction for Vehicle Safety Communications on a Highway	569
Sang Yeob Oh	
Detecting SIM Box Fraud Using Neural Network	575
Abdikarim Hussein Elmi, Subariah Ibrahim and Roselina Sallehuddin	

Overlapping-Node Removal Algorithm Considering the Sensing Coverage of the Node in WSN 583
Doo-Wan Lee, Chang-Joon Kim and Kyung-Sik Jang

Implementation of the Personal Healthcare Services on Automotive Environments 589
Kabsu Han and Jeonghun Cho

A Design of WSN Model to Minimize Data-Centric Routing Cost for Many-to-Many Communication 597
A. S. M. Sanwar Hosen and Gi-hwan Cho

Improving of Cache Performance using Trust Value of Nodes 605
Seung-Jae Yoo and Hwan-Seok Yang

Design of Microstrip Patch Antenna for Mobile Communication Systems Using Single-Feed 613
Chanhong Park

Microstrip Patch Antenna on UHF Band Using Multiple Meander for Metal Attached in U-City 623
Chanhong Park

Multi-Hop Relay Protocol Techniques for Performance Enhancement of LTE Downlink System 633
Chanhong Park

Performance Improvement Using Single Carrier-FDMA in Relay Based LTE Uplink System 645
Chanhong Park

An Efficient High-Speed Traffic Control Scheme for Real-Time Multimedia Applications in Wireless Networks 657
Moonsik Kang

Interference Assessment on the Circuit Domain in GSM-R Networks by Grey Clustering and Analytic Hierarchy Process 667
Si-Ze Li, Zhang-Dui Zhong, Yuan-Yuan Shi, Bo Ai, Jian-Wen Ding and Si-Yu Lin

Design of Transducer Interface Agent and its Protocol for WSN Middleware 677
Surgwon Sohn

Collaboration of Thin-Thick Clients for Optimizing Data Distribution and Resource Allocation in Cloud Computing 685
 Pham Phuoc Hung and Eui-Nam Huh

Mutual Exclusion Algorithm in Mobile Cellular Networks 695
 Sung-Hoon Park and Yeong-Mok Kim

An Effective Personalized Service Provision Scheme for Ubiquitous Computing Environment. 703
 Chung-Pyo Hong, Cheong-Ghil Kim and Shin-Dug Kim

Fast Data Acquisition with Mobile Device in Digital Crime. 711
 Chang-Woo Song, Jeong-Hyun Lim, Kyung-Yong Chung, Ki-Wook Rim and Jung-Hyun Lee

R²TP: Remote-to-Remote Transfer Protocols and Methods. 719
 Tae-Gyu Lee and Gi-Soo Chung

The Stack Allocation Technique on Android OS. 727
 Yeong-Kyu Lim, Cheong-Ghil Kim, Min-Suk Lee and Shin-Dug Kim

New Load Balancing Method for VoD Service 733
 Jinsul Kim, Kang Yong Lee and Sanghyun Park

Exploiting Mobility/Traffic Characteristics for PMIPv6-Based Distributed Mobility Management 743
 Ki-Sik Kong

A Multihoming-Based Vertical Handover Scheme 749
 Hee-Dong Park and Kyung-Nam Park

A Performance Prediction Model of Parallel DCT on Mobile Embedded Systems 755
 Yeong-Kyu Lim and Cheong-Ghil Kim

Implementation of a Low Cost Home Energy Saver Based on OpenWrt 761
 Se-Hwan Park, Hyun-Jun Shin, Myoung-Seo Kim and Cheong-Ghil Kim

Part VII Multimedia and Information Visualization

A Multimedia Authoring and Virtual Collaboration System Supporting Multi-Conferences for E-Learning 769
Yeongjoon Kim and Chuleui Hong

An Immersive Ski Game Based on a Simulator 777
Gil Ho Song, Won-Hyung Park, Eun-Jung Lim, Goo Cheol Jeong and Sang-Youn Kim

A Method of Combining Gaussian Mixture Model and K-Means for Automatic Audio Segmentation of Popular Music 787
Ing-Jr Ding

Human Action Classification and Unusual Action Recognition Algorithm for Intelligent Surveillance System. 797
Nae Joung Kwak and Teuk-Seob Song

Design of Configurable Pin Control Block for Multimedia System-on-a-Chip 805
Myoung-Seo Kim and Jean-Luc Gaudiot

Depression and Fatigue Analysis Using a Mental-Physical Model . . . 813
Xue-Wei Tian, Zhen-Xing Zhang, Sang-Hong Lee, Hee-Jin Yoon and Joon S. Lim

A Study of Non-photorealistic Rendering Method Using Orientation Sensor in Mobile Devices. 819
Sungtae Kim, Minseok Kang, Jiyeon Kim, Hongil Kim, Gukboh Kim and Jongjin Jung

A BCI Contents Development Method Based Templates 829
Yunsick Sung, Kyungeun Cho and Kyhyun Um

Pen-Ink Rendering for Traditional Building Images 837
Dokyung Shin and Eunyoungh Ahn

Context-Aware Statistical Inference System for Effective Object Recognition 843
Sung-Kwan Kang, Kyung-Yong Chung, Kee-Wook Rim and Jung-Hyun Lee

Adaptive Skinny Smudge Tool. 853
Noyoon Kwak and Eunyoungh Ahn

Modeling Student’s Handwritten Examination Data and Its Application Using a Tablet Computer. 861
 Youngjae Kim, Cheolil Lim, Haewook Choi and Minsoo Hahn

Advanced Media Measuring Method Using MPEG-2 Transport Stream for High Quality Broadcasting Management System. 867
 Sangkeun Kim

Proposed Media Signal Sharing Scheme Through NGN for Service Overlay Multimedia Framework 873
 Jungdae Kim

Realtime Sport Analysis Methodology for Extracting Target Scenes on Mobile Environment. 881
 Chung Young Lee and Jung Mo Kim

A Cost Effective Method for Matching the 3D Motion Trajectories. 889
 Hai-Trieu Pham, Jung-ja Kim and Yonggwon Won

Perceived Quality Model for Supporting Full Session Mobility in Multimedia Service Delivery Process 897
 Dongjun Suh, Jinsul Kim and Seongju Chang

A Study of Stereoscopic 3D Technology Development Trends on Mobile 905
 Cheong-Ghil Kim, Se-Hwan Park, Bong-Jin Back and Taeg-Keun Whangbo

Efficient Object Recognition Method for Adjacent Circular-Shape Objects 911
 Sung-Jong Eun and Taeg-Keun Whangbo

Part VIII Convergence Data Mining and Artificial Intelligence

Improved View Selection Algorithm in Data Warehouse. 921
 Jong-Soo Sohn, Jin-Hyuk Yang and In-Jeong Chung

A Novel Weighting Technique for Mining Sequence Data Streams. . . 929
 Joong Hyuk Chang and Nam-Hun Park

Analyzing Efficient Algorithms of Frequent Pattern Mining 937
 Unil Yun, Gangin Lee and Sung-Jin Kim

Efficient Isomorphic Decision for Mining Sub Graphs with a Cyclic Form 947
Gangin Lee and Unil Yun

Performance Evaluation of Approximate Pattern Mining Based on Probabilistic and Statistical Techniques 957
Unil Yun, Gwangbum Pyun and Sung-Jin Kim

Interactive Self-Diagnostic System Using Anatomical 3D Human Body 967
Jong-Hun Kim and Kyung-Yong Chung

Rule-Based Naive Bayesian Filtering for Personalized Recommend Service 977
Jong-Hun Kim and Kyung-Yong Chung

Design of an Actigraphy Based Architecture for Mental Health Evaluation 985
Mi-hwa Song, Jae-Sung Noh, Seung-Min Yoo and Young-Ho Lee

Efficient Detection of Content Polluters in Social Networks 991
Jin Seop Han and Byung Joon Park

A Prototype Selection Algorithm Using Fuzzy k-Important Nearest Neighbor Method 997
Zhen-Xing Zhang, Xue-Wei Tian, Sang-Hong Lee and Joon S. Lim

Enhanced Reinforcement Learning by Recursive Updating of Q-values for Reward Propagation 1003
Yunsick Sung, Eunyoung Ahn and Kyungeun Cho

Improved Method for Action Modeling Using Bayesian Probability Theory 1009
Yunsick Sung, Kyhyun Um and Kyungeun Cho

Decision Tree Driven Rule Induction for Heart Disease Prediction Model: Korean National Health and Nutrition Examinations Survey V-1 1015
Jae-Kwon Kim, Eun-Ji Son, Young-Ho Lee and Dong-Kyun Park

Data Mining-Driven Chronic Heart Disease for Clinical Decision Support System Architecture in Korea 1021
Eun-Ji Son, Jae-Kwon Kim, Young-Ho Lee and Eun-Young Jung

A Study on the Occurrence of Crimes Due to Climate Changes Using Decision Tree. 1027
 Jong-Min Kim, Awang-Kwon Ahn and Dong-Hui Lee

A Case Study for the Application of Storage Tiering Based on ILM through Data Value Analysis. 1037
 Chun-Kyun Youn

A Neural Network Mixture Model for Green Warranty Diffusion. . . 1055
 Sang-Hyun Lee, Sang-Joon Lee and Kyung-Il Moon

Part IX Web Technology and Software Engineering

Generation of User Interest Ontology Using ID3 Algorithm in the Social Web 1067
 Jong-Soo Sohn, Qing Wang and In-Jeong Chung

Collective Intelligence Based Algorithm for Ranking Book Reviews 1075
 Heungmo Ryang and Unil Yun

Ranking Techniques for Finding Correlated Webpages. 1085
 Gwangbum Pyun and Unil Yun

Square-Wave Like Performance Change Detection Using SPC Charts and ANFIS 1097
 Dong-Hun Lee and Jong-Jin Park

Hybrid Standard Platform for E-Journal Usage Statistics Management. 1105
 Youngim Jung and Jayhoon Kim

Study for Predict of the Future Software Failure Time Using Nonlinear Regression 1117
 Yoon-Soo Ra and Hee-Cheul Kim

Analysis of Threat-Factors for Biometric-Information Processing Systems According to Goal-Oriented Threat- Modeling 1125
 Su-Jin Baek, Jong-Won Ko and Jung-Soo Han

Distinct Element Method Analysis of Retaining Wall Using Steel Frame and Fill Material 1133
 Sam Dong Jung, Jung Won Park, Jong Hwa Won, Jeong Soo Kim and Moon Kyum Kim

Full-Scaled Experiment for Behavior Investigation of Reinforced Concrete Columns with High-Strength Wire Ropes as Lateral Spiral Reinforcement 1139
 Kyu Won Kim, Jong Hwa Won, Sam Dong Jung, Jung Won Park and Moon Kyum Kim

Local Deformed Diameter Analysis of a Pipe in Rigid Encasement for Water-crossings Application. 1147
 Jong Hwa Won, Gun Kim, Sam Dong Jung, Jung Won Park, Do Hak Kim and Moon Kyum Kim

A Study on Management System Design of Swimming Exercise Prescription by Using Fussy ANP 1157
 Kyoung-Hun Kim, Won-Hyun Kim, Tae-Won Kyung, Gyeng-Taek Yu and Chung-Sick Shin

Intelligent Recommendation System for Automotive Parts Assembly 1165
 Jong-Won Ko, Su-Jin Baek and Gui-Jung Kim

Model Transformation Verification Using Mapping Pattern and Model Transformation Similarity 1171
 Jong-Won Ko, Su-Jin Baek and Jung-Soo Han

Hierarchical Analysis of Steel House Material for 3D. 1179
 Jung-Soo Han and Myeong-Ho Lee

Multi-Faces Recognition Process 1185
 Jung-Soo Han and Jeong-Heon Lee

Software Performance Test Automation by Using the Virtualization 1191
 Gwang-Hun Kim, Yeon-Gyun Kim and Seok-Kyu Shin

Requirement Analysis for Aspect-Oriented System Development. 1201
 Seung-Hyung Lee and Hyun Yoo

System Analysis and Modeling Using SysML 1211
 Muzaffar Iqbal, Muhammad Uzair Khan and Muhammad Sher

Part X Green Convergence Services

Handover Latency Reduction Scheme for Railroad Communications in 4G Mobile Networks 1223
Ronny Yongho Kim and Baik Kim

Neo Energy Storage Technology: REDOX Flow Battery 1231
Sunhoe Kim

An Effective Interfacing Adapter for PRAM Based Main Memory via Flexible Management DRAM Buffer 1237
Mei-Ying Bian, Su-Kyung Yoon and Shin-Dug Kim

Erratum to: IT Convergence and Security 2012 E1
Kuinam J. Kim and Kyung-Yong Chung

Part I
Security Fundamentals

Development of an Automatic Document Malware Analysis System

Hong-Koo Kang, Ji-Sang Kim, Byung-Ik Kim and Hyun-Cheol Jeong

Abstract Malware attacks that use document files like PDF and HWP have been rapidly increasing lately. Particularly, social engineering cases of infection by document based malware that has been transferred through Web/SNS posting or spam mail that pretends to represent political/cultural issues or a work colleague has greatly increased. The threat of document malware is expected to increase as most PC users routinely access document files and the rate of this type of malware being detected by commercial vaccine programs is not that high. Therefore, this paper proposes an automatic document malware analysis system that automatically performs the static/dynamic analysis of document files like PDF and HWP and provides the result. The static analysis of document based malware identifies the existence of the script and the shell code that is generating the malicious behavior and extracts it. It also detects obfuscated codes or the use of reportedly vulnerable functions. The dynamic analysis monitors the behavior of the kernel level and generates the log. The log is then compared with the malicious behavior rule to detect the suspicious malware. In the performance test that used the actual document malware sample, the system demonstrated an outstanding detection performance.

Keywords Document · Malware · Automatic analysis system

H.-K. Kang (✉) · J.-S. Kim · B.-I. Kim · H.-C. Jeong
Team of Security R&D, Korea Internet and Security Agency, 78, Garak-dong,
Seoul, Songpa-gu, South Korea
e-mail: redball@kisa.or.kr

J.-S. Kim
e-mail: jisang@kisa.or.kr

B.-I. Kim
e-mail: kbi1983@kisa.or.kr

H.-C. Jeong
e-mail: hcjung@kisa.or.kr

1 Introduction

Malware attacks like Advanced Persistent Threat (APT) and spam mail using a document file have been rapidly increasing lately. These attacks are mostly used in the social engineering method, which uses a Web/SNS posting containing political and cultural issues, to induce the users download the malware, or that pretends to be a work colleague and that sends spam mail with document malware attached to it to infect the users with malware [1, 2]. Since most PC users routinely use document files, they are more vulnerable to document based malware than the existing types of PE (Portable Executable) malware. Moreover, the rate of this type of malware being detected by commercial vaccine programs is not that high. Since the commercial vaccine programs use the signature based detection method, which has a low rate of detecting document malware, the threat of document malware is expected to continue to increase [3, 4].

Therefore, this paper proposes an automatic document malware analysis system that will automatically perform the static/dynamic analyses of document files like PDF and HWP and that will provide the result. The static analysis of document malware identifies the existence of the script and the shell code that is generating the malicious behavior and extracts it. It also detects obfuscated codes or the use of reportedly vulnerable functions. The dynamic analysis monitors the behavior of the kernel level and generates the log. The log is then compared with the malicious behavior rule to detect the suspicious malware. In the performance test that used the actual document malware sample, the system demonstrated an outstanding detection performance.

2 Related Studies

The leading automatic malware analysis systems include Anubis [5], CWSandbox [6], and Wepawet [7].

Figure 1 shows the result of the malicious code analysis by Anubis. Anubis provides a Web based automatic malware analysis service. It provides the attribute data of the input file as well as the behavior data such as the registry, file, and process. It also provides the analysis result of the file and process that was derived from the malicious code [5]. However, Anubis does not provide the analysis of document based malicious codes, although it does provide the analysis of PE types of malicious codes.

Figure 2 shows the result of malicious code analysis by CWSandbox. Like Anubis, CWSandbox provides the automatic analysis of PE types of malicious codes. Unlike Anubis, CWSandbox provides the result of the PE file analysis through e-mail. Particularly, it provides the behavior data with the time that it occurred. [6]. However, like Anubis, CWSandbox does not provide the analysis of document based malicious codes.

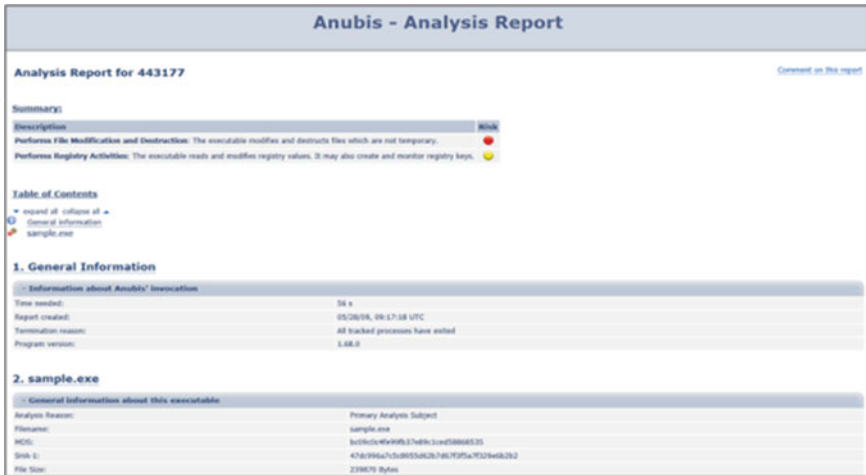


Fig. 1 The Anubis analysis result

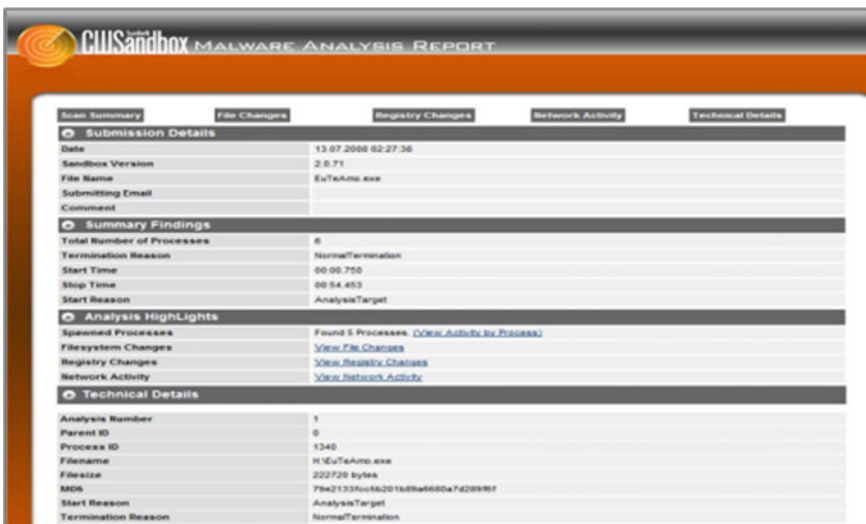


Fig. 2 CWSandbox analysis result

Lastly, Wepawet performs the static/dynamic analyses of the PDF and provides the behavior result [7]. Figure 3 shows the result of the PDF file analyzed by Wepawet.

Wepawet extracts the script/shell code that is contained in the PDF file and provides the behavior data of the extracted codes. For the generated file, it shows the result of applying the commercial vaccine program. However, Wepawet is limited in that it only provides the analysis of PDF format document based malicious codes.

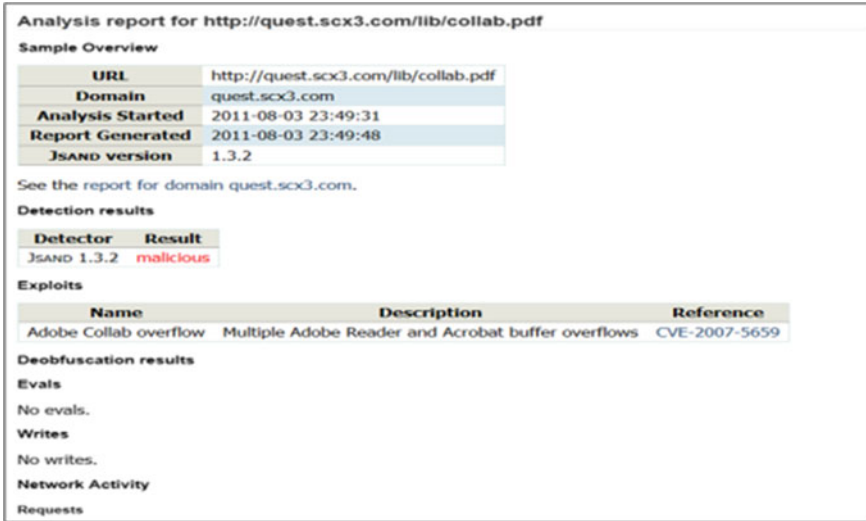


Fig. 3 PDF analysis result

3 System Design

The files targeted by the automatic document malware analysis system that are proposed in this paper are PDF, MS-Office, and HWP files. The system mainly consists of the analysis management module, static analysis module, and dynamic analysis module. Figure 4 shows the overall architecture of the automatic document malware analysis system.

As shown in Fig. 4, the analysis management module receives the request for analysis and management of the data. The static analysis module and dynamic analysis module retrieve the analysis request file from the DB, perform the analysis, and store the results in the DB. Figure 5 shows the analysis management module.

As shown in Fig. 5, the analysis management module performs the task of saving the analysis request data in the management DB so that the system will perform the static/dynamic analyses upon an analysis request by a Web user/external system.

The static analysis system uses the fact that most of the malicious behaviors in a document file are executed by the scripts/shell codes and checks if there is any script/shell code in the document file and extracts it if it exists. For example, it checks if/JS or/JavaScript naming is used in a PDF file and extracts the relevant java script. It extracts the VB scripts included in the macro of an MS-Office file. In a PDF file, it decodes five types of obfuscated codes that are applied to/Filter and detects if any reportedly vulnerable functions are used.

The dynamic analysis module checks if there is an analysis request in the DB and initializes the environment to begin the dynamic analysis. It then performs the

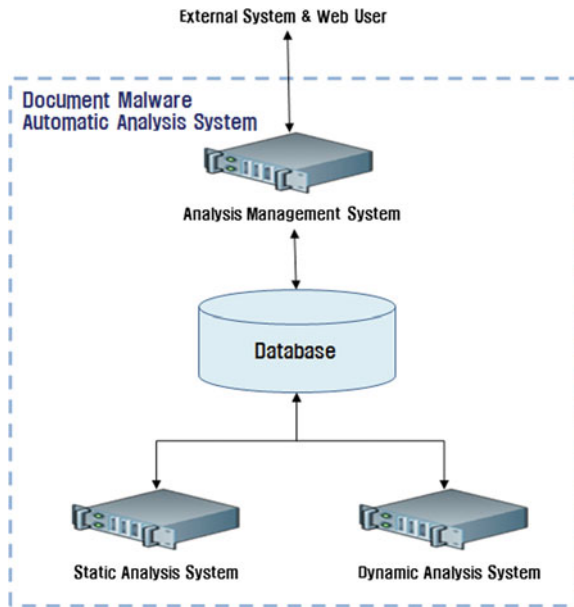


Fig. 4 Overall system architecture

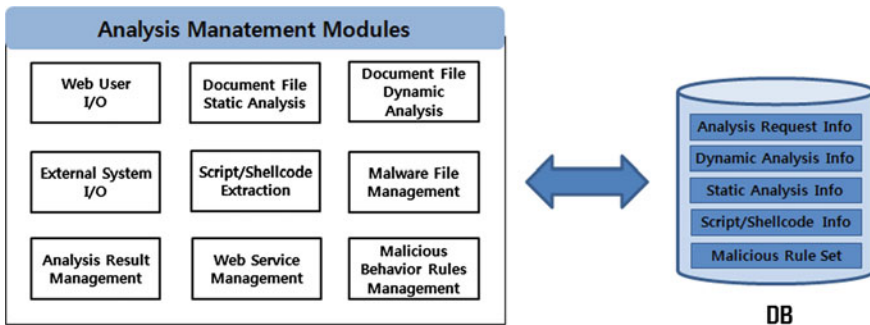


Fig. 5 Analysis management module

analysis of the document file and extracts the file/network/registry/process/memory behavior data as the analysis result. The extracted behavior data is compared with the malicious behavior rule, which is saved in the DB, to check potential maliciousness and the results are recorded in the DB. Figure 6 shows the dynamic analysis process for the document file.

Most malware document files generated the malicious codes, which perform malicious behaviors, during the file execution. Since the executable file generated by a document file is highly likely to be a malicious code, the files created by a

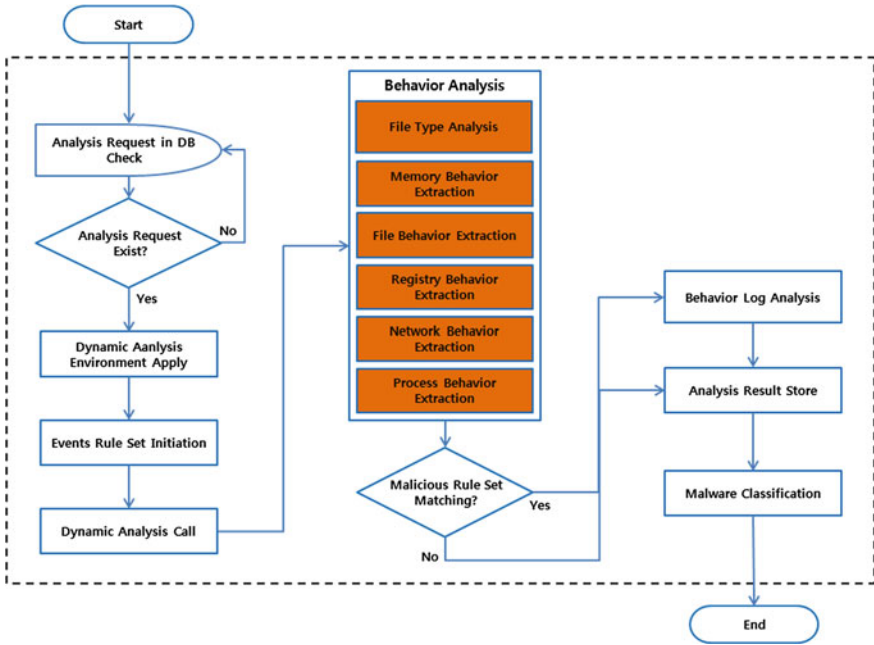


Fig. 6 The dynamic analysis process

document file need to be managed in the same group and be statically/dynamically analyzed. Figure 7 shows the process of analyzing the file generated by a document file.

As the static/dynamic analysis modules are configured as being the virtual environment, and they consist of many GeustOS systems. Each GeustOS performs the analysis of the input document file. Having many GeustOSs enables simultaneous analysis of multiple files.

4 System Implementation

The automatic document malware analysis system is deployed using the Web interface. A dotNet Framework and IIS Web server were deployed in the Windows 2003 server, and ASP was used to produce the Web pages. The automatic document malware analysis system provides not only the analysis of the document file uploaded by the administrators but also the I/O interface that can link with external systems. While the analysis is performed, the administrator can monitor the progress of the static/dynamic analyses in real-time. When the static/dynamic analyses of a document file are completed, the result is shown in the analysis result list. Figure 8 shows the document file analysis result.

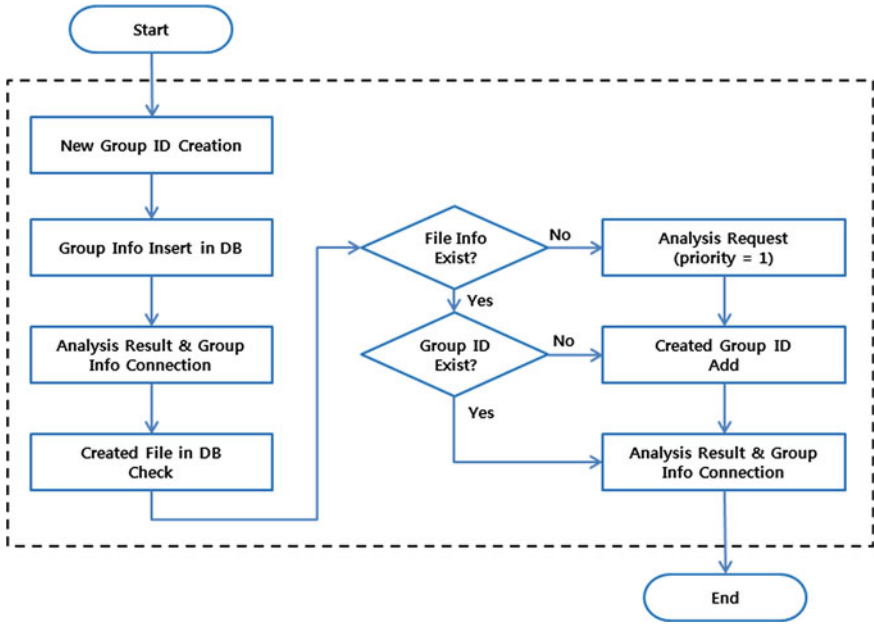


Fig. 7 Secondary generated file analysis process

The document file analysis result list, as shown in Fig. 8, displays a list of analyzed files. Users can query the files by various conditions like file type, hash value (MD5, SHA1), and period. Moreover, the detailed information of a specific file in the query can be checked. Figure 9 shows the detailed analysis information.

No.	아이디	유저	파일 형식	MD5	SHA1	악성 여부	위험도	분석 요청 시간	분석 완료 시간
372	KSA-7029	User	PDF	643F8612D593185646004840CC287	876056436611742669576602260c7c766a7e	악성	중	2012-09-03 17:16:09	2012-09-03 17:16:41
371	KSA-8917	User	PDF	20C21448F9FD4758E25E35451E3634	5e5ab6e6e75ab6c6f694645c0577645676e3	악성	중	2012-09-03 17:07:38	2012-09-03 17:09:11
370	KSA-7230	User	PDF	44933C4B17EE946FC3960303CA7E	86226ec8e726955a270e6229e1c039e6791c	악성	중	2012-09-03 17:07:23	2012-09-03 17:09:12
369	KSA-8163	User	PDF	523A76495D40642181C016A22F181C	6290481c95a5a0afefee2e3772074d97e5a	악성	중	2012-09-03 17:07:18	2012-09-03 17:09:29
368	KSA-8162	User	PDF	04E25834C5205C40379F24E21159	875426999a7e5f05c326a32d698323e42975	악성	중	2012-09-03 16:59:21	2012-09-03 16:59:52
367	KSA-8161	User	PDF	D1897834EECE9819C3C0C379181C06	ae1a5e6ba10a03934a6c02617550d05fc	악성	중	2012-09-03 16:57:44	2012-09-03 16:59:54
366	KSA-8160	User	PDF	D12104400C793240150A263148280	1a6f99022946c2394a6a1486a4744155a76a	악성	중	2012-09-03 16:57:44	2012-09-03 16:59:34
365	KSA-8159	User	PDF	C86E8B8144D095E837E91004180889	40c755567e8d8eae04b6b61c2af6a73594892	악성	중	2012-09-03 16:57:07	2012-09-03 16:59:34
364	KSA-8158	User	PDF	CA1C1C4B23E364E81849E909640C4	85c2096c337872261b1400b65c6a8b232c	악성	중	2012-09-03 16:57:07	2012-09-03 16:59:36
363	KSA-8157	User	PDF	C487102404E7489C84120D1CE8E8D49	794b264432e9e9a765e6f8900a67629802205e	악성	중	2012-09-03 16:56:28	2012-09-03 16:57:43
362	KSA-8156	User	PDF	8E2C878F8E81192321CE448A0900E	940c589706461a47e7e2213e6ca100116538	악성	중	2012-09-03 16:56:28	2012-09-03 16:57:46
361	KSA-8155	User	PDF	C280C05944D1C0858E9F151597942	63ab3954b119a1a2959d510249a78e3e3691	악성	중	2012-09-03 16:55:49	2012-09-03 16:57:02
360	KSA-8154	User	PDF	C2898C2640E3F4F391D58C370E83846	409c55833b72c15e0199646a13a542994271b	악성	중	2012-09-03 16:55:49	2012-09-03 16:57:06
359	KSA-8153	User	PDF	17A1998210E413F965235A859A80487	02c26c4030483ac3674e0c89a235a89a37611	악성	중	2012-09-03 16:55:16	2012-09-03 16:56:46
358	KSA-8152	User	PDF	8918507150E320ACE18C10D69F2D087	e6183654ae0487777773c0465a239a0c5e6d	악성	중	2012-09-03 16:51:59	2012-09-03 16:53:12
357	KSA-8151	User	PDF	842041102E7919121282D9E2C1A732	611998900142b3e6a475707a666704402cb	악성	중	2012-09-03 16:51:59	2012-09-03 16:53:16
356	KSA-8150	User	PDF	8A8F2C28321830C49D18E80440C30	91b1612ed0c038a5479623a2117a732447d67	악성	중	2012-09-03 16:48:54	2012-09-03 16:49:04
355	KSA-8149	User	PDF	A13F87649A8C81214083C70B0E40C	a4b7c3485437d1e1e524c07e19a6e3e199573	악성	중	2012-09-03 16:48:54	2012-09-03 16:48:10
354	KSA-8148	User	PDF	8A85218E664408908E285199F94403	8215c8e1404948E2794891a4a407405444	악성	중	2012-09-03 16:43:07	2012-09-03 16:44:31
353	KSA-8147	User	PDF	8917220F44C708C396821E283C230	99a6c1c554d47809203a6810c1897a85e09e	악성	중	2012-09-03 16:43:07	2012-09-03 16:45:57

Fig. 8 Document file analysis result list

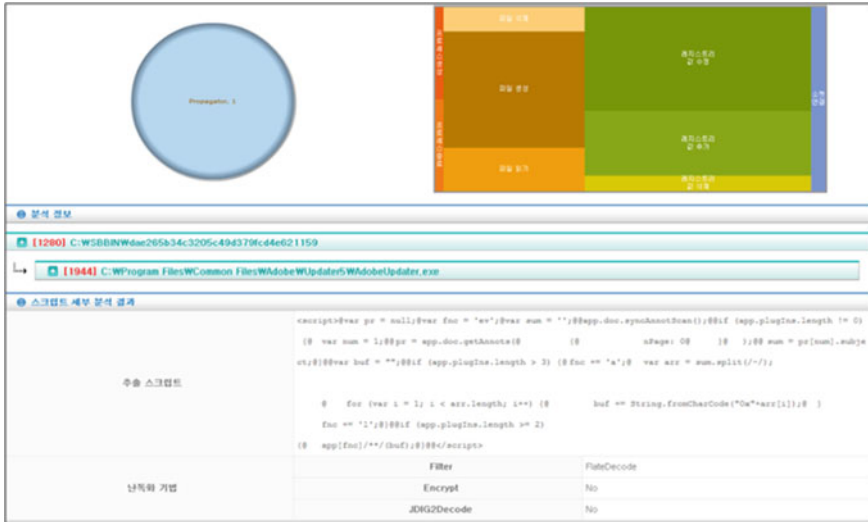


Fig. 9 Detailed document file information

Figure 9 shows the static/dynamic detailed analyses results of the document file. Users can check the scripts/shell codes that were extracted by the static analysis and whether code obfuscation and reportedly vulnerable functions were used. In the dynamic analysis, the result of the behavior analysis can be checked. The extracted behaviors are compared with the malicious behavior rule to determine the level of maliciousness. The malicious behavior rule is divided into the file, registry, process, network, and memory. The rule can be added or edited.

5 Performance Test

This paper used an actual document malware sample which was reported, for testing. A vaccine developer who is jointly studying this with KISC provided the samples that were used in testing. Table 1 shows the number of document malware samples used in the test.

Table 2 shows the number of malicious codes that were detected by the proposed automatic document malware analysis system from out of the document

Table 1 Number of samples

Type	No. of samples
PDF	50
MS-Office	28
HWP	10
Total	88

Table 2 Number of detected malware

Type	No. of detected malware
PDF	40
MS-Office	18
HWP	5
Total	63

malware samples in Table 1. Table 2 indicates that the automatic document malware analysis system proposed in this paper has the outstanding detection rate of 71.6 % for PDF, MS-Office, and HWP documents.

6 Conclusions

This paper proposed an automatic document malware analysis system that can automatically analyze document files. The static analysis extracted the scripts/shell codes from the document file and detected any obfuscation or use of reportedly vulnerable functions. The dynamic analysis monitored behaviors and determined the maliciousness based on the malicious behavior rule to detect the document files that were suspected of being malicious. The testing of the system on the actual document malicious code samples showed outstanding performance.

Although obtaining new samples is very important to increase the detection rate of document based malware, there is no efficient sample collection channel in Korea. In the future, a function to provide a Web based document file analysis service, like Wepawet, to general users is needed to secure the new samples.

Acknowledgments This research was supported by the KCC(Korea Communications Commission), Korea, under the R&D program supervised by the KCA(Korea Communications Agency)”(KCA-2012-(10912-06001)).

References

1. Park CS (2010) An email vaccine cloud system for detecting Malcode-Bearing documents. *J KMS* 13(5):754–762
2. Han KS, Shin YH, Im EG (2010) A study of spam spread malware analysis and countermeasure framework. *J SE* 7(4):363–383
3. BoanNews (2012) <http://www.boannews.com/media/view.asp?idx=31322&kind=1>, 2012
4. Ratantonio Y, Kruegel C, Vigna G, Shellzer (2011) a tool for the dynamic analysis of malicious shellcode. In: Proceedings of the international symposium on RAID, pp 61–80
5. Ulrich B, Imam H, Davide B, Engin K, Christopher K (2009) Insights into current malware behavior In: 2nd USENIX workshop on LEET, 2009
6. CWSandbox: Behavior-based Malware Analysis. <http://mwanalysis.org/>
7. Marco C, Christopher K, Giovanni V (2010) Detection and analysis of drive-by-download attacks and malicious JavaScript code. In: Proceedings of the WWW conference, 2010

Study of Behavior-Based High Speed Visit/Inspection Technology to Detect Malicious Websites

Ji-Sang Kim, Hong-Koo Kang and Hyun-Cheol Jeong

Abstract While the Web provides much convenience and many people all over the world use it almost every day, it is often misused as a medium for distributing malware without users' knowledge. Special care is particularly needed with regard to Websites that are popular with users, since their infection with malware can greatly extend the scope of any damage. Damage caused by malware can be minimized by detecting malicious sites and taking the necessary countermeasures early on. As attack techniques have been evolving, including the abuse of unknown vulnerabilities and the application of detection evasion technology, the advancement of detection technology is urgently required. Leading methods of inspecting the malware concealed in websites include low interaction Web crawling detection, which is fast but dependent upon the signature, and high interaction behavior-based detection, which offers a wide detection range and enables the detection of unknown attacks, although it is somewhat slow. This paper proposes a technology that can visit and quickly inspect large websites to more accurately detect unknown attacks and detection-evading attacks.

J.-S. Kim (✉) · H.-K. Kang · H.-C. Jeong
Team of Security R&D Korea Internet and Security Agency (KISA) Seoul,
Korea IT Venture Tower, Jungdaero 135, Songpa, Seoul 138-950, Korea
e-mail: jisang@kisa.or.kr

H.-K. Kang
e-mail: redball@kisa.or.kr

H.-C. Jeong
e-mail: hcjung@kisa.or.kr

1 Introduction

The technology used to inspect the maliciousness of websites can be categorized into the low interaction method, which uses a Web crawling tool; the high interaction method, which inspects infection by enabling a dynamic visit with a Web browser; and the hybrid method, which inspects a suspicious site using a Web crawler and then visits the site.

Since inspection by a behavior-based dynamic visit does not require a signature and inspects actual infection, it is highly accurate and has a high detection rate. However, the inspection of malware concealed in a website using the behavior-based dynamic visit requires 2–3 min for each website inspection, which includes virtual machine revert and analysis.

Considering the number of websites that are active on the Internet and the number of subpages of each site, the number of URLs to be inspected in Korea alone would amount to millions or even tens of millions.

To realistically inspect so many websites using the high interaction system, the current analysis environment, which requires 2–3 min to inspect each website, would have to be dramatically improved (i.e., an acceleration of 100 times or more).

This paper describes a high-speed website visiting technology that uses the multiplex browser and multi-frame, infection-attempt identification acceleration technology using the process-file-registry correlation analysis, and distributes URI tracking technology to enable such high-speed inspection capability.

2 Related Studies

Open source groups like Honeynet.org have released HoneyPot, a behavior-based malicious website analysis client tool. However, it has the limitation of not being able to analyze multiple websites simultaneously. MS developed Honey monkey, which can inspect malicious sites by running a snapshot before visiting them and then visiting and observing changes in the sites using multiple IE processes. Although it featured a relatively fast inspection performance of 8,000 URLs per machine per day, it still required too much time to inspect large sites. As such, the Honey monkey recommended a method of increasing the detection hit rate by preselecting potentially malicious URLs such as advertising sites as the inspection targets.

As part of its Monkey-Spider project, Mannheim University in Germany developed a system for detecting malware routes and sources using a crawler by organizing a honey pot-type network of malware analysis solutions and vaccines, and then analyzing the contents downloaded through the proxy server from the target website. However, the system still had such problems as duplicated URL analysis and recursive visit error due to the limitations of the open crawler used for

website content download. KISA in Korea is operating the Web crawling-based MC-Finder and a hybrid inspection system, which has enhanced the existing Web crawling technology by enabling it to collect malicious files by dynamically visiting a suspicious URL after scrawling it first.

3 High-Speed Website Visiting Technology

3.1 Website Alive Checking

Of the domains registered in Korea, more than 40 % are reportedly inactive. As such, executing the 'Alive' check of the domain first can minimize unnecessary visits and improve the performance. Such inactive domains can be checked through DNS query transfer and TCP Syn transfer. The procedure is described below.

- ① Sending of DNS query for a high-speed check and checking of the response.
- ② After acceptance of the response to DNS, Syn is sent to TCP port 90 and Ack is checked.
- ③ Assumes that the Web service is provided to the TCP port 80 when an Ack is received.

Since such an inspection method requires fewer CPU and network resources, multi-threads can be used for inspection. The use of multiple threads enables the advance checking of a large number of URLs on the list. (In the test using 17 CPUs, 100 threads were executed to inspect 1.8 million sites in 4 h. As a result, the number of targets to be investigated was reduced from 1.8 million to 1 million.)

3.2 High-Speed Visit Using Multiplex Browser and Multi-Frames

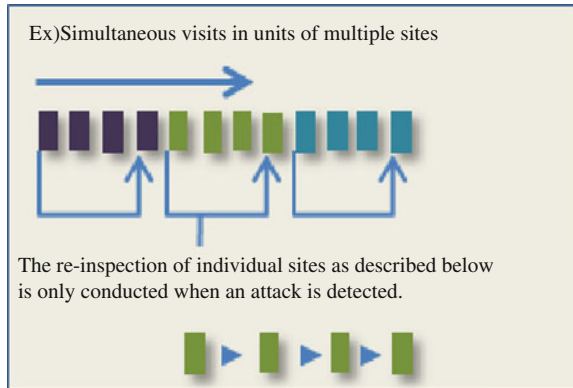
The high-speed inspection method introduced here uses the multiplex browser and multi-frames. It visits multiple websites by opening multiple Web browsers simultaneously. A main page is visited by 30 or more multiplex browsers simultaneously, while a visit to the subpages is accelerated by applying the multiplex browser and multi-frame visit techniques simultaneously.

When using 20 browsers with 5 frames, 100 sites (5×20) can be inspected simultaneously.

Multi-frames are used only to inspect the subpages.

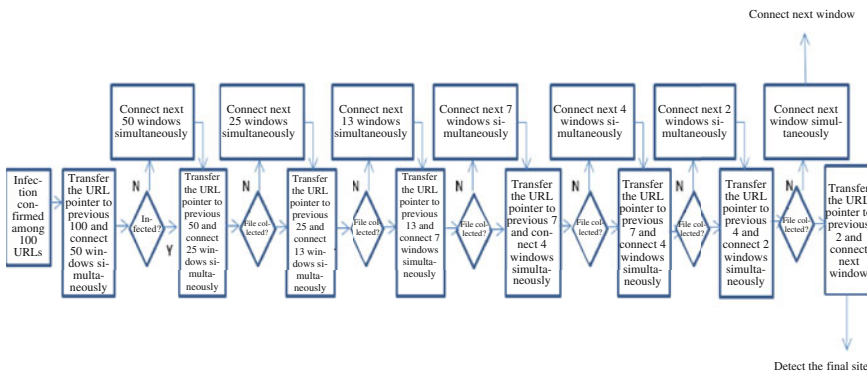
Sites are simultaneously visited using the multiplex browsers and multi-frames. If an infection attempt is not observed, then the net inspection target group is visited. If an infection attempt is confirmed, the suspicious site is tracked.

To track a suspicious site, the following tree method is used to quickly track the site with the minimum number of inspections.



If an infection attempt is detected among the 100 sites simultaneously visited, those sites are revisited in units of 50, i.e., 1/2 of the original number of sites. If an infection attempt is detected in a unit, then those sites are revisited in units of 25, i.e., 1/2 of the 50 sites. Inspection and re-inspection are recursively executed. Such a tree algorithm based re-inspection method can be greatly effective as the number of sites simultaneously visited increases. For example, when 100 sites are tracked, a malicious site can be identified in 7 inspections in the best case, 14 inspections in the worst case, and 10 inspections on average.

Compared to sequentially visiting one site at a time (once in the best case, 100 times in the worst case, and 50 times on average), this method can improve the performance fivefold on average.



3.3 Fast Malware Infection Attempt Identification Technology

3.3.1 Identification of an Infection Attempt by Analyzing the Correlation Pair of the Behavior Generated During the Visit

After rapidly inspecting the sites using the multiplex browsers, any vulnerability attack or malware infection attempt in the visited target site must be quickly identified.

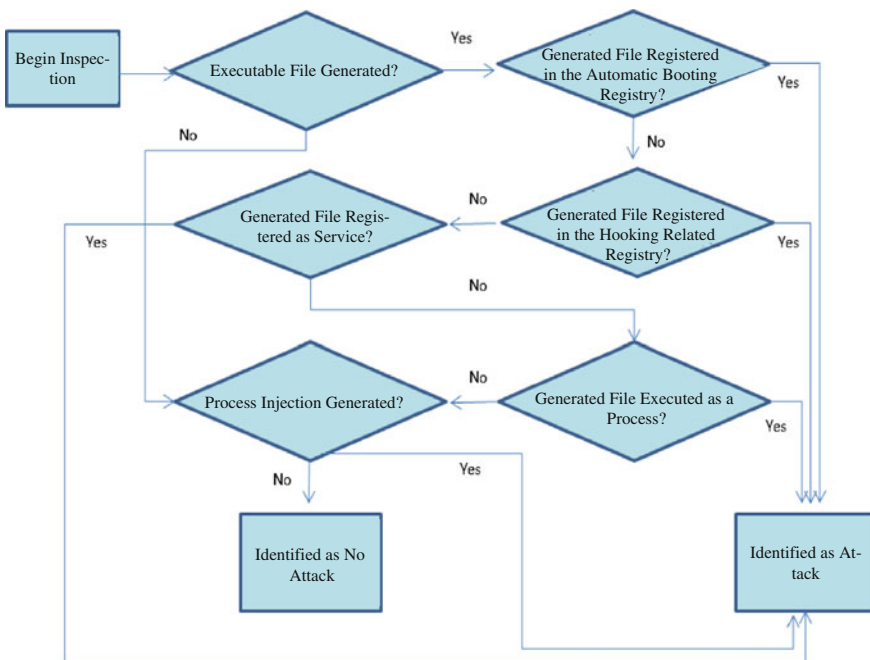
A Web browser limits the behaviors allowed after connecting a Web page to prevent security problems. The identification uses the feature to identify the infection attempt after visiting the website.

For example, one may suspect a malware infection attack if it detects executable file generation, registry registration, or process creation after a visit.

However, such behavior does not always mean malware infection has occurred since various files can be generated and processes loaded into the memory by a normal Website visit also.

Therefore, to correctly identify a malware infection attempt, correlation pair analysis is performed on the files, processes and registry registrations generated after a visit.

In other words, correlation analysis—such as the correlation of file generation and process load of the generated file, correlation of file generation and registry registration of the generated file, etc.—is used to accurately identify an infection attempt. Moreover, since the process injection can be considered as an attack on the vulnerable point, all injection generations are identified as attacks.

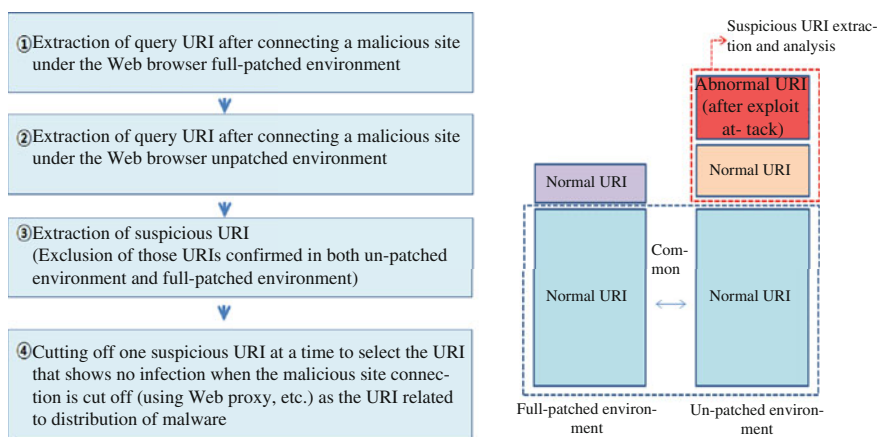


3.3.2 Malicious URI Tracking

When a malicious site is confirmed after high-speed inspection using multiplex browsers, the malicious URI within the malicious site needs to be checked.

Various codes exist in a malicious site, and it is very difficult to separate the attack codes from the normal codes. However, a malicious URI, such as malware distribution after an exploit attack, can be identified with the query session differentiation analysis of the Web browser full-patch environment and the un-patched environment, as shown below.

In the un-patched environment, an additional query such as malware download is generated after the exploit attack has been successfully executed. The detailed



procedure for tracking a URI is described below.

Of the session generated in the un-patched environment, those that cannot be observed in the full-patch environment are selected as suspicious URIs. The site is revisited after cutting off the URIs one at a time and checking the infection. If the infection is not generated after an URI has been cut off, then it is identified as the URI distributing the malware.

4 Performance Test

Tests showed that the environment described above enabled the high-speed behavior-based inspection and detection of many malicious sites. The detailed test results are shown below.

Test Performance

Condition	Performance
Tested system environment	Main page inspection
- CPU: i7v	-25,000 URL/day, 1host
- RAM : 16G	Subpage inspection
- Internet speed: 100 M	-65,000 URL/day, 1host

Domain Inspection Results

Detection system specification	Inspection target	Required period	No. of site domains inspected	No. of detected cases
1 Host - CPU: i7 - RAM: 16G - Internet speed: 100 M	1st inspection of service domains in Korea	48 h per host	More than 130,000	4 malicious site detected
	2nd inspection of service domains in Korea	48 h per host	More than 130,000	4 malicious site detected
	3rd inspection of service domains in Korea	48 h by host	More than 130,000	6 malicious site detected
	4th inspection of service domains in Korea	48 h per host	More than 130,000	0 malicious site detected
	5th inspection of service domains in Korea	48 h Per host	More than 130,000	8 malicious site detected
	6th inspection of service domains in Korea	48 h per host	More than 130,000	12 malicious site detected
	7th inspection of service domains in Korea	48 h per host	More than 130,000	4 malicious site detected

5 Conclusion

The need for high-speed, behavior-based identification technology is increasing in line with the advances made in techniques for concealing Web attacks and the ever increasing number of cases of exploitation of unknown vulnerabilities being reported. The use of multiplex browsers and high-speed identification technology is expected to help cope with malicious websites more effectively by overcoming the limitations of Web crawling to detect more malicious sites more quickly and by supplementing the existing Web crawler systems.

Acknowledgments This research was supported by the Korea Communications Commission (KCC), Korea, under the R&D program supervised by the Korea Communications Agency (KCA)”(KCA-2012-(10912-06001)).

References

1. Jamie R (2008) Server honeypot vs. client honeypot. The HoneyNet project. <http://www.honeynet.org/node/158>. Accessed Aug 2008

2. İkinci A, Holz T, Freiling F (2008) Monkey-spider: detecting malicious websites with low-interaction honeyclients. In: Proceedings of Sicherheit, Schutz und Zuverl, April 2008
3. Wang Y, Beck D, Jiang X, Roussev R, Verbowski C, Chen S, King S (2006) Automated web patrol with strider honeymonkeys: finding web sites that exploit browser vulnerabilities. In: 13th annual network and distributed system security symposium. Internet Society, San Die
4. New Zealand Honeynet Project Capture-HPC—capture—the high interaction client honeypot. <http://www.nz-honeynet.org/capture.html>
5. Kim BI, Cheong JI, Cheong HC Study of search keyword based automatic malware collection system
6. Kim BI Study of automatic collection of malware distributed through SNS. ISSN 1738-611X

One-Way Hash Function Based on Cellular Automata

Jun-Cheol Jeon

Abstract This paper proposes a secure and efficient one-way hash function based on a linear group and nonlinear non-group cellular automata (CA). The proposed architecture is based on parallelism and logical bitwise operation on CA so that our function forms remarkably simple structure. We show that the proposed one-way hash function based on a CA satisfies the secure properties and produces an excellent quality of message digest in spite of a low construction cost.

Keywords One-way hash function · Cellular automata · Confusion · Diffusion

1 Introduction

One-way hash functions play an important role in modern communication technology. The basic idea of cryptographic hash functions is that a hash-value serves as a compact representative image (sometimes called an imprint, digital fingerprint, or message digest) of an input string, and can be used as if it was uniquely identifiable with that string. Many cryptographic hash functions, all based on the so called MD4 initially proposed in [1], have received the greatest attention. However, in applications where speed is important and very large amounts of data have to be authenticated (e.g., electronic financial transactions, software integrity), hardware implementations are the natural solution. Thus dedicated cryptographic hash functions based on cellular automata are strongly recommended.

J.-C. Jeon (✉)

Department of Computer Engineering, Kumoh National Institute of Technology,
61 Daehak-ro, Gumi, Gyeongbuk, South Korea
e-mail: jcjeon@kumoh.ac.kr

Daemen et al. have persisted in vulnerability of scheme from [2] together with a new CA based hash function. Another research on CA based hash function has been reported by Mihaljecvic et al. [3] based on their previous report. They have proposed a family of fast dedicated one-way hash functions based on linear CA over $GF(q)$ in 1999. In a CA viewpoint, the above mentioned schemes are not real CA based hash functions since they did not provide any specific neighborhood and rules. Moreover, a compression function in [3] has only two times linear CA operations and other nonlinear functions are from HAVAL [4]. Thus well-defined and designed CA based hash function is exceedingly required. Though the previous papers have persisted in their security and advantages, they did not provide enough comprehension on security and experimental results. Moreover the previous works did not use specific rules so that it is hard to determine the characteristics of their schemes. One thing we know is that the security of hash functions is indeed based on confusion and diffusion. However, it is quite hard to explain a level of confusion and diffusion so that experimental results should be provided and compared with the previous well-known hash functions.

2 Cellular Automata

Cellular Automata (CA) is a collection of simple cells connected in a regular fashion. A CA was originally proposed by John von Neumann as formal models of self-reproducing organisms. Wolfram [5] pioneered the investigation of CA as mathematical models for self-organizing statistical systems and suggested the use of a simple two-state, three-neighborhood (left, self and right) CA with cells arranged linearly in one dimension. The CA structure investigated by Wolfram can be viewed as a discrete lattice of cells where each cell can assume either the value 0 or 1. The next state of a cell is assumed to depend on itself and on its two neighbors (three-neighborhood dependency). The cells evolve in discrete time steps according to some deterministic rule that depends only on local neighbors. In effect, each cell as shown in Fig. 1, consists of a storage element (D flip-flop) and a combinational logic implementing the next-state function.

In an m -state, k -neighborhood CA, each cell can exist in m different states and the next state of any particular cell depends on the present states of k of its neighbors. In this paper, we use a simple 2-state 3-neighborhood CA with the cells in one dimension.

Fig. 1 A typical CA cell

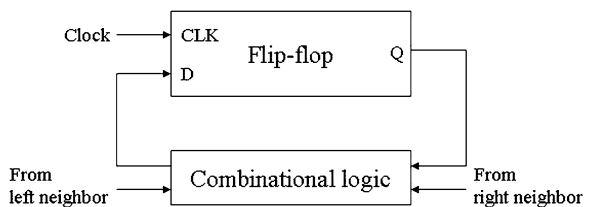


Table 1 State transition for rule 90, 202, 250 and 232 in 2-state 3-neighbor CA

Rule number	111	110	101	100	011	010	001	000
rule 90	0	1	0	1	1	0	1	0
rule 202	1	1	0	0	1	0	1	0
rule 150	1	0	0	1	0	1	1	0
rule 232	1	1	1	0	1	0	0	0

rule 90: $s_i(t+1) = s_{i-1}(t) \oplus s_{i+1}(t)$, rule 202: $s_i(t+1) = (s_{i-1}(t) \wedge (s_i(t) \oplus s_{i+1}(t))) \oplus s_{i+1}(t)$, rule 150: $s_i(t+1) = s_{i-1}(t) \oplus s_i(t) \oplus s_{i+1}(t)$, rule 232: $s_i(t+1) = (s_{i-1}(t) \wedge s_i(t)) \vee ((s_{i-1}(t) \vee s_i(t)) \wedge s_{i+1}(t))$ where \oplus , \wedge , and \vee denote the bitwise XOR, AND, and OR operations

Mathematically, the next state transition of the i th cell can be represented as a function of the present states of the i th, $(i+1)$ th and $(i-1)$ th cells:

$$s_i(t+1) = f(s_{i-1}, s_i, s_{i+1}),$$

where f is known as the rule of the CA denoting the combinational logic. For a 2-state 3-neighborhood CA, there can be a total of 2^3 distinct neighborhood configurations. For such a CA with cells having only 2 states there can be a total of $2^8 (=256)$ distinct mappings from all these neighborhood configurations to the next state. If the next-state function of a cell is expressed in the form of a truth table, then the decimal equivalent of the output is conventionally called the rule number for the cell [6].

Table 1 specifies four particular sets of transition from a neighborhood configuration to the next state. The top row gives all eight possible states of the three neighboring cells (the left neighbor of the i th cell, the i th cell itself, and its right neighbor) at the time instant t . Rows from second to fifth give the corresponding states of the i th cell at time instant $(t+1)$ for four illustrative CA rules.

If the rule of a CA cell involves only XOR logic, then it is called a linear rule. A CA with all the cells having linear rules is called a linear CA, whereas a CA with AND-OR logic is a nonlinear CA. If a state transition of a CA contains only cyclic states, then the CA is called a group CA; otherwise it is a nongroup CA. The rule applied on a uniform group CA is called a group rule; otherwise it is a nongroup rule [7].

3 Cellular Automata Based One-Way Hash Function (CAH-256)

Our scheme employs the MD (Merkle-Dameggard) structure which is well-known as a secure model [8]. Given a message M to be compressed, CAH-256 pads M first. The length of (i.e., the number of bits in) the message after padding is a multiple of 256, and padding is always applied even when the length of M is already a multiple of 256. The last block of the padded message contains the number of bits in the unpadded message. Now suppose that the padded message is

$M^{(0)} \dots M^{(n-2)} M^{(n-1)}$, where each $M^{(j)}$ is a 256-bit block. CAH-256 starts from the block $M^{(0)}$ and the all-zero 8-word (256-bit) string IV and processes the message $M^{(0)} \dots M^{(n-2)} M^{(n-1)}$ in a block-block way. More precisely, it compresses the message by repeatedly calculating $H^{(0)} = IV$ and $H^{(j+1)} = H_{CAH}(H^{(j)} \oplus M^{(j)})$, where j ranges from 0 to $n-1$ and H_{CAH} is called the updating algorithm of CAH-256. Finally $H^{(n)}$ is the hash result. In summary, CAH-256 processes a message M in the following three steps:

Step 1. Pad the message M so that its length becomes a multiple of 256. The last block of the padded message indicates the length of the original (unpadded) message M .

Step 2. A 256-bit buffer is used to hold the constant IV , intermediate and final results of the hash function. Another 256-bit buffer is required to hold the constant K .

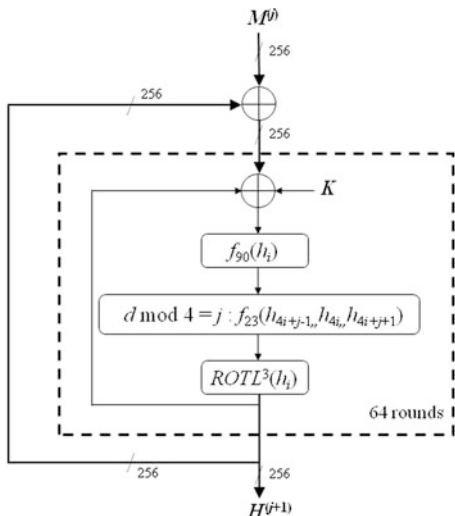
Step 3. Compute repeatedly $H^{(j+1)} = H_{CAH}(H^{(j)} \oplus M^{(j)})$ for j from 0 to $n-1$, where $H^{(0)}$ is a 8-word (256-bit) all-zero bit-string and n is the total number of blocks in the padded message M .

The main purpose of padding is security reason as used on the MD structure. The other purposes of padding are two-fold: to make the length of a message be a multiple of 256 and to let the message indicate the length of the original message. CAH-256 uses a 64-bit field to specify the length of an unpadded message. Thus messages of up to $(2^{64}-1)$ bits are accepted which is long enough for practical applications. CAH-256 pads a message by appending a single bit '1' followed by the necessary number of 0-bits until the length of the message is 192 modulo 256. Then it appends to the message the 64-bit field. Two constant vectors, IV and K which are 256-bit each, are considered. IV is all zero bit-string and K is the first thirty-two bits of the fractional parts of the cube roots of the first sixteen prime numbers such like SHA-256.

The heart of algorithm is a module that consists of processing of 64 rounds. All rounds have the same structure which is composed of XOR operations with the constant K , two CA rule functions and 3-bit shift operation. In order to design a concrete hash function, we use combinations of a linear group rule and nonlinear non-group rule. A linear group rule provides a collision resistance from present states to next states and a nonlinear non-group rule provides one-way property and nonlinearity. Rule 150 based on periodic boundary condition is only a linear group rule for a message with 256-bit length, and it has a highest dependency from neighborhood in the middle of the whole linear rules. Meanwhile, we choose rule 23 for a nonlinear non-group CA operation since rule 23 provides not only a high nonlinearity but also a special transition form.

The computation can be considered as a 4-step transformation of H_{CAH} . The calculations in each step are done simultaneously on all bits of H . Let $m_0 m_1 \dots m_{255}$ denote the bits of $M^{(j)}$ and $h_0 h_1 \dots h_{255}$ denote the bits of $H^{(j)}$, an intermediate message value during a round, and $k_0 k_1 \dots k_{255}$ denote the bits of constant K and d denotes the number of round. Before starting every rounds, the computation, $h_i = h_i \oplus m_i$ ($0 \leq i \leq 255$) is preprocessed. Fig. 2 illustrates a single step of the updating function.

Fig. 2 CAH-256 operation in a single round



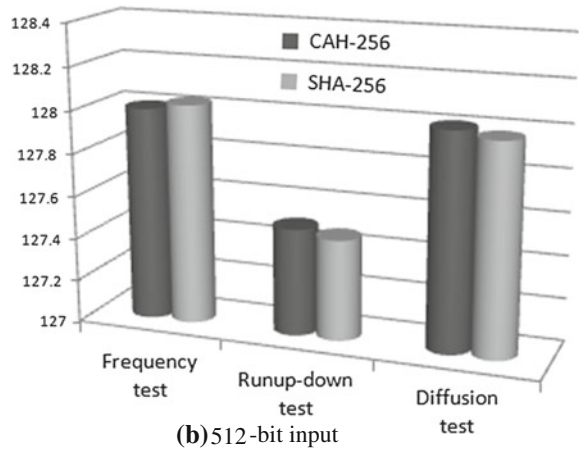
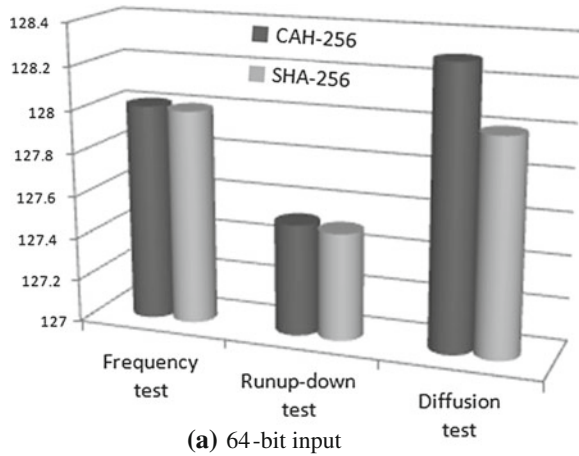
4 Security and Efficiency Analysis

Let H be an iterated hash function with MD-strengthening. Then preimage and collision attacks on H have roughly the same complexity as the corresponding attacks on compression function H [8]. Thus the security of our scheme depends on the updating function H_{CAH} .

The XOR operation with K blocks that the CA operation generates repeated patterns on the characteristics of CA operation. The linear group CA function, rule 90, generates a distinctive output result according to a different input based on group property so that it blocks a primary collision in the updating function. Let the state transition for rule 23 with three neighborhoods, s_{i-1} , s_i , and s_{i+1} be $f_{23}(s_{i-1}, s_i, s_{i+1})$, then the result from the possible states combinations, 111 to 000, is $\{00010111\}$. Suppose that a state of one of neighborhoods is complemented then we obtain the following results: $f_{23}(\neg s_{i-1}, s_i, s_{i+1}) = \{01110001\}$, $f_{23}(s_{i-1}, \neg s_i, s_{i+1}) = \{0100 1101\}$, $f_{23}(s_{i-1}, s_i, \neg s_{i+1}) = \{00101011\}$. Now we find some specific property among the results that Hamming distances among four strings are exactly 4 of 8-bit string by pairs. It guarantees that the present states via rule 23 would be updated with the same transition probability, 1/2 so that a changed input state makes the next states with a half difference. This property also makes impossible to find previous bit values from attackers.

Our function employed the rule 23 as a nonlinear rule which is balanced and guarantees a high nonlinearity. The rule 23 has another good property to make a balanced bit-string. Let $1^n 0^m$ be the three neighborhood states, and n and m are the number of 1s and 0s, respectively, where $n + m = 3$ and $0 \leq n, m \leq 3$, then there exist four different types such as $1^3 0^0$, $1^2 0^1$, $1^1 0^2$ and $1^0 0^3$. If $n \geq 2$ or $m \leq 1$,

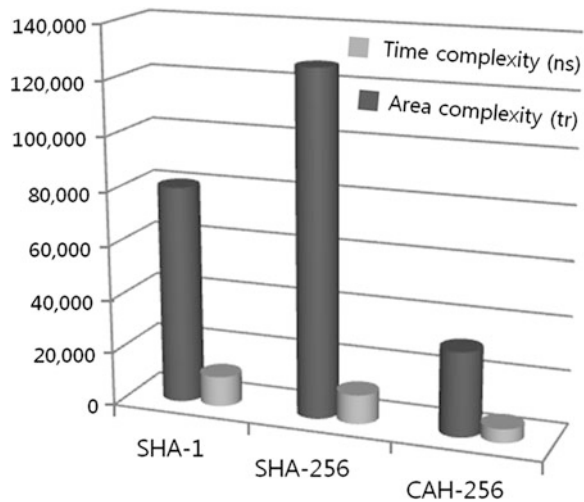
Fig. 3 Comparison of quality tests according to the number of input-bit between the proposed function and SHA-256



the next state becomes 0, otherwise 1. We have applied rule 23 to a quarter of a block (64 bits) in different bit positions at each round. Thus every bit position is equivalently applied to the nonlinear operation 16 times.

Confusion is caused by the high nonlinearity and constant K based on the repeated structure of the cellular automata mechanism. The nonlinear CA operation can generate 1's from a zero background. On the input of the next iteration, these would give rise to characteristics with high confusion effect. Hence simple difference patterns in $H^{(j)}$ gives rise to a vast amount of possible difference patterns in $H^{(j+1)}$. Each bit of $H^{(j)}$ depends on nearly 4 bits of the previous round by two CA functions and XOR operations after 3-bit shift operation so that the influence increases by 4-bit a round during the continuous 64 rounds. Thus the influence of the first injection of a message bit has spread out over all bits of $H^{(j+1)}$ with same

Fig. 4 Comparison of area and time complexity for proceeding 512-bit message



transition probability by the time of the last injection. Hence it satisfies the diffusion property. The actual message bits injection in H is realized to be diffused and confused in subsequent rounds.

In order to compare the efficiency among the schemes, we have chosen SHA-1 and SHA-256 which are known as the best quality hash functions and the current U.S. federal standard. In order to compare the quality of hash function, we made an experiment on several points of view as shown in Fig. 3. The specified test methods based on randomness test in [9] is suitably determined to examine and compare the quality of hash functions. The results show that both functions have produced good results. A frequency, runup-down and diffusion test returns the number of 1's, the number of run-up and run-down and the number of changed bits in output according to changing 1-bit input respectively. We simulated the mentioned tests using 1-million random data set.

We are usually trying to find the design that will best satisfy a given set of design requirements when we implement arithmetic unit design. We consider construction simplicity, defined by the number of transistors needed for its construction and the time needed for the signal change to propagate through gates [10].

Area is assumed to be totally contributed by the number of transistors in gates and registers required to compute a find hash result. The cost due to time consists of the delay time of the gates and registers for proceeding a 512-bit input message block. As shown in Fig. 4, our scheme based on cellular automata has outstanding complexity compared to the other well-known schemes. Consequently, the proposed CAH-256 has more than 2 times less area and time complexity than SHA-1 and SHA-256 for proceeding 512-bit message, respectively. The description allows a straightforward chip implementation. Based on parallelism and logical bitwise operation of a CA, our scheme makes extremely high speed possible.

5 Conclusions

This paper has proposed a secure and efficient cryptographic hash function. We conclude that the proposed cryptographic hash function based on a CA satisfied the secure properties and produced an excellent quality of message digest though it has an exceedingly low construction cost. Therefore, we expect that the proposed function will be efficiently used for preserving the integrity of a potentially large message.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2011-0014977).

References

1. Rivest RL (1991) The MD4 message-digest algorithm. *Crypto*, LNCS 537:303–311
2. Daemen J, Govaerts R, Vandewalle J (1993) A framework for the design of one-way hash functions including cryptanalysis of Damgard's one-way function based on a cellular automaton. *Proc Asiacrypto'91*, LNCS 739:82–96
3. Mihaljevic M, Zheng Y, Imai H (1999) A family of fast dedicated one-way hash functions based on linear cellular automata over $GF(q)$. *IEICE Trans Fundam E82-A(1)*:40–47
4. Zheng Y, Pieprzyk J, Seberry J (1993) HAVAL—a one-way hashing algorithm with variable length of output. *Auscrypt*, LNCS 718:83–104
5. Wolfram S (1983) Statistical mechanics of cellular automata. *Rev Modern Phys* 55:601–644
6. Jeon JC, Yoo KY (2008) Montgomery exponent architecture based on programmable cellular automata. *Math Comput Simul* 79:1189–1196
7. Jeon JC (2010) Nonlinear and nongroup cellular automata chaining technique for cryptographic applications. *Math Comput Modell* 51:995–999
8. Damgarrd IB (1989) A Design Principle for Hash Functions. *Crypto*, LNCS 435:416–442
9. Knuth DE (1997) *The art of computer programming, seminumerical algorithms*, vol 2, 3rd edn. Addison-Wesley Longman Publishing Co., Inc., Reading
10. Gajski DD (1997) *Principles of digital design*. Prentice-Hall International Inc., Prentice Hall

A Novel Malware Detection Framework Based on Innate Immunity and Danger Theory

Mohamed Ahmed Mohamed Ali and Mohd Aizaini Maarof

Abstract Artificial immune system (AIS) is a computational system inspired by the principles and processes of the Biological immune system which has the capabilities to learn, adapt, self tolerance and memories actions, which make it a good example that we can take for solving some major problems in many fields, including the problem of malware detection in the field of computer security. The main idea is to detect any type of files that trying to harm the computer system by infecting some executable software when these files running, spread it to other files or computers. In this paper, we proposed a framework to detect malware using the innate immune system combined with danger theory to eliminate tow major drawbacks of current malware detection methods; detection accuracy and high false positive alarms.

Keywords Innate immune system • Danger theory • Malware detection

1 Introduction

The main obstacles facing the traditional malware detection methods were the high rate of creating new malware, the ability to change their shapes from time to time and from place to place (polymorphic malware) which make the detection use the

M. A. M. Ali (✉)

Faculty of Mathematical Sciences, University of Khartoum, Khartoum, Sudan
e-mail: mamautm@gmail.com;mama@uofk.edu

M. A. Maarof

Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia,
UTM Skudai 81310 Johor, Malaysia
e-mail: aizaini@utm.my

normal model for detecting malware based on the saved data (Signature-base model) a useless job [1]. However, in the last two decades the field of the artificial immune system (AIS) creates a new research area help the researchers to overcome efficiently some problems in the field of computer science like pattern recognition, data mining, intrusion detection and malware detection [2]. The biological immune system (BIS) is a system of biological structures and processes within an organism that protects against disease by identifying and killing pathogens and tumor cells. It detects a wide variety of agents, from viruses to parasitic worms through the integration of its two parts, innate and adaptive. It needs to distinguish pathogens from the organism's own healthy cells and tissues in order to function properly [3, 4]. Detection is complicated as pathogens can evolve rapidly; producing adaptations that avoid the immune system and allow the pathogens to successfully infect their hosts, but with the main characteristics of the biological immune system like: adaptability, self- tolerance, diversity, distributable and saved memory make it easier to defeat any invaders were trying to harm the organism [5]. Artificial immune system (AIS) inherits these characteristics to overcome many problems in the field of computer security. In section two we introduce the concept of innate immune system, section three discuss the danger theory concept and its benefits, then we introduce the novel framework in section four.

2 Innate Immune System

The innate immune system represents the first line of defense in the human immune system containing some external parts like skin, mucous and stomach acids to keep pathogens out of the body and internal parts like the inflammatory response and phagocytes. Phagocytes are a class of cells (part of the white blood cells) can engulf pathogens through its surface receptors which had the ability to connect to the proteins on the pathogen surface. After this connection is happen the phagocytes cut the bacteria or virus protein into small parts called peptide to attach them to major histocompatibility complex type 2 (MHC II) to present this complex on the phagocytes surface. Phagocytes called antigen presenting cells (APCs) when they present the complex of MHC II and the peptide on its surface. Phagocytes cells comprise macrophages, neutrophils and dendritic cells. Macrophages and neutrophils are phagocytes (cellular engulfment) the invading pathogen, then killing them through a respiratory burst. The neutrophils are the numerically superior cells of white blood cells (WBC) and the faster one to receive the infected tissue. Dendritic cells include the basophils and the eosinophils, although they are categorized as phagocytic cells, they are not killing the pathogen by phagocytosis. The basophils mediate the allergic reaction, while the eosinophils kill the invader pathogen by secreting highly toxic proteins added to the Toll like receptors (TLR) which are a pathogen recognition receptors found in the cell membrane also activate the immune cell response [5–7].

3 Danger Theory

The concept of danger theory initialized by Matzinger disprove that the immune system defense mechanism depend on the definition of what is part of the organism cells and what is not, what we called (self nonself theory SNS) which treat any things coming from outside as an invader [8]. Danger theory declares that interaction of the B cell receptors with antigen initiates a signal, this signal initiate the immune response. B cells secrete specific antibodies that recognize and react to stimuli. Another type of cell, the T (killer) lymphocyte, is also important in different types of immune reactions. The Danger model added another layer of cells and signals proposing that antigen presenting cells (APCs) are activated by danger alarm signals from injured cells, such as those exposed to pathogens, toxins, and so forth [9]. Alarm signal scanned to be constitutive or inducible, intracellular or secreted, or even a part of the extracellular matrix. Because cells dying by normal programmed processes are usually scavenged before they disintegrate, whereas cells that die necrotically release their contents, any intracellular product could potentially be a danger signal when released. Inducible alarm signals could include any substance made, or modified, by distressed or injured cells [10].

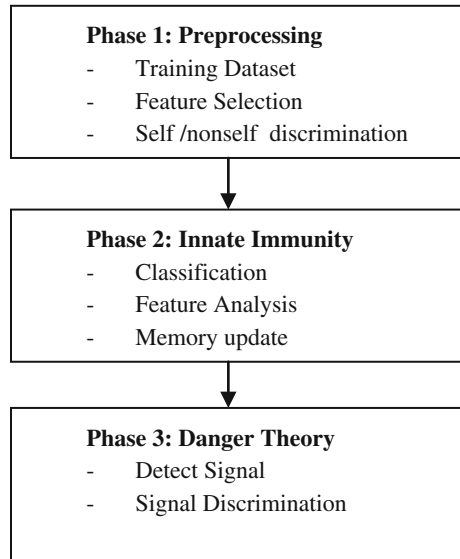
4 A Novel Malware Detection Framework Based on (Innate-Danger)

The proposed framework in Fig. 1 is composed of three main phases: the pre-processing phase, the innate immunity phase, and the danger theory phase.

4.1 Phase 1: Preprocessing

The preprocessing phase objectives is to define the self from non self depend on the data collected as a training data set to select the features that is suitable for classifying the malware from benign files and create an adaptive memory to store all the information about known malware as a signature based detection. The feature selection process means that specific features that are affected by the malware and for different types of attacks should be given the higher priority in selection compared to the features that keep out of change. By the end we will get the most important features that will help in differentiating between malware and benign files. By the end of this phase, we get the features and information about self and non self.

Fig. 1 The proposed framework



4.2 Phase 2: Innate Immunity

After complete the preprocessing phase we go through the innate immunity phase where we make the classification based on the features selected in the previous stage. These features help us to take advantages of the innate characteristics to distinguish between the harmful file and the benign file. Fetching the data in this phase leading to two processes, the first process is making a decision if the file is a malware or not depend on the signature (Pathogen Associated Molecular Patterns-PAMPs) associated with that file in hand [11]. If it is a malware then we stimulate the system to increase the detection ability of that type of malicious software and update the memory. The second process taking place if it is not a known malware we then look at the features selected from the previous stage whether the file have some of the malware features or not, if it had some of these features then maybe this is a malware and maybe not, but we must to make sure before take a decision to avoid the false positive alarms if the file is not a malware. In this case we move forward to the danger theory phase.

4.3 Phase 3: Danger Theory

In this phase we take the data from the previous phase to avoid the problem that facing the most of the accurate detection models which is the high rate of the positive alarms as a result of classifying wrongly some benign files as malware due

to the lack of information about the file being scanned. In that case we increase the percentage of the accuracy with time tradeoffs spending in dealing with wrong alarms. So we add this phase to decrease the false positive alarm. By taking the output of the innate phase we come with some files have a number of malware features. By applying the maturation examination to the signals and executables file we can eliminate the false alarms and make the detection status not positive until the triggering of signal happened. Selection of the signal depends on the features like processor (CPU) usage and memory usage or any other feature. We here select the file that make a high memory usage as a danger signal.

5 Conclusions

Malware detection is a major task nowadays not only because the importance of the information and the resources store and transfer and process these information, also because the big evolution in the creation of malware and the related malware detection industry. A lot of models and frameworks proposed during the last two decades, but have their limitations because the accuracy and false positive alarms tradeoffs. In this paper a novel malware detection framework based on innate immunity and danger theory to overcome the low detection accuracy and the high rate of false positive alarms. This work is exploratory in which many experiments will be conducted to verify the viability of the framework.

Acknowledgments This work and research is done by support of Faculty of Computer Science and Information Systems, Universiti Teknologi Malaysia, Malaysia, Faculty of Mathematical Sciences, University of Khartoum, Sudan.

References

1. Christodorescu M, Jha S, Seshia SA, Song D, Bryant RE (2005) Semantics-aware malware detection. In: IEEE symposium on security and privacy, 2005
2. Castro LND, Von Zuben FJ (1999) Artificial immune systems: part I—basic theory and applications. Technical Report, RT–DCA 01/99, Dec 1999
3. Timmis J, Knight T, Castro LND, Hart E (2004) An overview of artificial immune systems. 2004
4. Andrews L (2008) Immunity, St. Martin's Minotaur 2008
5. Kuby J (1994) Immunology. vol 2nd edn. 1994
6. Parkin J, Cohen B (2001) An overview of the immune system. *The Lancet* 357(9270):1777–1789
7. Medzhitov R (2001) Toll-like receptors and innate immunity. *Nat Rev Immunol* 1(2):135–145
8. Matzinger P (1994) Tolerance, danger, and the extended family. *Annu Rev Immunol* 12:991–1045

9. Ali MAM, Maarof MA (2012) Malware detection techniques using artificial immune system. In: Kim KJ, Ahn SJ, (eds) Proceedings of the international conference on IT convergence and security 2011, Springer, Netherlands, pp 575–587
10. Matzinger P (2002) The danger model: a renewed sense of self. *Science* 296(5566):301–305
11. Janeway CA (1989) Approaching the asymptote? Evolution and revolution in immunology. *Cold spring harbor symposia on quantitative biology*, vol 54 Pt 1, pp 1–13

Forensic Evidence Collection Procedures of Smartphone in Crime Scene

Jeong-Hyun Lim, Chang-Woo Song, Kyung-Yong Chung,
Ki-Wook Rim and Jung-Hyun Lee

Abstract As the Smart phone becomes gradually generalized and expands its influence on daily life, the digital evidential matter could be an important clue to prove criminal charge in the forensic process of criminal accident. Since the digital evidential matter could be easily spoiled, fabricated and disappeared, it needs to secure the criminal related evidence promptly as applicable according to clear procedures when investigate the initial scene of accident. Thus, this paper induces forensic procedures and items which a digital forensic investigator should take when it seizes, searches and verifies the Smart phone in the scene of accident considering characteristics of the Smart phone and establishes a criminal related

J.-H. Lim
School of International Affairs and Information, Dongguk University,
26, Pil-dong, Jung-gu, Seoul, Korea
e-mail: rockq81@dongguk.edu

C.-W. Song
HCI Laboratory, Yonghyeon 1, 4-dong, Nam-gu, Incheon, Korea
e-mail: ph.d.scw@gmail.com

C.-W. Song · J.-H. Lee
Department of Computer Science and Engineering, Inha Univ,
Yonghyeon 1, 4-dong,
Nam-gu, Incheon, Korea
e-mail: jhlee@inha.ac.kr

K.-Y. Chung (✉)
School of Computer Information Engineering, Sangji University,
83 Sangjidae-gil,
Wonju-si, Gangwon-do, Korea
e-mail: dragonhci@hanmail.net

K.-W. Rim
Department of Computer Science and Engineering, Sunmoon University,
Galsan-ri, Tangjeong-myeon, Asan-si, Chungcheongnam-do, Korea
e-mail: rim@sunmoon.ac.kr

search database and shows what kind of evidential matter for criminal charge could be collected through the applications implemented based on the said search database.

Keywords Smartphone forensic · Evidence collection · Procedure

1 Introduction

The Smartphone provides more convenient and various functions than the mobile phone does. It is possible to obtain necessary information at any time at any place by accessing the internet via Wi-Fi or 3G mobile communication network. In addition, it enables us to download and execute various free and pay applications posted in online markets and easily establish a social relation such as free communication and information sharing and making personal connections through SNS including Kakao Talk or My People etc. Now the number of Smart phone subscribers is reported to be 30 million as of September 2012 increasing suddenly since it has passed 5 million, a datum point of popularization in October 2010 [1, 2]. So, as the Smart phone becomes gradually generalized and expands its influence on our daily life, the digital evidential matter could be an important clue to prove criminal charge in the forensic process of criminal accident. Though it is difficult for the digital evidence to obtain a probative power under Criminal Procedure Act, but requirements for the seizure, search and verification are highly reinforced in Criminal Procedure Act Revision in January 2012, so the collection of digital evidence in the scene of accident may play an important role in the case [3].

This paper is composed as follows. [Section 2](#) explains about the digital forensic and classifies evidential matters by criminal type through cases. [Section 3](#) proposes behavior patterns of the investigator when it seizes searches and verifies the Smart phone in the scene of criminal accident by supplementing the forensic procedure and demonstrates whether necessary information could be substantially collected by establishing a database with relative evidential matters of criminal charges. [Section 4](#) shall draw a conclusion.

2 Digital Forensic and Evidential Matter

Digital evidential matters are so various that it needs to systemize them by type of criminal. Accumulating these systemized data shall be of assistance a lot for the investigator in directing the prompt digital evidence collection for criminals in future. [Table 1](#) shows digital evidential matters by type of criminal [4].

Digital forensic is defined by DFRWS (Digital Forensics Research Workshop), a popular academic society in the field of digital forensic [4, 5]. As digital

Table 1 Crimes in the computing environment

Name of crime	Potential computer evidence
Child abuse	-Internet history logs. -Chat logs. -Internet searches. -Images. -Movies files. -calendars/notes
Murder	-Calendars/notes. -Internet history logs. -Address books. -Images. -Financial/asset records. -Medical records. -Reproductions of signature
Harassment	-calendars/notes. -Internet history logs. -Address books. -Images. -Financial/asset records. -Internet searches about victims
Identity theft	-Credit card information. -Electronic money transfer. -Financial records. -Online banking software. -Reproductions of signature. -Forged document
Counterfeiting	-Credit card information -Financial records. -Reproductions of signature
Narcotics	-Credit card information -Electronic money transfers. -Financial records. -Fictitious identification. -Photographs of drugs and accomplices. -Unfilled prescriptions
Terrorism	-Credit card information -Electronic money transfers. -Financial records. -Fictitious identification. -VOIP software

apparatuses become various and generalized, the data in digital apparatus are used as evidence and are referred to terms such as Computer Forensic, Network Forensic, Mobile Forensic and Smart phone Forensic according to objects of survey. In addition, objects of survey shall be the activated data, file system, database, code and hidden data; various log data and the trace of using application programs according to type of data.

3 Forensic Evidence Collection Procedures of Smartphone

The regulation for digital evidence collection and its analysis of Supreme Prosecutor's office in Korea presents following procedure. Seizure, search and verification of digital apparatus and collection of digital evidential matters shall be conducted by a digital forensic investigator. When it seizes an information processing system such as a computer, it shall seize the storage medium only removing it from the information processing system in principle, but may seize the whole information processing system when it is not possible to achieve the purpose of investigation by seizing only a storage medium or the digital apparatus or digital materials could be damaged or lost. An identification paper shall be prepared and attached to confiscated articles which shall be sealed and confirmed by the signature of owner [6].

3.1 Procedure to Collect Digital Evidence of Smartphone

The regulation for digital evidence collection and its analysis of Supreme Prosecutor’s Office in Korea presents following procedure. Seizure, search and verification of digital apparatus and collection of digital evidential matters shall be conducted by a digital forensic investigator. When it seizes an information processing system such as a computer, it shall seize the storage medium only removing it from the information processing system in principle, but may seize the whole information processing system when it is not possible to achieve the purpose of investigation by seizing only a storage medium or the digital apparatus or digital materials could be damaged or lost. An identification paper shall be prepared and attached to confiscated articles which shall be sealed and confirmed by the signature of owner [7].

For a Smart phone, since the potential to spoil and damage digital evidential matters is big due to its easy portability and usability, it needs to secure criminal related evidence promptly in the scene of accident. Thus, this paper shall supplement such problems by complying with regulations of Supreme Prosecutor’s Office based on general procedure to collect digital evidence of Smart phone [8, 10, 11] and reinforcing evidence collection stages. Figure 1 shows the procedure to collect digital evidence of Smartphone. It is possible to collect and analyze evidences more quickly by using search keywords using data from Smartphone as illustrated in Fig. 1 and an accurate direction to investigate in the initial investigation.

Confiscation of Smart phone: The investigator shall obtain the signature of participant on the agreement and check any failure or significance of the Smart



Fig. 1 Procedure to collect digital evidence of smartphone

phone confiscated. In addition, it shall collect information such as type, OS and password of the Smart phone with the user [9].

Check the Battery: The investigator shall prevent the Smart phone confiscated from being used by initializing the Smart phone by factory set and removing batteries in the scene because just cutting the power off while power on may erase the data which could be the digital evidential matters or cause a fragmentation of file storage space to make it difficult to analyze them afterward [10]. So, with this reason, alternative batteries shall be connected to the Smart phone in order that the Smart phone may keep the power.

Isolation of Frequency: The investigator shall prevent damage of integrity of the Smart phone against vicious transfer of Bluetooth, WI-Fi, 3G or other unexpected frequency using the frequency isolation device.

Analysis of Collected Evidences: It needs to collect the part where there exists criminal relation out of digital evidential matters when collecting evidences and concurrently to conduct the imaging work for original data of collected digital evidence. In addition, it needs to collect and analyze the digital evidential matters by minimizing damage of evidence in order to maintain the atomicity and integrity of collected data.

3.2 Collecting Criminal Charges according to Pattern

The investigator shall collect the data which may provide relation to the accident from potential evidential matters using the case pattern search.

Dump Image: With characteristics of Smart phone, it has a specific part which is difficult to access in ordinary way when collecting evidences. Thus, such a method including Android Rooting or iPhone Jailbreak shall be needed. In this case, a Dump Image needs to be created before collecting and analyzing evidences in order to prevent spoil and damage of digital evidences. For this Dump Image, the investigator shall obtain the signature of participant in order to verify the identity and integrity by applying Hash Function. The original copy shall be preserved and the analysis work shall be conducted with duplicate copy which has the same Hash Function value.

Collecting the latest Update List of Smart phone: Since there are many manufacturers of Smart phones, various OS, and vulnerable data according to type of application. It needs above all to promptly collect the latest update lists of Smart phone at the scene.

Collecting System Log of Smart phone: Collecting the system log of Smart phone could be made at the stage of detailed analysis after transfer from the scene of accident because it has a strong element of non-vulnerability but doing it at the scene may prepare a clue to solve the accident by setting a direction to investigate in the initial investigation and presuming the suspect by obtaining criminal related digital evidential data quickly.

3.3 Data Acquisition

After implemented application is connected to smart phones, if Dump image is extracted and analyzed, items are shown on the screen as shown in figure. Pre-analysis is carried out based on table related to criminal charge. These methods can reduce the entire analysis time. Data such as smart phone basic information, a list of contacts, call list/log, SMS/MMS, memos/events and traces of using the Internet (URL, cookies, and bookmarks) can be found easily using the pattern search.

In matching process, the evidential matter collects object data information continuously from the Crime Scene and periodically sends to the “Send Information” in a binary (digital) data format. Here, send Information indicates the data sender. In offline mode, “Send Information” extracts all feature points from the binary image data. After that it sends these feature data, directly to the “Pattern Modeling and Analysis Tool” for matching and waits for the result.

4 Conclusions

For Smart phone which has a high usability in the scene of criminal accident, it has a merit to contain many data to secure but a demerit to have a high potential to spoil and fabricate data more easily than the general digital evidential matter. Provisions related to the evidence of criminal out of provisions of Criminal Procedure Act Revision in Korean and foreign countries reflect the significance of collecting digital evidences related to the criminal in the initial scene of accident. Therefore, in this Paper, methods to secure criminal related evidences promptly according to clear procedure are presented when the investigator collects evidences related to the criminal in the initial scene of accident by supplementing general digital forensic procedure considering characteristics of Smart phone. We found that it shall be easier to collect data if they search patterns of criminal through the application after establishing a database by creating a criminal charge related table. It is highly recommended to supplement this paper by conducting a professional verification on a criminal charge related table and applying various patterns in future study.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (No. 2012-0004478).

References

1. Korea Internet and Security Agency (2012) The first half of Smartphone Survey 2012
2. Korea Communications Commission (2012) Statistics wired and wireless subscribers for 2012. 7

3. Kang CH (2012) A study on the improvement of search and seizure of digital evidence Ph.D. thesis, Sungkyunkwan University, 2012
4. Gary P, MITRE Corp (2001) A road map for digital forensic research. Report from the first digital forensic research workshop
5. Phillips A, Nance KL (2010) Computer forensics investigators or private investigators: who is investigating the drive? In: Proceedings of the IEEE international workshop on systematic approaches to digital forensic engineering, pp 150–157
6. Lee KW (2012) Data integrity verification method using smartphone forensic. Master's thesis, Ajou University
7. National Institute of Justice (2009) Electronic crime scene investigation: an on-the-scene reference for first responders
8. Chao HC, Peng SH (2011) Research of digital evidence forensics standard operating procedure with comparison and analysis based on smart phone. In: Proceedings of international conference on broadband and wireless computing, communication and applications, pp 386–391
9. Lee JH, Chun WS (2011) Digital evidence collection and analysis research for smartphone forensics. Korea Inst Inform Secur Cryptol 21(6):56–65
10. Lee JH, Park DW (2011) A study of the preparation and procedures by smartphone mobile forensic evidence collection and analysis. Korea maritime information and communication sciences 2011 fall conference, Hoseo Graduate School of Venture, pp 269–272
11. Barrere M, Betarte G, Rodriguez M (2011) Towards machine-assisted formal procedures for the collection of digital evidence. In: Proceedings of the annual international conference on security and trust, pp 32–35

NAC System Analysis and Design for Improvement Model

Seung-Jae Yoo, Jeong-Mo Yang and Hwan-Seok Yang

Abstract Many companies are using a large network, and this helps facilitate information exchange and business corporate. But the use of the intranet cause often the secrets leaked or waste of computer resources of corporate and then this cause the huge social losses. In order to protect the company's high technology and information the companies should take appropriate information security measures for information security themselves. In this paper, we would like to emphasize the necessity of establishing the concept of Network Access Control (NAC) system. For this, we analyze the features of commercial systems, and we propose the enhanced NAC model with its own security features.

Keywords Network access control • Authentication • Access control

1 Introduction

All modern corporations have unique information for the development of the business, which are the company core assets. This information includes the state-of-the-art technology, as well as the information of the employees and the

S.-J. Yoo · J.-M. Yang
Department of Information Security Engineering, Joongbu University,
101 Daehakro, Chugu, Geumsan-gun, Chungnam, South Korea
e-mail: sjyoo@joongbu.ac.kr

J.-M. Yang
e-mail: jmyang@joongbu.ac.kr

H.-S. Yang (✉)
Department of Information Security Engineering, Joongbu University,
101 Majeon-ri, Chubu-myeon, Geumsan-gun, Chungnam, South Korea
e-mail: yanghs@joongbu.ac.kr

companies and individuals involved. In recent years, as well as digital equipment against this information for network and system breaches, including cybercrime, this has been a general increase [1–3].

Order to protect information technology businesses should take appropriate information security measures for information security at the enterprise level.

In this paper, we want to insist on the need to build a network access control system with the sequences, current network system status and its problems, NAC concepts to meet the needs of the users, NAC capabilities from the manufacturer's view, NAC's pros and cons analysis.

2 Related Works

2.1 Network Access Control System

The NAC is network access control technology to minimize security risks, checking whether it compliances user authentication and security policy before the user terminal (end-point) access to the network and NAC technology verifies the authenticity and integrity in accordance with the procedures in the current network management policies.

The procedure method includes the host (user/system) certification, inspection policies, procedures, methods isolated, treatment, and forced. And ultimately NAC technology aims to implement advanced defense network security system to block the security threats that are diffusing.

NAC is a security infrastructure that is able to enforce the security of access to the network being accessed terminals and is a system to protect the system as a whole by blocking the infected with malware PC, laptop, or mobile device to connect to the company network basically.

In the diversified environment as network and terminal device technology development, the expansion of the company's business environment and change the type of attack, we need much more these technologies which are the new foregone security paradigm [4, 5].

2.2 Network System Status and Vulnerability

Commonly it is able to access the unauthorized access to the network by unauthorized users (terminals) in the most of existing network system. It is a serious vulnerability since this means that visitors or partner staffs may access to the internal critical servers [6].

Even if we should also raise the user level security systems such as anti-virus or personal firewall, it can be used to attack by the vulnerability of security management such as not performing the appropriate security patches to the operating system (OS) or not doing update the security products.

Recent prominent types of attacks are attacking to server system or network using the vulnerable users PC. This shows directly how vulnerable the user level is.

Therefore administrators require managing the user level and moreover it needs to be done at the network level for the total protection of network system.

We refer the proper functions of NAC from the administrator perspective (NAC_AdP) and manufacturer perspective (NAC_MaP).

3 NAC Functionality Concepts and Principles

3.1 Functions of NAC_AdP

Some requirements analysis to solve the problems of user authentication, of security policy verification, of risk monitoring and detection, of policy enforcement authority control and of add-ons which appear due to the problems with the existing Network as follows [1–3]:

- All network environmental supports and interlocking with pre-built authentication schemes through user authentication.
- Terminal security system control functions, advisory or prohibited-process control functions, up-to-date antivirus, and the security patches to maintain on latest state through the verification of security policies.
- Terminal/network services using the policy enforcement device and graduating control function according to the user rights selection.
- Abnormal traffic control functions arising from terminals and blocking of the bypassing path to the network through the monitoring and risk detection.
- Monitoring records for user IP allotment, IP management using DHCP, disaster preparedness capabilities and various terminal OS support.

Now, for some NAC commercial models we analyze their functions and compliance the requirements mentioned above.

3.2 Functions of NAC_MaP

In terms of a manufacturer, NAC is defined by the network access control technology to minimize security risks which is checking whether the user authentication and security policies have been complied before the user terminal (end-point) access to the network.

Accordingly, NAC has two major functional elements, authentication and terminal integrity verification (enforcement) and also has the traffic validation function for adding. Also we can see the differences of the perspectives and goals between the two models ‘A’ and ‘B’ in Table 1.

Table 1 Comparison of ‘A’ and ‘B’ models

	Model ‘A’	Model ‘B’
Policy/ verification/ enforcement	○	○
Authentication	○	○
Access control	Wireless LAN/tethering control	–
Users rights management	○	○
Aims	Cover the area of managerial security vulnerability	Combine network security with endpoint security technology

The key aspects of model ‘A’ is focused as followings [2];

- Importance of Policy verification (enforcement) through Security policies enforceable infrastructure,
- Importance of authentication through unauthorized user (terminal) control.
- Users rights management depending on the types.
- Importance of access control via a wireless LAN access control.

In the case of model ‘B’, we find the following four major perspectives [3].

- Unauthorized user access control by internal network management and control (authentication method suitable for the operating network).
- AV and the latest MS Patch install compulsory and shared folders and password policy compliance force through the device control policy compliance control.
- Tracking of internal malicious traffic generating system through Abnormal system detection and isolation and secure the network stability the through the network system isolation.
- Internal network and major system changes management, IT assets and IP management via intelligent IT Infrastructure Management.

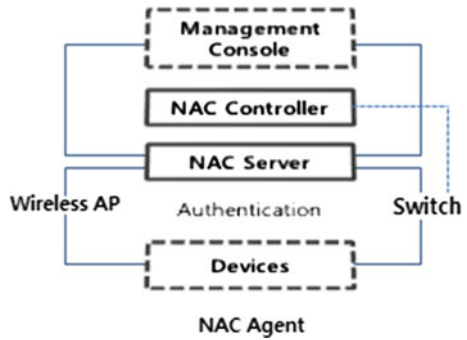
Through these perspectives, we find that they aim to build a strong security response system by the combining network security with Endpoint security technology.

4 Analyze of NAC models

4.1 Models Configuration and Operation Procedures

Model ‘A’ is composed of policy server, block server, Agent and Console. Policy server undertakes the user authentication and NAC policy management, which can be configured for installation location that can communicate with agent if possible (Fig. 1). Block server is a network control device as In-line network equipment,

Fig. 1 Configuration of model 'A'



and to provide access control on the network when it is not 802.1x support environment Agent should be installed on the user PC, receiving NAC policy and enforcement.

And the console opens and runs a Web page with the policy, even if you do not use the Internet server in the ON position, the screen can be seen. And also it is equipment configuration, policy setting, user management, log, statistics, integrated management page.

NAC action process of commercial model 'A' follows the procedures User Authentication, Security Policy Verification, Quarantine, Remediation and ACL.

When the users access to the network, it will be checked the agent installation status. If the agent is installed already, it will immediately attempt to authenticate, but if it is not already installed, the authentication attempt is made after the installation induce. Successful authentication leads to undergo checking the integrity, and comply with the policy before using the network.

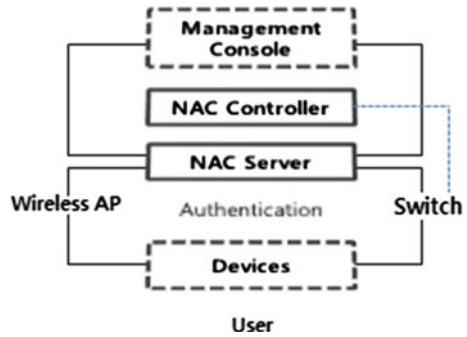
If it fails to comply with the policy, it the network will be able to use after enforcing policies via isolation or warning only. But the network will be blocked in the case of an authentication failure.

NAC action process of commercial model 'B' constitutes the configuration with policy server, blocking server and console (Fig. 2). Policy server manages user authentication and NAC policy management, and it is possible to construct at any location where can communicate with the users. Block server is a network control device as In-line network equipment, and to provide access control on the network when it is not 802.1x support environment.

NAC action process of commercial model 'B' follows the procedures, User Authentication, Security Policy Verification, Quarantine, Remediation and ACL. PC gets out to the (default) gateway via controller (block server). But here these methods are used a configuration of the product without using malicious ARP spoofing, hacking techniques.

Finally console provides a screen where you can learn simple first page. Also it is the integrated management page which controls equipment configuration, policy setting, user management, log statistics.

Fig. 2 Configuration of model ‘B’



In other words, Controller has the gateway MAC value before PC goes to (default) gateway via ARP spoofing. All PC recognize by the gateway to Controller, so have an action process which goes to the controller first.

In these controllers, NAC controls the equipment and user on the network, and then executes the network risk detection. Also it will be responsible for the management and policy settings for each PC on the server.

4.2 Comparison Analysis of NAC Action Processes

We can see some common and difference points of the two models A and B. As common points, they control the network with NAC detection system as defined by the administrator and the network is able to control the source when do not comply the security rule including the personal security solutions. And also prevent to access to internal network by going to the induced page.

As a difference, ‘A’ controls packets even under the NAC Agent not installed, and it is of agent-based. Thus agent installation presence and user authentication procedures should also be seen as one of the basic elements of the NAC. Due to the presence of NAC Agent, it has the audit trail ability of each PC. Also external log server and audit logs of each PC can be linked, and the number of bands of the internal network is not affected because policy inspection and certification inspection using NAC Agent installed on each PC.

In contrast, model ‘B’ contains two components, Controller and Policy Server. And under the NAC security policy violation, user authentication procedures do not think the basic elements of the NAC since there is no NAC Agent. Also it is needed one controller per C Class Network bandwidth for which it controls the internal network bandwidth by bringing each PC’s Default Gateway MAC.

Accordingly if you have a large number of internal networks, additional management costs are incurred by increasing the number of controllers.

In addition, the proper version control of agent makes it easy the additional programs to install, and Windows security patches update via patch management, and can leave a vaccine interlock function and system constraints inspection, Audit

records by showing the status of the PC agent, also makes it easy IP management by sequentially granting for IP hard to manage, and by using the IP records of granted each PC.

Also by isolating the internal problems causing system from the network and removing network problems, it can improve the availability of the internal network, and by the establishment of enterprise network risk management and monitoring system, it can manage the real-time status of network assets and trace management of the change of unauthorized equipment, and IP management.

5 Conclusion

Implementation of the NAC system has the following four major advantages;

- Security Enhancement by Complying User Security Policies,
- Security Enhancement of Internal Main System,
- Availability Improvement of Internal Network,
- Enterprise Network Risk Management and Monitoring System Establishment.

In fact, through enhanced security by the user security policy compliance, it can restrict and control the network access against the unauthorized equipment (user). And it attains the following efficiencies;

- Inside the main asset and information protection from unauthorized parties by enhancing the main system security.
- Enhanced security for internal security by Business-specific network access Ramification. Also by isolating the internal problems causing system from the network and removing network problems, it can improve the availability of the internal network, and by the establishment of enterprise network risk management and monitoring system, it can manage the real-time status of network assets and trace management of the change of unauthorized equipment, and IP management.

As the disadvantage, agent program should be installed on each PC and so increases conflict and the maintenance burden in accordance with Agent installation. Essentially it is complex on demand of 802.1x implementation, and increases the points of failure due to these many installations. Also a network change is inevitable in the in-line equipment and it cause confusion to the user when a device failure.

Based on these, we present the NAC improved future model. And then to redeem the current network status problem, we have to install NAC system which is installed a security program already.

Therefore, as a model for future NAC products, if we enclose multiple endpoint security products in a single product and add it to NAC, then it provides its own security, there is no need to build an additional layer of security solutions even if you install only one product.

Also it is able to access control for network, and even if it is not installed agent, all of the IP-based policies can be applied to a wide range of devices.

If it takes the equipment Out-of-Band placed instead of In-line way, since it is unnecessary to change the network structure, it is needed devices that do not affect the existing network. Also in the enlarged constantly changing complex environment, it will be necessary to improve the NAC model by mixing a variety of options to protect the infrastructures.

References

1. <http://www.symantec.com/ko/kr/>
2. <http://www.unetsystem.co.kr/>
3. <http://www.geninetworks.com/>
4. López G, Gómez AF, Marín R, Cánovas O (2005) A network access control approach based on the AAA architecture and authorization attributes. In: Proceedings of the 19th IEEE international parallel and distributed processing symposium (IPDPS'05)
5. Suzuki S, Shinjo Y, Hirotsu T (2005) Capability-based egress network access control for transferring access rights. In: Proceedings of the third international conference on information technology and applications (ICITA'05)
6. IEEE802 (2001) Port-based network access control, IEEE Std 802.1X-2001

Part II
Industrial and Business
Information Security

Analysis of Methods for Detecting Compromised Nodes and Its Countermeasures

Fangming Zhao, Takashi Nishide, Yoshiaki Hori
and Kouichi Sakurai

Abstract The increased application of sensor network introduces new security challenges. In this paper, we analyze the detection methods of compromised nodes and its countermeasure in the sensor network. We first review common attacks in the sensor network application which can compromise a valid, resource-constrained node (or called device). Then, we introduce several standard detection approaches to show their characteristics in different applications of the sensor network. Finally, we summarize and discuss existing countermeasures to the compromised nodes.

Keywords Sensor network · Security challenge · Detection method · Compromised node · Countermeasure

F. Zhao (✉) · T. Nishide · Y. Hori · K. Sakurai
Kyushu University, 744 Motoooka, Nishi-ku, Fukuoka 819-0395, Japan
e-mail: fangming.zhao@toshiba.co.jp

T. Nishide
e-mail: nishide@inf.kyushu-u.ac.jp

Y. Hori
e-mail: hori@inf.kyushu-u.ac.jp

K. Sakurai
e-mail: sakurai@inf.kyushu-u.ac.jp

F. Zhao
Toshiba Corporate Research & Development Center, 1, Komukai-Toshiba-cho,
Saiwai-ku, Kawasaki 212-8582, Japan

1 Introduction

A sensor network typically consists of one base station and hundreds or even thousands of small, low-cost nodes distributed over a wide area. The base station manages join and leave of nodes, collects data, and broadcasts control information through the Internet to (from) all nodes. Sensor nodes are always designed with the goals of being simple and relatively cheap so that lots of nodes can be deployed in different environments. This has led to these sensors having constraints in terms of low computation, limited memory and bandwidth.

The increased application of sensor network introduces new security challenges, especially related to privacy, connectivity and security management, causing unpredicted expenditure and disaster to the whole system. In this paper, we review common attacks in the sensor network where a node may be compromised, and we introduce several standard detection approaches to show their characteristics in different applications of the sensor network. Finally, we summarize and discuss existing countermeasure for the compromised nodes.

2 Attack Scenarios

2.1 Attack Overviews

Basically, two kinds of attacks exist in the sensor network, (i) *Attack over the network*, (ii) *Attack on the node*.

- **Attack over the network:** this kind of attack mainly intercepts a message between two nodes. Attackers may reside between any two nodes and then they can eavesdrop the message, modify the message or even inject malicious data into the communication channel.
- **Attack on the node:** the attacker mainly targets physical nodes in the sensor network. Attackers always want to leave potential back-door for future's compromises, to extract or to destruct cryptographic keys. If attackers successfully intrude into a node, they can decrypt, observe or even generate fake messages and then send to other nodes.

We summarized several existing attacks over the network and their details in Table 1. Those attacks are mainly targeting two security properties: *Secrecy*¹ and *Authenticity*² [1].

¹ A protocol preserves the secrecy of some term M if an adversary cannot obtain M by constructing it from the outputs of the protocol.

² If the initiator thinks the protocol has been executed with another group member, then it really has, and similarly for a responder.

Table 1 Existing attacks (over the network)

Attacks	Target security properties	Action of attackers	Preventable
Replay attack	Authenticity	Replay or delay data packet	Yes
Man-in-the-middle attack	Secrecy, authenticity	Eavesdrop, packet injection, data modification	Yes
Impersonation attack	Authenticity, secrecy	Impersonate valid user	Yes
Denning-Sacco attack	Authenticity, secrecy	Impersonate valid user by replay past data	Yes

To overcome the *Attack over the network*, we can implement cryptographic primitives to guarantee the authenticity and secrecy of communication between legitimate nodes, e.g. an authentication protocol. Zhao et al. [1] gives an authenticated key exchange protocol to establish and authenticate a session key for both the initiator and the responder: two nodes first generate a common ephemeral secret key (called session key) based on a pre-shared secret key, and they use a cryptographic keyed hash method to authenticate each other's session key, and finally, they can establish a secure communication channel for future's data transmission. On the other hand, *Attack on the node*, which can compromise a remote node, can be launched mainly from physical attackers. Unlike other attacks mentioned above, physical attacks destroy sensors permanently, so the losses are irreversible. For instance, attackers can extract cryptographic secrets, tamper with the associated circuitry, modify programs in the sensors, or replace them with malicious sensors under the control of the attacker. It is very complicated and difficult to detect and prevent such attacks only using security protocols discussed above, since a compromised node may have the secret key of a valid node, and it can authenticate itself to the sensor network. We will discuss several scenarios of compromised nodes in next subsection, and we also summarize detection methods for compromised nodes in Sect. 3.

2.2 Node Compromising Scenario

In this subsection, we mainly review physical attacks which can compromise a valid node in the sensor network. We summarize three main attacks that compromise a valid node depending on the goals the adversary wants to achieve.

- Remove or replacement of nodes. A simple removal or replacement of a valid node can be easily launched by attackers who even lack the knowledge to access the internal system of that node. Because in the sensor network, a node is always placed remotely from the administrator, it is difficult to monitor each node in real time.

This kind of attack can be detected by the base station or another valid node through a periodic remote check, for example, a node authentication protocol [1]

using the node key, to ensure whether the node holds a valid ID and a valid node key which are set up in the node's joining step.

- Modification of the node's internal state. The purpose of this attack can be, for example, to modify the algorithm of its original function or the default configuration data in order to generate fake messages or commands, or, to modify the routing algorithm in order to change the network topology.

If the attacker does not modify the cryptographic keys (node key) or the original node ID, it is difficult to detect such compromised nodes using some cryptographic authentication protocol like [1]. In Sect. 3, we will discuss both a software-based and a hardware-based attestation technique to verify and establish the absence of malicious changes to the node's internal state.

- Access of the node's internal state. This passive attack consists of accessing the internal state of the captured node without changing anything of the node. If any node is compromised by such attacks, the adversary can actually retrieve all of its secret data including cryptographic keys, and then he can analyze the local traffic and eavesdrop all the confidential data through this node.

The remote detection of this passive attack is difficult. One candidate method is sending maintenance staffs periodically to check the remote node. However, such method obviously raises the management cost of the whole sensor network.

3 Existing Detection Approaches

In this section, we summarize existing approaches to detect compromised nodes in the sensor network. To check whether a node is compromised or not can be accomplished via node attestation approach, a process whereby a trusted entity (e.g., a base station or a node) verifies that a node is indeed running the expected application code and, hence, has not been compromised. We introduce both software and hardware based approaches hereafter.

3.1 Software based detection

The software based attestation technique can verify the trustworthiness of resource-constrained nodes, without requiring dedicated tamper-resistant hardware or physical access to the device. It performs challenge-response protocols that can verify the integrity of code memory of a remote node. A node (called prover) computes a checksum of its memory along with a challenge sent from the remote base station (called verifier).

Seshadri et al. [2] proposed the SoftWare-based ATTestation technique (SWATT), a code attestation algorithm that is executed solely through software means. Their technique was designed with the intention of creating a method to

externally verify the code running on embedded devices. A trusted verifier is the key component in achieving this goal of their algorithm. The malicious node will contain at least one line of code that is different from the expected code running on normal sensors. The verifier has a copy of the memory contents residing in valid (uncompromised) nodes. The verifier sends a challenge to the node, which it uses as the input to a pseudo-random generator to create random memory addresses. A checksum is performed in the device on each of these memory addresses. The verifier runs the same verification procedure locally to compute the expected value of the checksum. This expected value is compared to the value returned by the node in question. A compromised node that has altered the memory contents would have to discern whether each memory location created by the pseudo-random generator has been changed.

Their SWATT method works well enough to be used in sensor networks where average nodes are not provided with secure hardware like the TPM. Those compromised nodes which resulted from *Remove or replacement of nodes* and *Modification of the node's internal state* as described in Sect. 2.2, can be detected using such software-based techniques. However, this approach does not work for the compromised node as described in *Access of the node's internal state*.

3.2 Hardware Based Detection

The hardware based attestation technique has stronger security properties and the implementation of such an approach must rely on secure hardware.

TPM is a secure coprocessor that stores an integrity measure of a system according to the Trusted Computing Group (TCG) [3]. Shi et al. [4] is based on the TPM's ability to report the system status to a remote party. This approach is mainly developed for non-resource constrained computer systems and requires all communication partners to perform public key cryptography. The complete system configuration in the Platform Configuration Registers (PCRs) of the attesting entity (a node), must be transmitted to the verifying entity (a base station). Then the verifying entity evaluates the trustworthiness of the attested entity by comparing the received Stored Measurement Log (SML) and PCR values with given reference values. However, in many applications of sensor network, most sensor nodes do not possess enough resources to perform public key cryptography and the transmission of large messages increases the energy consumption significantly.

Similarly to the software based approaches, those compromised nodes which resulted from *Remove or replacement of nodes* and *Modification of the node's internal state* can be detected by such hardware based techniques. This approach does not work for the compromised node as described in *Access of the node's internal state*. This approach relies on a per-device TPM and the availability of a trusted BIOS that begins attestation at boot time. For the sensor network with resource-constrained nodes (devices), the wide implementation of this approach is not quite practical.

4 A Discussion on Countermeasures

After a node is compromised and detected, the system manager (e.g. the base station) must revoke that node as soon as possible to decrease the damage to the whole sensor network. Once a node is revoked, it is no longer capable of reading any confidential message. Other nodes must disconnect all communication channels from/to that node when the compromised node is identified. Since all nodes' communication and data transmission are assumed to be based on the cryptographic protocols, to revoke a compromised node actually means to revoke that node's cryptographic key for encryption, decryption and authentication. The mechanism of revoking a key always depends on the cryptographic algorithm in which the key is used, more specifically, a symmetric key cryptography, or, a public key cryptography. Hereafter, we introduce both revocation mechanisms.

4.1 *Revocation Mechanism for Symmetric Key*

The revocation mechanism for the symmetric key was studied for a long time in the content distribution system. In the content distribution system, the multimedia content can be packaged and distributed to end users in such a way that no one other than a set of privileged users defined by the content owner should be able to use it. Basically, the multimedia content is encrypted based on the symmetric key encryption before being distributed to users. The content encryption is performed by the licensed distributor, and the content decryption is performed by the end user's licensed player. When any player is compromised, the content distribution system must revoke that player's keys to ensure only valid users can use the content but those compromised cannot.

In the content distribution system, traitor tracing [5] is used for the detection of compromised playback devices. It allows the authority to trace piracy back to its original source, and whatever devices are involved in piracy can then be revoked. Traitors are those people/devices that aid pirates in attempting to gain access to an encrypted broadcast digital content. A traitor tracing scheme is employed by the AACs [6]. If a tracing authority determines a playback device has had its device keys compromised, then the compromised device keys can be revoked based on the subset difference revocation scheme from broadcast encryption [7, 8]. The responsibility of broadcast encryption is to facilitate the distribution of multimedia content in such a way that only a predefined set of users are able to play back the content. The subset difference revocation scheme based on broadcast encryption has ability to revoke certain licensed players in the event that these players are compromised by an attacker. Once these players are detected and then revoked, they are no longer capable of decrypting any newly released broadcast. Therefore, the attacker can no longer use the secret information stored in the players to pirate

any newly released multimedia content. Please refer to [7, 8] to for the details of a subset difference revocation scheme.

The broadcast encryption based revocation scheme also works for revoking compromised nodes in the sensor network which utilizes the symmetric key cryptography based security protocols. To apply the revocation scheme to sensor network, each node should get its node key (or device key) from a trusted authority (e.g. in base station) in the node setup phase. Then the trusted authority manages all node keys for all nodes in the sensor network. If any compromised node is detected, for example, by software/hardware detection mechanism we introduced above, the trusted authority can revoke the node key of that compromised node. An example of using broadcast encryption based revocation in the smart grid, an implementation of sensor network, is proposed in [1].

4.2 Revocation Mechanism for Public Key

In public key infrastructure (PKI) [9], certificate revocation schemes are well studied for a long time. A (digital) certificate is a signature by a trusted certificate authority (CA) that securely binds together several quantities. Typically, these quantities include at least the name of a user and its public key. The certificate revocation process starts with a request to revoke or suspend a certificate. A certificate should be revoked if the private key is assumed to be compromised. We discuss the two most common ones: Certificate Revocation Lists (CRL) and the Online Certificate Status Checking Protocol (OCSP).

The CRL is currently the most widespread implementation mechanism for certificate revocation within a public-key infrastructure. A CRL is simply a list of certificates that have been revoked before their intended expiration date. The CA issues this list periodically, together with its signature. Usually, a CRL will be quite long if the CA has many clients. Nonetheless, the complete list is transmitted to any party that wants to perform a certificate check. If the CRL based revocation scheme is applied to nodes in sensor network, two problems should be considered for those nodes that may have low computation, limited memory and bandwidth: (i) the transmission cost of the CRL among nodes, (ii) the computational cost of the verification of digital signature in the CRL.

In OCSP, the CA responds to a certificate status query from a client by generating a digital signature on the certificate's status. This method reduces transmission costs to a single signature per query, but it increases computation costs for both clients and the CA. If the CA is centralized, it becomes highly vulnerable to denial-of-service (DoS) attacks. When the OCSP is applied to sensor network, both the heavy computation cost problem for nodes and the DoS attack for the CA should be considered.

For all countermeasures discussed in Sects. 2 and 4, except for avoiding utilizing heavy computational PKI related process in revocation protocols, the resource-constrained sensor nodes need to decrease/avoid using timestamp and

random number based protocols: (i) Because timestamp is highly depending on the network synchronization accuracy, it is very difficult to synchronize accurate time through all the sensor nodes. (ii) Since the generation of pseudorandom numbers always costs lots of resources of sensor nodes, the decrease of the generation of pseudorandom numbers is preferable.

5 Conclusions

In this paper, we mainly analyzed the detection and countermeasures of the compromised nodes in the sensor network. After summarizing those attacks which can compromise a sensor node, we review both the software based detection method and the hardware based detection method. After a compromised node is detected by such detection methods, as the countermeasures, we analyze two revocation schemes, which are widely studied in digital right management and public key infrastructure, to revoke the cryptographic keys of that compromised node. Finally, we also discuss challenges of both revocation schemes when they are applied to sensor network.

Acknowledgments This research is partially supported by JAPAN SCIENCE AND TECHNOLOGY AGENCY (JST), Strategic Japanese-Indian Cooperative Programme on Multidisciplinary Research Field, which combines Information and Communications Technology with Other Fields, entitled “Analysis of Cryptographic Algorithms and Evaluation on Enhancing Network Security Based on Mathematical Science”. This research is also partially supported by Grants-in-Aid for Scientific Research (B) (23300027), Japan Society for the Promotion of Science (JSPS).

References

1. Zhao F, Hanatani Y, Komano Y, Smyth B, Ito S, Kambayashi T (2012) Secure authenticated key exchange with revocation for smart grid. In: Proceedings of the third IEEE power and energy society conference on innovative smart grid technologies (ISGT), 2012
2. Seshadri A, Perrig A, Doorn L, Khosla P (2004) SWATT: SoftWare-based attestation for embedded devices. In: Proceedings of the IEEE symposium on security and privacy, 2004
3. Trusted Computing Group. <http://www.trustedcomputinggroup.org>
4. Shi E, Perrig A, Doorn L (2005) BIND: a fine-grained attestation service for secure distributed systems. In: Proceedings of the 2005 IEEE symposium on security and privacy, 2005
5. Chor B, Fiat A, Naor M, Pinkas B (2000) Tracing traitors. *IEEE Trans Inf Theory* 46(3):893–910
6. AACS Specifications (2010) Introduction and Common Cryptographic Elements Book Rev 0.951
7. Fiat A, Naor M (1993) Broadcast encryption. In: Proceedings of Crypto 1993. LNCS, vol 773. Springer-Verlag, Berlin, pp 480–49191
8. Naor D, Naor M, Lotspiech J (2001) Revocation and tracing schemes for stateless receivers. In: Proceedings of Crypto. LNCS, vol 2139, Springer, Heidelberg, pp 41–62
9. NIST SP 800-32 (2001) Introduction to Public Key. Technology and the Federal PKI

Protecting Advertisers Against Click Frauds

Rattikorn Hewett and Abhishek Agarwal

Abstract Click frauds are Internet crimes where clicks are deliberately performed to increase the publisher's earnings or to deplete an advertising budget of the advertiser's competitor. This paper presents an approach to automatically detecting click frauds by using a mathematical theory of evidence to estimate the likelihood of the frauds from click behaviors. Unlike most existing work, our approach provides online computation that incorporates incoming behaviors on real-time. It is theoretical-grounded, easily extensible and employs only data available at the advertiser's site. The latter makes the approach feasible for the advertisers to guard themselves against the frauds. The paper describes the approach and evaluates its validity using real-world case scenarios.

Keywords Click fraud detection · Online advertisements · Theory of evidence

1 Introduction

Web advertisements are important sources of revenue for both *brokers* (e.g., Google, Yahoo, MSN, Ask.com etc) who provide the technical platform for online advertisements and *publishers* who display ads on their websites. *Click frauds* are Internet crimes in Web advertisement where click is made with intents of generating “illegal” revenue or causing monetary loss to the competing advertisers [1, 2]. Detecting

R. Hewett (✉) · A. Agarwal

Department of Computer Science, Texas Tech University, Lubbock, Texas, USA
e-mail: rattikorn.hewett@ttu.edu

A. Agarwal

e-mail: abhishek.agarwal@ttu.edu

click frauds is a challenging problem. First, fraud behaviors evolve and continually change over time. Second, both humans and software bots can carry out the frauds. The former requires understanding of human intents. Third, it is difficult to track user identities with their IP addresses as each user's IP address can change over time. Finally, the advertisers typically only have access to the data from their servers but not from the broker's. Thus, they have limited information.

Various techniques have been proposed [3–7]. Most either rely on (1) data that are attainable only by the brokers (e.g., user's activity on publisher site prior to the click), or (2) unrealistic assumptions (e.g., past click behaviors dictate future behaviors), or (3) a batch computation that does not handle new incoming data in real-time and that requires a large historical data set for the results to be statistically valid.

This paper addresses the above issues and proposes an online automatic approach to detecting click frauds using a widely used mathematical theory of evidence, called *Dempster-Shafer (D-S) Theory* [8]. Since our approach employs only data available for the advertiser, it can be useful for both advertisers and brokers. For each incoming click of a given IP, we can estimate our belief of the likelihood whether the click is a fraud or not.

The rest of the paper is organized as follows: [Section 2](#) presents related work and [Sect. 3](#) gives preliminaries including terms, assumptions and an overview of the D-S Theory. [Section 4](#) describes the mass functions, the main contribution of our theory of evidence for click fraud detection. [Section 5](#) evaluates the approach using data from a real-world scenario. The conclusion of the paper is given in [Sect. 6](#).

2 Related Work

Click fraud detection has only been researched recently [3–7]. Google's approaches to detecting invalid clicks include rule-based, anomaly-based and classifier-based techniques [4]. The first two rely on expertise knowledge on known abnormal patterns or conditions to detect malicious acts. Thus, they are unable to detect unseen cases. On the other hand, the classifier-based approach uses historical data of user click behaviors and unrealistic assumptions. It also tends to have low accuracy as fraud mechanisms change over time. Google's approach also uses the datasets that are only at the disposal of brokers. Our proposed approach does not require such assumptions and privileged data.

A burst detection algorithm [3] focuses on high frequency user activity in short time periods but it does not handle other click fraud patterns. Work in [5] uses bluff ads for detecting sources of click frauds. For example, a click by an IP address in Australia on a bluff ad of a special offer on pizza in New York City should indicate a fraud possibly by a bot or a naive fraudster. However, this approach requires the broker implementation and may not work for sophisticated frauds.

The D-S Theory of evidence is applied in [7] to a click fraud detection system. However, the approach focuses on using the D-S Theory for combining evidence resulted from other approaches, i.e., rule-based, outlier detection, and click map. Our approach, however, computes a belief of evidence by using the traditional approach of formulating mass functions in the D-S theory. The approach also requires data from the client's site server, which may not be available (e.g., if cookies are disabled). It also maintains an online database of suspicious click sources. Unfortunately, the paper does not provide sufficient details or sources that we can run or replicate the system. Therefore, a comparison study is infeasible.

3 Preliminaries

3.1 Web Advertisement Terms

In web advertisement, an advertiser pays for his ads to be displayed in order to sell his products or services. These ads may create more traffic and revenue for the advertisers. An *ad-site* is the advertiser's website, which may contain internal links to multiple pages. An Internet *user* (or *agent*) can visit the ad-site via Internet search, URL on the advertiser's browser, bookmark, or the ad on a publisher site. An *ad-visit* is a visit of a user to ad-site by clicking an ad and otherwise it is a *non-ad visit*. A *session* is a continuous period of time that a user navigates within the advertiser's site where a user maintains an active HTTP connection with the server. *Publishers* are the websites that host ads for the advertisers. The publishers get paid for each click on the ads they host. A *broker* provides the technical platform for online advertisements including Internet search engine (e.g., Google, Yahoo, AOL, Ask.com) and geographical location etc. Google runs a pay-per-click program called *AdSense* using *Gclid*, which is a unique user ID corresponding to the server log for each click made on Google ads.

3.2 Assumptions and Limitations

This section discusses necessary assumptions and limitations. First, we assume that a user who makes a purchase is not a fraudster. This is reasonable because a fraudster who makes a purchase to confuse a detection system will likely find that his action does not pay off. Second, to cope with the fact that IP addressing changes over time, our approach is designed to identify click frauds per window of period W . This provides the flexibility to analyze click frauds in various window sizes. In general, the change of IPs makes it near impossible to detect identity of fraudsters over a long period of time. In this paper, W is specified to be 30 min, which is a small enough window that the user IP address will not get changed

unless it is deliberately done by the user to hide his/her IP (e.g., proxy IPs via proxy servers). In such a case, the frequency of these malicious acts is likely to be low for a sufficiently small window size.

Third, to estimate user session durations, we sum the durations between consecutive HTTP requests and assume 30 s spent on the last page. This is due to the fact that HTTP is a stateless protocol and thus, it is difficult to accurately estimate the session duration. Lastly, our approach assumes a specific pay-per-click model similar to Google's *AdSense* and the usage of *gclid*, to uniquely identify clicks on the ads in addition to an IP address. This follows Google's approach to make distinction among visits on different ads. It also automatically eliminates redundant clicks (e.g., double ad-clicks is counted as one). In addition, *gclid* allows us to separate ad-clicks and non-ad clicks. Note that this assumption is introduced only to relate our approach to realistic online advertising. It does not limit our approach in terms of its applicability to other payment model and other user IDs.

3.3 Mathematical Theory of Evidence

This section describes the *Dempster-Shafer (D-S) Theory* [8] for uncertainty reasoning that assigns a probability to a *set* of atomic elements (or hypotheses) rather than an atomic element alone. The theory is a generalization of Bayesian probability theory. It is used widely to combine evidences from multiple sources to support the *belief* of a specific hypothesis.

Let U be a finite set of all hypotheses (atomic elements) in a problem domain. A *mass function* m provides a probability assignment to any $A \subseteq U$, where $m(\emptyset) = 0$ and $\sum_{A \subseteq U} m(A) = 1$. The mass $m(A)$ represents a belief *exactly* on A . In the context of click fraud detection, $U = \{fraud, \sim fraud\}$ represents a set of two hypotheses of a click of being a fraud or not a fraud, respectively. Elements of the power set of U representing $A \subseteq U$ are all possible hypotheses, namely $\{fraud\}$, $\{\sim fraud\}$, U , and \emptyset . The belief of each of the hypotheses lies between 0 and 1. The property of the mass function implies that $m(\{fraud\}) + m(\{\sim fraud\}) + m(\{fraud, \sim fraud\}) = 1$ since $m(\emptyset) = 0$. When there is no information regarding U , $m(\{fraud\}) = m(\{\sim fraud\}) = m(\emptyset) = 0$ but $m(\{fraud, \sim fraud\}) = 1$. The last case includes all possible hypotheses and therefore, its truth is believed to be certain.

For every mass function, there are associated functions of *belief* and *plausibility*. The degree of belief on A , $bel(A)$ is defined to be $\sum_{X \subseteq A} m(X)$ and the plausibility of A , $pl(A)$ is $1 - bel(\sim A)$. For example, $bel(\{fraud\}) = m(\{fraud\}) + m(\emptyset) = m(\{fraud\})$. In general, $bel(A) = m(A)$ for any singleton set $A \subseteq U$ and in such a case the computation of *bel* is greatly reduced. In this paper, we use the terms *likelihood* and *belief* synonymously.

A mass function can be combined for multiple evidences using the *Dempster's Rule of Combination*, which is a generalization of the Bayes' rule. For X, A ,

$B \subseteq U$, a combination mass functions m_1 and m_2 , denoted by $m_1 \oplus m_2$ (or $m_{1,2}$) is defined as the following:

$$m_{1,2}(X) = m_1 \oplus m_2(X) = \frac{\sum_{A \cap B = X} m_1(A)m_2(B)}{1 - K}$$

$$\sum_{A \cap B = \phi} m_1(A)m_2(B)$$

where $K = m_1 \oplus m_2(\emptyset) = 0$.

The combination rule can be applied in pairs repeatedly to obtain a combination of multiple mass functions. The above rule emphasizes the agreement between multiple sources of evidence and ignores the disagreement with a normalization factor. See more details in [8].

4 Proposed Approach

This section describes our main contribution, the development of mass functions of corresponding types of evidences. For mass function m_i of evidence type i , α_i and β_i represents the strength of the evidence to support the belief and disbelief of a fraud, respectively. These values range between zero and one, and are commonly obtained by experiments.

Evidence based on number of clicks on the ad

For a given IP address in a given window, if the number of clicks on the ad is high then the likelihood of this user being a fraud is high. This can be referred to short bursts. The mass function m_1 of this evidence supports a *belief* of *fraud* where the belief value depends on the number of clicks. Let n be a total number of clicks during the window period. Thus, the likelihood of each click being a fraud is $1 - 1/n$ and we have:

$$m_1(\text{fraud}) = \alpha_1(1 - 1/n)$$

$$m_1(\sim \text{fraud}) = 0$$

$$m_1(U) = 1 - m_1(\text{fraud}) = 1 - \alpha_1(1 - 1/n).$$

Evidence based on time spent on browsing

If the time spent by a user at the *ad-site* is high then the user is less likely to be a *fraud*. Fraudsters are likely to have smaller sessions since they are not interested in the product. As a user continues to spend more time at the *ad-site*, the belief that he/she is *~fraud* will increase. Thus, the mass function m_2 of this evidence support *~fraud* where the belief value depends on the time spent at the *ad-site*. Let t be the time spent by a given user in all visits in the time window W . Thus, $0 < t \leq W$. The mass function m_2 can be computed as follows:

$$m_2(\text{fraud}) = 0$$

$$m_2(\sim \text{fraud}) = \beta_2(t/W)$$

$$m_2(U) = 1 - m_2(\sim \text{fraud}) = 1 - \beta_2(t/W).$$

Evidence based on non-ad visits followed by ad-visits

If a user clicks on an ad after having visited the *ad-site* via direct URLs (i.e., *non-ad visit*), then the user is likely to be a *fraud*. Consider all types of visits of a given user in a window period, the mass function m_3 of this evidence can be formulated in three cases: *non-ad visits without following ad-visits*, *non-ad visits followed by ad-visit*, and *ad-visits only*. We formulate the mass functions below. Let m and k be a number of *non-ad visits* and *ad-visits* so far, respectively. Let p be a number of *ad-visits* after *non-ad visits*, q be a number of visits after the first *non-ad visit*, and d be the strength of case 3 (e.g., d is set to 10 in our experiment).

case 1: Non-ad visits without following ad-visits.

$$m_3(\text{fraud}) = 0, m_3(\sim \text{fraud}) = \beta_3 * m/(m+k), \quad \text{and } m_3(U) = 1 - \beta_3 * m/(m+k)$$

case 2: Non-ad visits followed by at-least one ad-visit.

$$m_3(\text{fraud}) = \alpha_3 * p/q, m_3(\sim \text{fraud}) = 0, \quad \text{and } m_3(U) = 1 - \alpha_3 * p/q$$

case 3: Ad-visits only.

$$m_3(\text{fraud}) = \alpha_3 * k/(d * (m+k)), m_3(\sim \text{fraud}) = 0, \quad \text{and } m_3(U) = 1 - \alpha_3 * k/(d * (m+k))$$

Evidence based on place of click origin

If a click occurs at a location outside the target business region of the advertiser then it is likely to be a *fraud*. For example, a car dealer company will advertise certain models to customers in neighboring cities. The click from different country will likely be a fraud. The mass function m_4 for this evidence supports *fraud* if the click occurs outside advertiser's business region, and $\sim \text{fraud}$, otherwise. We define X to be one if location of the click is in a target region of the advertiser's business, otherwise zero. The mass functions are:

$$m_4(\text{fraud}) = \alpha_4 * X$$

$$m_4(\sim \text{fraud}) = \beta_4 * (1 - X)$$

$$m_4(U) = 1 - m_4(\text{fraud}) - m_4(\sim \text{fraud})$$

The strength of the mass functions can be determined via experiments that systematically refine the parameter values to fit with observations. Due to space limitation, we omit this part and only report results. In this paper we use $\alpha_1 = 0.8$, $\beta_2 = 0.99$, $\alpha_3 = 0.6$, $\beta_3 = 0.2$, $\alpha_4 = 0.4$, $\beta_4 = 0.1$, respectively. The combination rule of the D-S theory is then applied to the belief values until all evidences have been incorporated into the likelihood estimation of each hypothesis. The final step

of our detection system is to make a final *fraud certification* based on the belief values obtained. We use simple thresholds to classify combined belief values into three levels: *~fraud* (<0.5), *suspicion* (>0.5 and <0.65), and *fraud* (≥ 0.65).

5 Evaluation

5.1 Experimental Data

To evaluate our proposed approach, it is necessary for us to use synthetic data for several reasons. Most importantly, it would be hard to determine if the proposed approach produces correct results even with real weblog data since we have limited resources to validate the correctness of the findings. Furthermore, the click data are typically maintained and monitored privately. However, one can verify if the click detection works correctly by working backward on synthesized data with known outcomes (e.g., by injecting click fraud behaviors to test whether the detection system can identify the frauds as expected).

We evaluate our approach using a data set representing a scenario that contains 40 clicks in a window of 30 min. Clearly, analysis of consecutive windows can be performed in the same fashion. For each user, we pre-process the raw web server logs and extract relevant data in real time including (1) IP address of the remote computer requesting the web page; (2) time and date of request; (3) the requested page; (4) the referrer page, and (5) the Gclid. The region from which the click originated can be extracted from the IP address by using one of the many geo location services.

5.2 Case Study Scenario

Consider a scenario of click activities of a sophisticated fraudster in a 30-minute window from 1:45 am to 2:15 am as shown in Fig. 1. A user makes ad-visit in three brief sessions of a few minutes (each indicated by a dark grey block). In each session, he clicks once on the ad and performs other activities (e.g., browsing, open links, etc.) Then he makes a non-ad visit session (indicated by a light grey block), where he logs in and puts an item in a shopping cart (annotated by icons in Fig. 1). After that the user makes another four ad-visit sessions, each of which he clicks on the ad. It is obvious that in the first three sessions, each click in a short time makes him look increasingly suspicious. However, the suspicion of fraud decreases in the next session where he spends more time at the site signing on as a registered customer and putting an item in the cart. These behaviors conform to a genuine customer rather than a fraudster. However, the fact that his shopping activity is in a non-waking hour (especially, if the IP is invoked from non-target

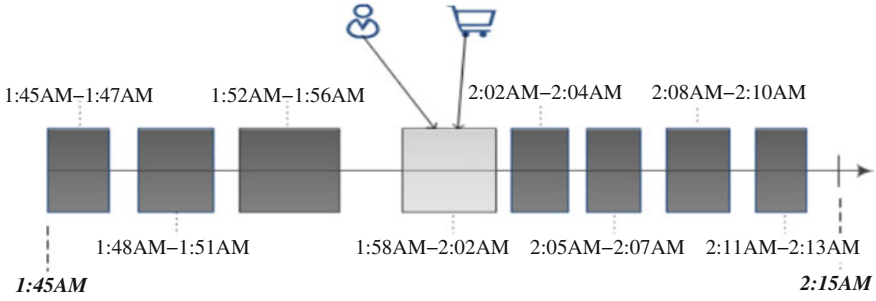
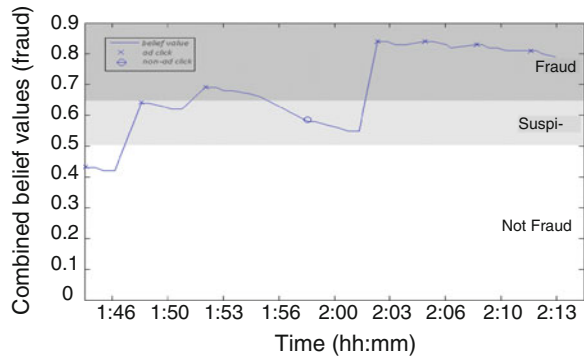


Fig. 1 Click activities between 1:45 and 2:15 am

Fig. 2 Belief values of click fraud behaviors over time



customer regions or countries) is conflicting evidence that supports the contrary. Moreover, his next four clicks on the ad via ad-visits (after he already knew how to get to the commercial/advertising site via a direct URL) make him becomes highly suspected as a fraudster.

Applying a combination rule from the D-S theory to belief values computed from the mass functions of evidences introduced in Sect. 4, Fig. 2 shows the final belief values obtained. It is clear that the belief values are dynamically changing as new incoming evidence is incorporated into the estimation of fraud likelihood.

As shown in Fig. 2, initially, the belief of fraud after the first click is a little over 0.4, which is believed to be *~fraud*. After second and third clicks, the belief of fraud increases and the detection decision moves from *~fraud* to *suspicious* and to *fraud*, respectively. In the fourth visit, the user does a non-ad visit and thus, the belief value drops back giving a certification of *suspicion*. Had the user stopped clicking on the ad at this point, this would have been the decision about the user. However when the user clicks on the ad again via ad-visit in the next few visits, the belief increases to support *fraud*. The belief of fraud continues to be high and this is certified as a case of *fraud* as a final decision. The results obtained agree with our reasoning. We can validate that the results obtained are consistent with our conclusion observed earlier. Thus, in this scenario, the approach works correctly.

6 Conclusion

This paper presents an automated approach to detecting click frauds that is based on a well-grounded theory for novel detection of unanticipated patterns. It uses on-line computation, which can adapt well to real-time systems and is effective in analyzing each incoming click at the advertiser's disposal. Furthermore, the approach does not require advertisers to maintain large historical databases. This makes the approach beneficial for use by advertisers. Finally, the framework is flexible in that new mass functions can be added for refinement.

Future work includes experiments to gain understanding of the characteristics of the proposed approach, including how window sizes impact the detection capability, whether our approach works for specialized bot attacks.

References

1. Kshetri N (2010) The economics of click fraud. *IEEE Secur Priv* 8(3):45–53. http://libres.uncg.edu/ir/uncg/f/N_Kshetri_Economics_2010.pdf
2. Li X, Liu Y, Zeng D (2011). Publisher click fraud in the pay-per-click advertising market: incentives and consequences. In: Proceedings of the IEEE international conference on intelligence and security informatics, pp 207–209
3. Antoniou D, Paschou M, Sakkopoulos E, Sourla E, Tzimas G, Tsakalidis A, Viennas E (2011) Exposing click-fraud using a burst detection algorithm. In: Proceedings of the ISCC on computers and communications, IEEE symposium, pp 1111–1116
4. Tuzhilin A (2006) The Lane's gifts vs. Google report. http://googleblog.blogspot.com/pdf/Tuzhilin_Report.pdf
5. Haddadi H (2010) Fighting online click-fraud using bluff ads. *ACM SIGCOMM Comput Comm Rev* 40(2):21–25
6. Walgampaya C, Kantardzic M, Yampolskiy R (2010) Real time click fraud prevention using multi-level data fusion. In: Proceedings of the world congress on engineering and computer science, San Francisco, 20–22 Oct 2010
7. Walgampaya C, Kantardzic M (2011) Cracking the Smart ClickBot. In: Proceedings of the 13th IEEE symposium on web systems evolution, pp 125–134
8. Shafer G (1976) *A mathematical theory of evidence*. Princeton University Press, Princeton

A Novel Intrusion Tolerant System Based on Adaptive Recovery Scheme (ARS)

Seondong Heo, Jungmin Lim, Minsoo Lee, Soojin Lee
and Hyunsoo Yoon

Abstract Nowadays, as many information systems are connected to Internet and provide useful services to people through Internet, this openness makes the systems as targets of attackers. Even though conventional security solutions such as intrusion detection system (IDS) or firewall were designed to protect such attacks, it is impossible to block all the attacks. The researches on intrusion tolerant system (ITS) have been conducted in order to keep the proper services in the threatening environments. In this paper, we propose a novel Adaptive Recovery Scheme (ARS) which can be applied to intrusion tolerant architecture. ARS has proactive recovery scheme and reactive recovery scheme including self-recovery and emergency recovery. ARS selects appropriate recovery scheme according to internal and external factors to maintain required security and performance level. Additionally, ARS protects an integrity of critical files through snapshot technology. The performance of ARS is compared with existing recovery-based intrusion tolerant system by CSIM 20.

Keywords Intrusion tolerant system (ITS) · Adaptive recovery scheme (ARS) · Virtual machine (VM) · Performance

S. Heo · J. Lim · M. Lee · H. Yoon (✉)
Department of Computer Science, Korea Advanced Institute of Science and Technology,
291 Daehak-ro, Yuseong-gu, Daejeon 305-701, South Korea
e-mail: sdheo@nslab.kaist.ac.kr

S. Lee
Department of Defense Information Science, Korea National Defense University,
33 Je2Jayuro, Goyang-si, Gyeonggi-do 412-706, South Korea

1 Introduction

Internet has been widespread as a platform supplying various services. Most of useful services such as banking, social network, and search are provided through Internet. This openness of information system offers convenience to users, while it gives chances to intrude or compromise the system for attackers. Vulnerabilities of the systems have been increasing due to the complexity of the systems, thus attackers can easily discover vulnerabilities which might be exploited in a malicious ways.

Traditional protection solutions such as an intrusion detection system (IDS) or an intrusion prevention system (IPS) needs information about attacks to defense. However it is almost impossible to find out all the information ahead of attacks, so ITS takes different approach to protect information system. Main focus of ITS is to provide services in the face of intrusions [1].

Even though there are many studies on ITS, recovery-based approach [2–4] using virtual technology have been the most actively studied in recent years. Most of those researches present proactive recovery mixed with reactive recovery in order to build a secure system. Proactive recovery can remove invisible intrusions while reactive recovery is useful to restrict compromised node's malicious activity. However, both are not considering the performance of the system when active node stops working by reactive recovery.

In this paper, we propose a novel ITS based on ARS whose security and performance is critical. ARS concentrates on maintaining performance of the system in the threatening environments. It periodically conducts proactive recovery to eliminate intrusions and proper reactive recovery depending on circumstances. We present two reactive recovery schemes, compared to existing scheme, which terminates the compromised VM immediately: self-recovery and emergency recovery. In ARS, VM whose critical files are modified or deleted by attacks can be recovered by self-recovery, and it doesn't require halting the VM. And emergency recovery is utilized when VM must be cleansed, but halt of VM causes degradation of the system performance. Experiment with CSIM20 [5] verifies effectiveness of these schemes, which significantly decrease the effects of attacks to system performance. The contributions of our work are as follows:

- ARS can maintain required level of performance and security with multiple recovery schemes: proactive recovery, reactive recovery including self-recovery and emergency recovery.
- Critical files of VMs are protected by a snapshot. Attacks that compromise the integrity of the files are almost detected by snapshot comparison. In addition, deleted or modified files can be repaired by self-recovery.

The rest of this paper is organized as follows: [Section 2](#) introduces related works on ITS from 1990s to present. [Section 3](#) presents ARS and its recovery schemes including state transition. In [Sect. 4](#), we present performance test result with CSIM 20. This paper ends with a conclusion in [Sect. 5](#).

2 Related Work

Researches on intrusion tolerance architecture can be classified into three types depending on the centric scheme: middleware-based ITS, hardware-based ITS, and recovery-based ITS. Both Middleware-based architecture [1, 6] and hardware-based architecture [7, 8] use IDS or detection techniques to find intrusions. But it is impossible that detecting all kinds of attacks. To solve this problem, recovery-based architecture performs proactive recovery periodically to eliminate hidden attacks. The Self-Cleansing Intrusion Tolerance (SCIT) researches [2] use virtualization techniques to implement proactive recovery. There are four states in SCIT architecture: active, grace, cleansing, and live spare. Central controller executes state transition of all VMs. In SCIT, central controller must be separated from VMs which are in active or grace state to prevent stealth attack. SCIT-HES [3] proposes SCIT enhanced with hardware which completely prevents communication between VMs exposed to external network to central controller.

P. Sousa et al. [4] proposed the architecture using proactive and reactive recovery. In this architecture, each replica consists of two parts: payload and wormhole. Payload is the part to provide services, and wormhole performs communication between VMs and reactive recovery in order to detect intrusions. Proactive recovery is done in a manner similar to the SCIT, in that all replicas are periodically rejuvenated. Reactive recovery is activated when replica j receives two crucial variables used in this research: $W_suspect(j)$ or $W_detect(j)$. $W_suspect(j)$ means that the replica j is suspected of being faulty and $W_detect(j)$ means that the replica j is faulty without doubt. If replica j receives a detection message from $f + 1$ different replicas, the replica j is recovered immediately. Otherwise, if there are $f + 1$ suspicions, recovery must be scheduled with the periodic proactive recoveries in order to guarantee system availability.

Proposed system uses reactive recovery and central controller by the following reasons: ① System requires more resources to employ reactive recovery, but it can limit compromised VM's action to disturb the system's operation. ② Using central controller can enforce VM's recovery strongly because central controller is a hypervisor of all VMs and owns every permission on recovery. ③ It can reduce communication cost for snapshot transmission. Thus, we introduce reactive recovery scheme for guarantying availability and performance.

3 ARS (Adaptive Recovery Scheme)

3.1 System Architecture

The architecture of the system is illustrated in Fig. 1. States of all VMs are managed by the central controller. The central controller is responsible for states of each VM, system performance and responding to attacks. In normal case, VM's

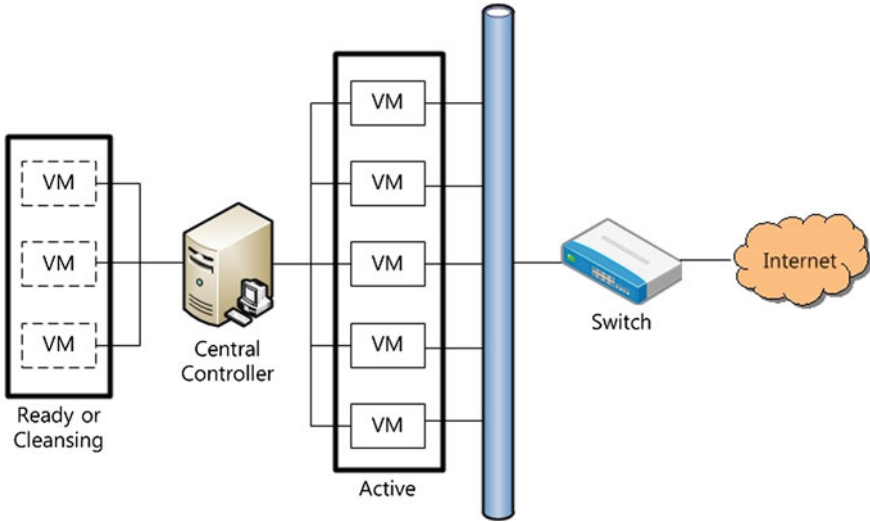


Fig. 1 System architecture

state is periodically switched by the central controller: active \rightarrow cleansing \rightarrow ready \rightarrow active [2]. When VM changes its state from ready to active, VM takes a snapshot of its critical files and sends it in the central controller. And in active state, VM takes a snapshot of it periodically and sends it to central controller through internal network for checking integrity of critical files. Each VM supplies services and processes requests incoming from external network. All requests are randomly delivered to each VM. The following assumptions are applied to the design of the architecture.

- The internal network is completely separated from the external network, so it is impossible to sniff or attack central controller directly from outside.
- All states of VMs are controlled by the central controller. In other words, the central controller can change all VM's states whenever it needs, though VM cannot receive requests due to an intrusion.
- If VM doesn't send its snapshot to the central controller within the designated threshold, controller regards it as a compromised VM.

3.2 State Transition

Figure 2 shows state transition of a VM in ARS. Exposure time is a predefined value which means how long each VM is exposed to the external network and provides services.

- Active: a VM is exposed to the external network and provides services. VM periodically sends its critical file's snapshot to the central controller. If a VM is

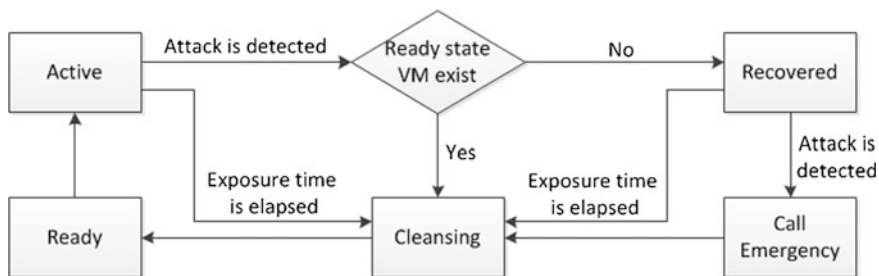


Fig. 2 State transition diagram

attacked, the central controller checks if there is a VM in ready state. If there is a one, central controller changes its state to active while changing corrupted VM’s state to cleansing state. Otherwise, the central controller sends a snapshot to the compromised VM and changes its state to recovered state.

- Cleansing: a VM is offline. Central controller restores the state of VM to pristine state.
- Ready: a VM is offline. It is ready to provide services.
- Recovered: a VM utilizes snapshot received from central controller to recovery. If attack is detected in recovered state, central controller changes a VM’s state to cleansing, and conducts emergency recovery.

3.3 Proactive Recovery

Table 1 shows the parameters associated with ARS. In ARS, central controller recovers VMs in active state by exposure time in order to remove undetected attacks. Exposure time (T_e) can be derived by T_c , N_{max} and N_{min} as the Eq. (1).

$$T_e = \frac{T_c \cdot N_{min}}{N_{max} - N_{min}} \tag{1}$$

And the probability that the attack is not occurred in time T is $P(X = 0) = e^{-\lambda T}$. Therefore, the probability that VM is attacked at least once in T_e becomes: $P(X > 0) = 1 - e^{-\lambda T_e}$. Since $e^{-\lambda T_e}$ is a monotonically decreasing function of T_e , small T_e is required in order to make more secure system. But T_c and N_{min} are constants, so increasing N_{max} is the only way to decrease T_e .

Table 1 List of parameters used in ARS

Parameter	Description
T_e	Time duration in which VM is exposed to external network
T_c	Time duration in which VM cleanses
N_{max}	Number of VMs can be operated
N_{min}	Number of VMs required for performance
λ	Poisson parameter of attack successes among incoming requests

3.4 Reactive Recovery

Proactive recovery can eliminate stealth attacks properly, but constructing a secure system only with proactive recovery is a costly method. So recent researches are focused on reactive recovery. Even though additional resources are required to implement reactive recovery, it is much smaller than resources required for building a system that has the same security level only with proactive recovery.

There are three different reactive recovery schemes in ARS depending on circumstances. First, if the central controller detects an attack on a VM which was not discovered before, the central controller examines existence of a VM in ready state. If ready state VM exists, the central controller changes its state to active state and set T_e to restored VM's remaining exposure time. Otherwise, in second case, the central controller sends snapshot to a compromised VM, then the VM utilizes it to recover itself. Even though a VM may have its snapshot before an attack, it is unreliable. The central controller is separated from the external network, so snapshot from the central controller is much more credible. Third, if the central controller detects an attack on a VM in a recovered state, central controller conducts an emergency recovery. In this case, central controller changes state of attacked VM to cleansing and increases exposure time by $T_c/(N_{max} - N_{min})$ of active VMs that have less exposure time than attacked VM in order to reduce performance degradation. ARS applies emergency recovery, so it can rapidly regain performance and availability as compared to existing systems.

4 Simulation and Results

4.1 Experimental Environment

The CSIM 20 simulator was used for estimating efficiency of ARS. CSIM is a simulation toolkit, which consists of simulation engine and a comprehensive set of tools that can be used to develop realistic models of complex systems. By using CSIM 20 simulator, we have simulated ARS in several conditions for evaluating performance with C++ language in Visual studio 2008.

We generate requests at intervals of 0.2 s and set processing time to 0.9 s. The request generation time and processing time follow the exponential distribution with the mean value 0.2 and 0.9, respectively. λ is set to 0.00002, and successful attack halts the operation of VM. Consequently, attacked VM cannot receive requests from the external network. We assumed that all the attacks will be detected because intrusion detection is not a main work in this paper. The probability of successful self-recovery is set to 0.5. Other values used in experiment are shown in Table 2.

Table 2 Values used in experiment

Parameter	Value	Parameter	Value
T_e	600 s	N_{min}	5
T_c	120 s	λ	0.00002
N_{max}	6		

Table 3 Response time and queue length

	Normal (not attacked)	Attacked (attacked)	ARS (attacked)
Response time (sec)	7.91	11.88	8.12
Queue length	6.59	9.90	6.75

4.2 Results

Table 3 shows the results of simulation with million requests. ‘Normal’ system wasn’t attacked, and ‘attacked’ system is the system that employs only proactive recovery or uses conventional reactive recovery which restores compromised VM immediately. Response time of ARS is slightly increased by 2.7 %, but it is very small compared to the response time of attacked system which is increased by 50.2 %. In simulation, we don’t restrict max queue length so there is no packet drop by queuing problem. But in the real world, requests sent to the attacked system can be dropped by timeout or full queue.

In order to emphasize the impact of an attack, we artificially send one attack packet in every thousand requests with other same parameters. Figure 3 illustrates the result of simulation with 10 thousands requests. Attacks are represented as dashed line. As seen in Fig. 3, ARS reduces the impact of attacks with same environment. If self-recovery succeeds in correct recovery, response time doesn’t increase in spite of attacks.

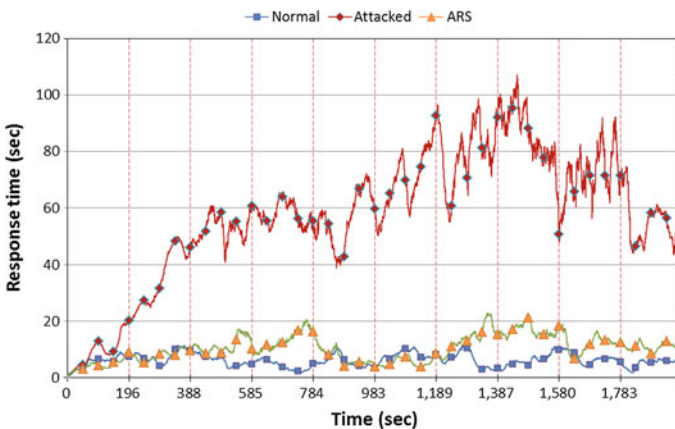


Fig. 3 Response time

5 Conclusion

Recent researches on ITS have been focused on the proactive and reactive recovery. However, there are little considerations on maintaining performance and availability when active VM is recovered by reactive recovery.

In this paper, we introduce a novel intrusion tolerant architecture based on ARS which runs adaptive recovery schemes according to the internal and external factors: proactive recovery, reactive recovery including self-recovery and emergency recovery. It has been shown that ARS can significantly reduce the impact of attacks that affects performance and availability of system. Especially, ARS maintains a desired level of performance under the unpredictable savage attacks.

Acknowledgments This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the CYBER SECURITY RESEARCH CENTER supervised by the NIPA (National IT Industry Promotion Agency), NIPA-H0701-12-1001.

References

1. Wang F, Gong F, Sargor C, Goseva K, Trivedi K, Jou F (2001) Scalable intrusion tolerance architecture for distributed server. In Proceedings of the second IEEE SMC information assurance workshop, 2001
2. Huang Y, Sood A (2002) Self-cleansing systems for intrusion containment. In: Proceedings of workshop on self-healing, adaptive, and self-managed systems (SHAMAN), 2002
3. Arsenault D, Sood A, Huang Y (2007) Secure, resilient computing clusters: self-cleansing intrusion tolerance with hardware enforced security (SCIT/HES). In: Proceedings of the second international conference on availability, reliability and security (ARES 2007), 2007
4. Sousa P, Bessani AN, Correia M, Neves NF, Ver'issimo P (2010) Highly available intrusion-tolerant services with proactive-reactive recovery. *IEEE Trans Parallel Distrib Syst* 21(4):452–465
5. Schwetman H (2001) CSIM19: a powerful tool for building system models. In Proceedings of the 2001 winter simulation conference, pp 250–255
6. Saidane A, Nicomette V, Deswarte Y (2008) The design of a generic intrusion tolerant architecture for internet servers. *IEEE Trans Dependable Secure Comput*, 2008
7. Just JE, Reynolds JC (2001) HACQIT (Hierarchical adaptive control of QoS for intrusion tolerance). In: Proceedings of the 17th annual computer security applications conference, 2001
8. Chong J, Pal P, Atighetchi M, Rubel P, Webber F (2005) Survivability architecture of a mission critical system: the DPASA example. In: Proceedings of the 21st annual computer security applications conference, pp 495–504

Design and Implementation of Linked Network Security System Based on Virtualization in the Separate Network Environment

Dong-Hwi Lee and Kyong-Ho Choi

Abstract In this study a Linked Network Security system based on Virtualization (LNSV) is proposed to effectively perform data transmissions in a network separated environment under the aspects of management, operation, and cost. The LNSV proposed in this study represents an open architecture in accessing its system through network connectors for all users in individual networks and can be used as a general purposed system for storing all data to be transmitted. It is possible to prevent the access of unauthorized users because the stored data files include source IP/PORT, destination IP/PORT and Hash Values. Also, it can guarantee the security of communication through transmitting and receiving data using encryption/decryption functions. Thus, the LNSV can provide safe connection services between separated networks.

Keywords Access control · Network security · Hypervisor · Virtualization · Separate network

D.-H. Lee

Department of Industrial Security, Kyonggi University, San 94-6, Iui-Dong, Yeongtong-Gu, Suwon-Si, Gyeonggi-Do, South Korea
e-mail: dhclub@naver.com

K.-H. Choi (✉)

Center for Industry, Kyonggi University, San 94-6, Iui-Dong, Yeongtong-Gu, Suwon-Si, Gyeonggi-Do, South Korea
e-mail: cyberckh@gmail.com

1 Introduction

The present day is an age that promotes works using computers and networks and creates high value added businesses. Governments, enterprises, and individuals have already accepted cyberspace as an essential element for their works and lives. Thus, they have much interests on defending various attacks in this cyberspace [1]. In spite of much interests on this issue, however, there have been lots of threats and that increasingly occurs serious damages. Thus, studies on the issues of intelligent phishing attacks [2], attacks on important facilities using Stuxnet [3], and leaks of internal information [4] and its countermeasures have been largely conducted.

In addition, Intranet and Internet can be separately operated to protect important information. In this case, an extra data transmission architecture is required because the communication between these separated networks cannot be implemented. Figure 1 shows an example of this case in which a Windows based system updating method is presented under the separated network environment guided by Microsoft.

For transmitting data to separated networks, it is necessary to satisfy various requirements such as storing medium like CDs or USBs, authentication systems, access control systems, and encryption/decryption functions [5]. Recently multiple systems, networks, and security systems have been established and operated for satisfying such requirements. The present situation under the separated networks increases investments on IT properties and degrades business efficiencies of system and network managers and other users.

Thus, in this study a Linked Network Security system based on Virtualization (LNSV) is proposed to effectively perform data transmissions in a network separated environment under the aspects of management, operation, and cost. For achieving it, Sect. 2 investigates data transmissions in a network separated environment. Section 3 describes a security system designed in this study. Section 4 represents an analysis of the proposed system. Section 5 concludes this study.

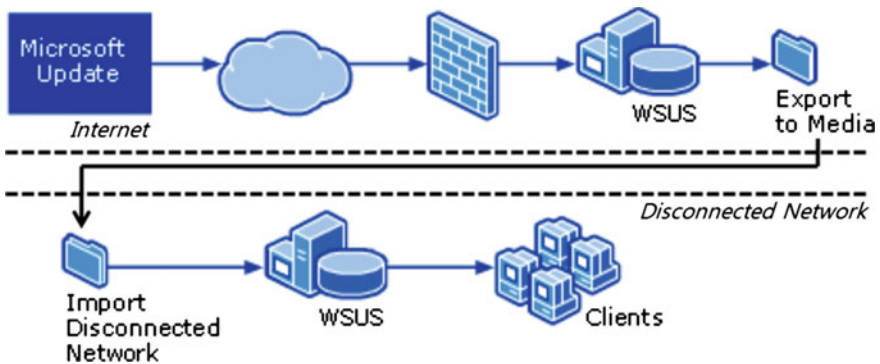
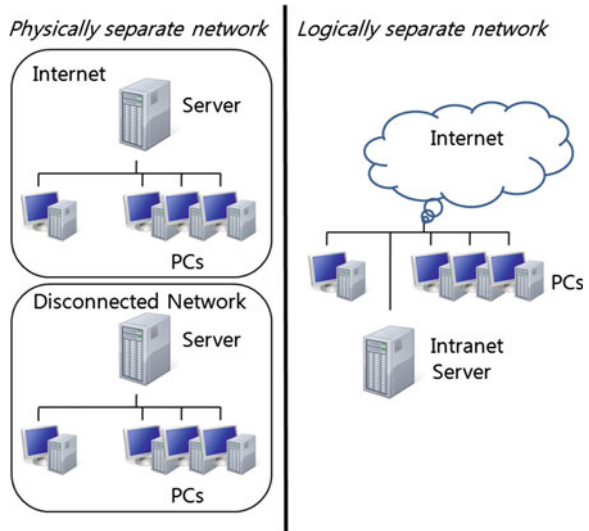


Fig. 1 Distributing updates on an isolated segment based on WSUS (Microsoft)

Fig. 2 Physically and logically separate network environment



2 Related Work

Network separation for protecting cyber threats in networks and leaks of internal information can be divided into physical and logical ways based on its construction methods [6]. The physical network separation separates all systems and networks based on its hardware. The logical network separation is performed using a software technology. Figure 2 represents the physical and logical network separation.

In network separated environments, CDs and USBs can be usually used to transmit data. However, it shows disadvantages in leaking data due to its losses, user inconveniences for connecting a device to a PC, and distributing malicious codes and requires an additional system for USB authentication and encryption/decryption related. In addition, a storage based connection server can be used in which the configuration is presented as shown in Fig. 3 [7]. The disadvantages of this server are to establish and operate extra multiple servers and storing devices and an information security system that manages accesses and encrypts transmitting and receiving data.

3 Design of LNSV

In this study, a single security system, LNSV, that guarantees safe data transmissions in a separated network environment is proposed. The proposed system is designed using a virtualization technology.

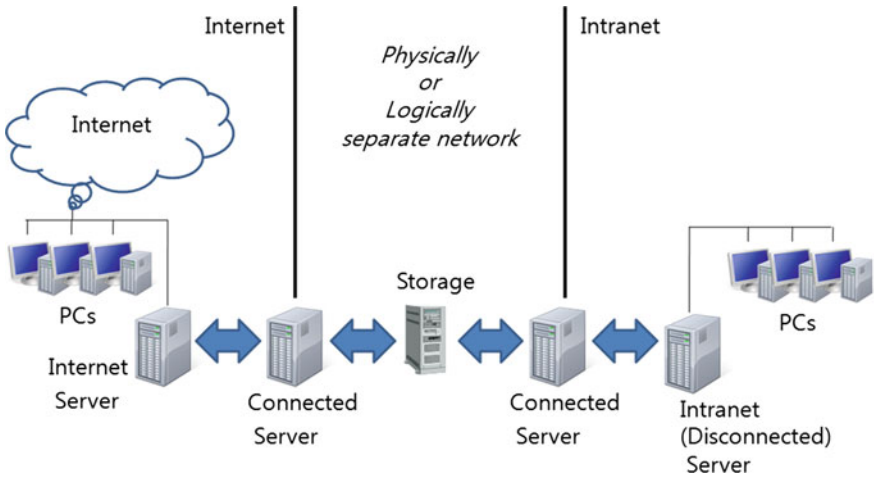
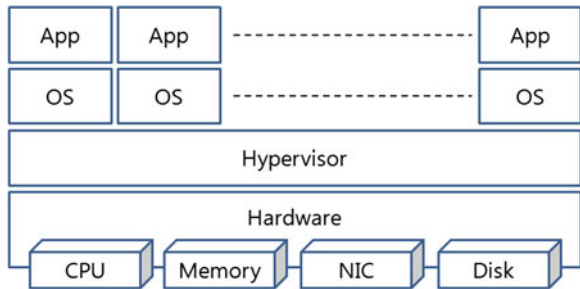


Fig. 3 Connected server based on storage architecture

Fig. 4 Concept of hypervisor



Virtualization is a technology that makes possible to share and use system resources [8]. A core technology in this virtualization is a hypervisor that allows multiple operational systems in a single system [9]. Figure 4 illustrates the concept of this technology.

By using the hypervisor, it is possible to provide various functions in a single system. Thus, the requirements for implementing safe data transmissions in a network separated environment can be integrated in a single system using multiple operational systems.

For implementing data transmissions between separated networks, an accessible connection module and a storing space are required in an individual network. Users can store the data to be transmitted to a storing space using a connection module directly. Also, it is necessary to prevent unauthorized accesses while the stored data is transmitted/received and the security of the data is to be ensured under encryption/decryption functions in order to protect the data from unauthorized users. The architecture of LNSV proposed in this study is presented in Fig. 5.

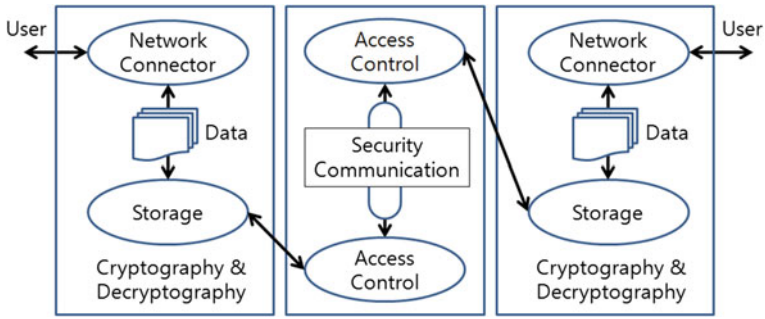


Fig. 5 Conceptual design of LNSV

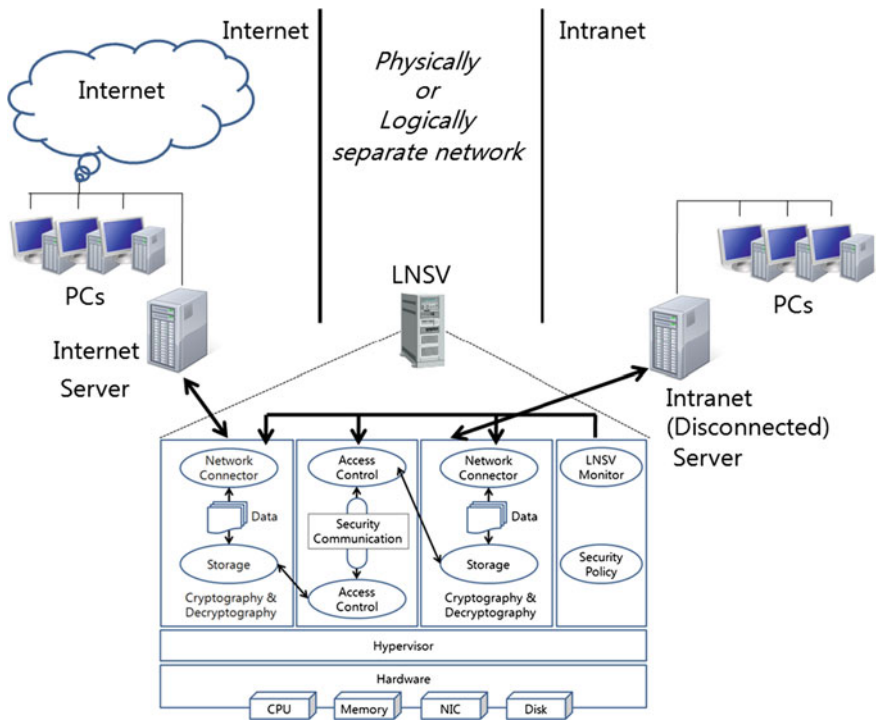


Fig. 6 Installed LNSV in separate network for data delivery

4 Analysis of LNSV

Figure 6 shows the application of the designed LNSV to a network separated environment. The implemented LNSV consists of four virtual systems. One of the systems performs safety communication and controlling accesses. Two of them support user accesses in individual networks and storing the data to be transmitted.

While storing data, source IP/PORT, destination IP/PORT and Hash Values are included. Then, the LNSV can implement communications separately to users who transmit/receive data. Thus, there are no sessions for the communication between users in separated networks. As the individual networks presented in Fig. 6 are two different networks, Intranet and Internet, the system that provides the connection module and storing space is also configured by two different parts. The remained system represents the overall operation of the LNSV and plays a role of supporting the security policy for each system in the LNSV.

The LNSV shows no leaks of data due to media losses because it does not use CDs or USBs. Also, it has no disadvantages of user inconvenience for connecting it to a PC and of distributing malicious codes. In addition, it does not need an extra system for authenticating USBs and encrypting it. Because it satisfies all security functions required to transmit data in a network separated environment, it is possible to focus on managing and operating a single system and that leads to effectively promote the works of system and network managers. Moreover, it shows an advantage that needs small initial investment for ensuring security in a network separated environment.

5 Conclusion

The LNSV proposed in this study represents an open architecture in accessing its system through network connectors for all users in individual networks and can be used as a general purposed system for storing all data to be transmitted. It is possible to prevent the access of unauthorized users because the stored data files include source IP/PORT, destination IP/PORT and Hash Values. Also, it can guarantee the security of communication through transmitting and receiving data using encryption/decryption functions. Thus, the LNSV can provide safe connection services between separated networks.

In recent years controls systems for major infrastructures have been operated through connecting them to external networks [10]. As the control systems exhibit large damages and ripple effects caused by violations, it requires a more strong security system. Thus, it is expected that the LNSV proposed in this study will represent high-usability and valuable uses.

Acknowledgments This work was supported by a grant from Kyonggi university advanced Industrial Security Center of Korea Ministry of Knowledge Economy.

References

1. Yang SJ, Stotz A, Holsopple J, Sudit M, Kuhl M (2009) High level information fusion for tracking and projection of multistage cyber attacks. *Inf Fusion* 10(1):107–121
2. Shahriar H, Zulkernine M (2012) Trustworthiness testing of phishing websites: a behavior model-based approach. *Future Gener Comput Syst* 28(8):1258–1271

3. Nicholson A, Webber S, Dyer S, Patel T, Janicke H (2012) SCADA security in the light of cyber-warfare. *Comput Secur* 31(4):418–436
4. Blasco J, Hernandez-Castro JC, Tapiador JE, Ribagorda A (2012) Bypassing information leakage protection with trusted applications. *Comput Secur* 31(4):557–568
5. Joe I, Lee S (2011) Design and implementation of storage-based data sharing system in the separate network environment. *Korea Inf Commun Soc* 36(5):477–483
6. Jee J, Lee S, Lee S, Bae B, Shin Y A logical network partition scheme for cyber hacking and terror attacks. *J Korean Inst Inf Scientists Eng* 39(1):95–100
7. Kim J, Kim J, Han Y (2010) EAI-based architecture design and implementation for software development between disconnected networks. *Korean Soc Internet Inf*, pp 257–258
8. Rodríguez-Haro F, Freitag F, Navarro L, Hernández-sánchez E, Farías-Mendoza N, Guerrero-Ibáñez JA, González-Potes A (2012) A summary of virtualization techniques. *Procedia Technol* 3:267–272
9. Lee G, Ji J-W, Chun H-W, Lee K-W (2011) Design of an authentication system proper for hybrid cloud computing system. *J Inf Secur* 11(6):31–36
10. Kim KA, Lee DS, Kim KN (2011) ICS security risk analysis using attack tree. *J Inf Secur* 11(6):53–58

A Study About Security Awareness Program Based on RFID Access Control System

Kyong-Ho Choi and DongHwi Lee

Abstract In this study, a security method that protects data from an information system accessed by unauthorized persons through physical ways like eye contacts is proposed. In addition, it is applied to security awareness programs for improving security recognition and to contribute to the protection of important information. The security awareness program proposed in this study is able to detect the violation in physical security policies and to implement additional training related to warning messages or such violated security policies and that leads to present positive effects of changing user security awareness and corresponding works. It is expected that these effects are to be spreaded to the whole organization and influence for all members.

Keywords Security awareness · RFID · Access control · Information security

1 Introduction

Recently people use various information and communication devices in their daily lives and works. Such information and communication devices used in collecting, storing, and analyzing information effectively accumulate valuable data and its

K.-H. Choi

Center for Industry, Kyonggi University, San 94-6, Iui-Dong, Yeongtong-Gu,
Suwon-Si, Gyeonggi-Do, South Korea
e-mail: cyberckh@gmail.com

D.-H. Lee (✉)

Department of Industrial Security, Kyonggi University, San 94-6, Iui-Dong,
Yeongtong-Gu, Suwon-Si, Gyeonggi-Do, South Korea
e-mail: dhclub@naver.com

importances have been increasingly emphasized. Thus, technical, physical, and managerial efforts for protecting computers and networks from cyber threats have been made. By organizing these efforts, security policies and its implementations strengthen the capability of protecting information in organizations and that makes possible to protect such valuable information.

However, systematic security policies and plans can be degraded due to users' inattention and unconcern. For instance, as a manager who charges personal medical information with a low awareness in security, it may release personal information and violate privacy [1, 2]. Thus, the awareness in security should be recognized as an important element to reduce user mistakes [3] and programs for increasing such awareness are to be prepared.

There are various cases in releasing important information due to users' inattention and unconcern. Also, there are some unexpected information leaks like a unknown eavesdropping by a third person for secret businesses in a public space or opening some secret reports to a person with a low security level. In particular, under recent business environments managing important information using computers a third person or a person with a low security level may steal information from resources through glancing it furtively, Piggy-backing, Shoulder surfing or other untechnical ways [4, 5].

In this study, a security method that protects data from an information system accessed by unauthorized persons through physical ways like eye contacts is proposed. In addition, it is applied to security awareness programs for improving security recognition and to contribute to the protection of important information. For achieving it, [Sect. 2](#) examines access control using RFID. [Section 3](#) introduces a security awareness program proposed in this study. [Section 4](#) analyzes the proposed security awareness program. [Section 5](#) concludes this study.

2 Related Work

It is possible to perform a differentiated management that allows authorized persons who charge specific works in allowed sections by separating sections physically using Radio Frequency Identification (RFID) [6]. In particular, determining the indoor or outdoor condition of a user who has an RFID tag and recognized by an RFID reader represents [7] a definite control of authorized persons. However, there exist some problems that unauthorized persons follow an authorized person closely and piggy-backing or glancing a monitor or documents and shoulder surfing on a desk furtively to steal such data.

Thus, it is necessary to prepare auxiliary devices or assistants to count ingress and egress persons in order to control unauthorized persons and to protect data. Infrared ray sensors and detection devices can be used to count such ingress and egress persons [8]. Actually, by comparing the number of RFID recognized persons and the number of infrared detector counted persons the number of unauthorized persons are verified in a specific area.

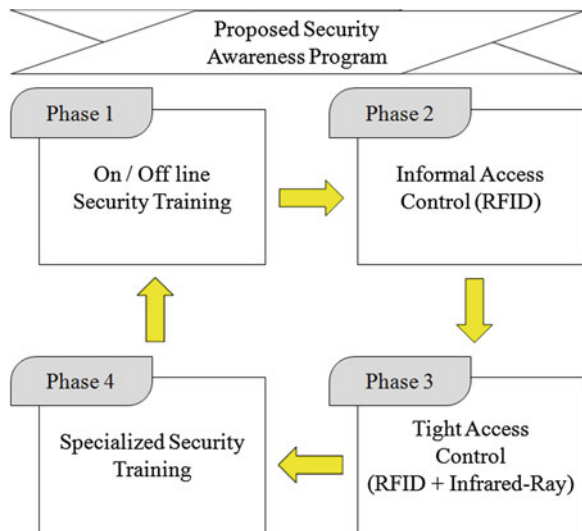
In this study, based on the fact that protects some threats of releasing data from unauthorized persons through combining RFID and infrared detectors, a security awareness program for establishing more safe information security environments is proposed.

3 Proposed Security Awareness Program

The proposed security awareness program consists of four steps. The first phase is to teach off-line security training at a lecture room different from on-line training using Internet home pages, e-mail, and mobile devices. The second phase operates an access control system using RFID sensors and readers. The third phase applies a strong access control using RFID sensors and infrared detectors as specific days, periods, and conditions are required. The final phase implements training for the security awareness program related to control unauthorized persons to the persons who charge such a control process. The proposed security awareness program is presented in Fig. 1.

For achieving the strong access control in the Phase 3, an RFID tag is to be recognized to an RFID reader and then a person who gets the RFID is verified. Then, the actual number of gate passed persons is also counted using infrared sensors. If the number of persons detected by the infrared detectors exceeds the number of persons recognized by the RFID sensors, the excessive persons are determined as unauthorized persons. In addition, it is possible to verify the security levels applied to users by determining different security levels for individual spaces in an RFID control system. The configuration of the physical spaces in the Phase 3 is presented in Fig. 2.

Fig. 1 Proposed security awareness program process



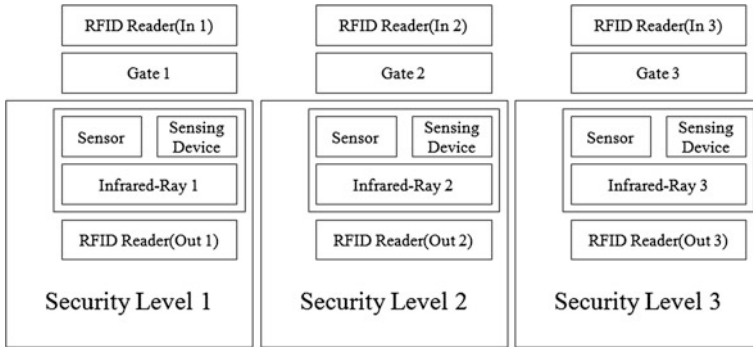


Fig. 2 Physical environment of tight access control

4 Analysis of Proposed Program

The information required for implementing the proposed security awareness program for each user are Employee Identification Number, Allowed Office Number, RFID Tag Number, and its Security Label. Table 1 represents such information in the Y enterprise that is a subject in this study.

The second floor of the Y enterprise consists of three different offices. Thus, the security levels for each office are different according to their businesses. It makes an easy control of ingress and egress persons.

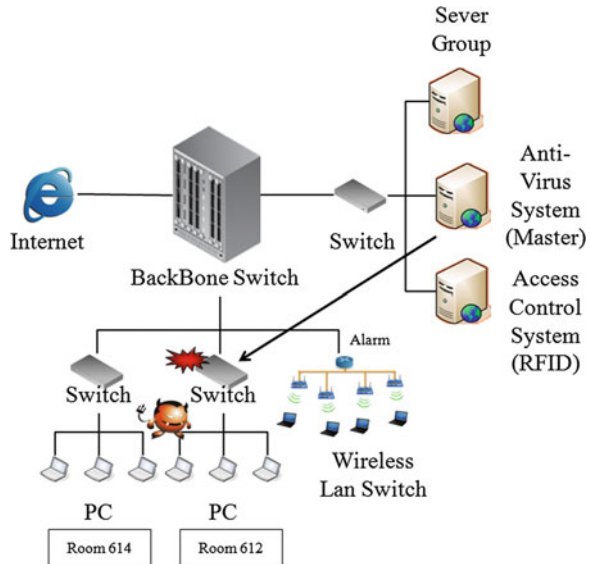
The experiment was implemented to detect unauthorized persons who try to enter the offices and to verify the proper security levels applied to members. As a problem occurs in an office while monitoring the access control at the Security Management Center, a warning message is transmitted to the users in the office using a security agent. In this case, it is possible to sent a more strong message to a specific user by covering its monitor using a general image file or controlling the movement of a mouse cursor. Then, a security training of protecting unauthorized persons was implemented for the members who engaged in the office. The network structure used in this process is presented in Fig. 3.

The proposed security awareness program shows an advantage that maximizes the secureness of internal information. It does not allow to access information by unauthorized persons or accessors with a low security level even a part of information.

Table 1 Information of access control for tight access control policy

	Employee identification number	Name	Allowed office number	RFID tag number	Security label
1	0048	Mr. Lee	612, 614	0048	1
2	0062	Mr. Choi	612	0062	2
3	0627	Ms. Kim	612, 614	0627	1
4	0707	Ms. Kang	612	0707	2

Fig. 3 Network structure for proposed program



In addition, it makes possible to grant the role and responsibility of users who try to access their own resources and to improve the understanding and implementation of the security policy by promoting such security awareness based on the warning message personally delivered to members in an organization. It reveals that the proposed method includes people, procedures, and technologies regarded in information security management [9]. In this study, procedural characteristics in establishing a security policy for each individual business phase and some technical elements including RFID, infrared ray, and anti-virus system are presented.

5 Conclusion

The information security management is a type of system and cannot be simply determined as a technical reference [10]. The evaluation of the security policy that is established and implemented is important. The evaluation can improve the security awareness of the members in an organization and develop it based on some issues pointed in the evaluation.

The security awareness program proposed in this study is able to detect the violation in physical security policies and to implement additional training related to warning messages or such violated security policies and that leads to present positive effects of changing user security awareness and corresponding works. It is expected that these effects are to be spreaded to the whole organization and influence for all members.

Acknowledgments This work was supported by a grant from Kyonggi university advanced Industrial Security Center of Korea Ministry of Knowledge Economy.

References

1. Jeong S-J, Kim M, Lee C-K, Lee C-H, Son B-C, Kim D-H, Kim J-H, Lee J-T, Chang G-R, Hong K-H, Suh C-H (2011) Cognition and practice on medical information protection of industrial health care manager of small and medium sized enterprises. *Korean J Occup Environ Med* 23(1):53–63
2. Kim J-J, Kweon E-H (2012) A study on the therapists' protective actions of medical information privacy—with a focus on physical and occupational therapists. *Korea Inst Electron Commun Sci* 7(2):447–455
3. Drevin L, Kruger HA, Steyn T (2007) Value-focused assessment of ICT security awareness in an academic environment. *Comput Secur* 26(1):36–43
4. Goucher W (2011) Look behind you: the dangers of shoulder surfing. *Comput Fraud Secur* 2011(11):17–20
5. Johnny L, Scott P, Jack W, Kevin DM (2008) *No tech hacking*. Syngress, Rockland
6. Kim MS, Lee DH (2012) A way of securing the access by using PCA. *J Inf Secur* 12(3):3–10
7. Choi KH, Kim JM, Lee D (2012) Network 2-factor access control system based on RFID security control system. *J Inf Secur* 12(3):53–58
8. Kim JM, Choi KH, Lee DH (2012) Network group access control system using piggy-backing prevention technique based on infrared-ray. *J Inf Secur* 12(4)
9. Eminagaoglu M, Ucar E, Eren S (2009) The positive outcomes of information security awareness training in companies—a case study. *Inf Secur Tech Rep* 14(4):223–229
10. Broderick JS (2006) ISMS, security standards and security regulations. *Inf Secur Tech Rep* 11:26–31

A Framework for Anomaly Pattern Recognition in Electronic Financial Transaction Using Moving Average Method

Ae Chan Kim, Won Hyung Park and Dong Hoon Lee

Abstract Nowadays, security incidents of financial IT services and internet banking hacking against the financial companies have occurred continuously, resulting in a loss of the financial IT systems. Accordingly, this paper based on ‘framework standards of financial transaction detection and response’ was designed to propose of anomaly Electronic Financial Transaction (EFT) pattern recognition and response for the method to detect anomaly prior behaviors and transaction patterns of users. It was applied to moving average based on the statistical basis.

Keywords Electronic financial transaction • Pattern recognition • Moving average

1 Introduction

Financial business can be said to be IT-intensive industry due to its early use of information technologies. With its utilization of digital technologies and internet, customer contact points, online payment services and exert systems have recently

A. C. Kim

Department of Financial Security, Graduate School of Information Security, Korea University, 145 Anam-dong 5-ga, Seoul, Seongbuk-gu 136-713, South Korea
e-mail: holytemple@korea.ac.kr

W. H. Park (✉)

Department of Information Management, Far East University,
Wangjang-ri, Gangok-myeon, Chungbuk, Eumseong-gun 369-700, South Korea
e-mail: whpark@kdu.ac.kr

D. H. Lee

Graduate School of Information Security, Korea University, 145 Anam-dong 5-ga,
Seoul, Seongbuk-gu 136-713, South Korea
e-mail: donghlee@korea.ac.kr

been introduced, and new products and services have come to be provided at a lower transaction cost. However, IT systems have widely been applied to the financial companies, security incidents of financial IT services such as financial transaction sniffing, spoofing and internet banking hacking against the financial companies have occurred continuously, resulting in a loss of the financial IT systems. To respond effectively to the financial IT security incidents, this paper suggested anomaly pattern recognition method to minimize the damages.

2 Related Work

Cognitive psychology define recognition of exterior object acting on human as pattern recognition, which match input-information with information of long term memory, then make decision about how to classify input-information into certain known pattern [1]. Pi Younguo [2] suggested that how to apply achievement of cognitive science to traditional machine pattern recognition by combining with characteristic of machine pattern recognition was discussed.

Chan [3] and Zhang [4] adopted wavelet analyzing method that their approaches have yielded some effects in data volume reduction and sequence matching. Tianqing Zhu [5] applied the idea of EMD to develop strategies to automatically identify the fluctuation variance in financial time series in order to detect suspicious transaction behavior. Recently, there is a standard for EFT detection and response in Korea [6]. It contains framework for fraud and anomaly financial transaction to detect and response.

In this paper adopted an approach to create an attack graph based on statistic moving averaged for prediction of anomaly financial detection. The concept of using statistic at framework was proposed in [7]. It is possible to detect of prior anomaly transaction behaviors and detection of transaction pattern information. Also, it could be utilized as the basic data to establish effective response to detect anomaly transactions through data-mining.

3 Anomaly Pattern Recognition

The anomaly pattern recognition, those of which are out of their normal range, which are the transaction user terminal, the prior transaction behavior, and the transaction pattern when a user uses the electronic financial transaction. In addition, it means determining the anomaly transaction by collection and analyzing user transaction information for online transactions. It should be comprehensively determined by the prior transaction behavior, and the transaction pattern [6]. Table 1 presents determination targets and methodology proposals for development of the comprehensive pattern recognition.

Table 1 Anomaly pattern recognition methodology

Item	Detail	Collecting and detecting information	Methodology proposal	Purpose
Anomaly prior behavior	Attempt of anomaly prior behavior for financial transaction	Log-in, error limit exceeded, change of personal information	Pattern matching based on moving average (IF-THEN Rule based)	Detection of anomaly behavior
Transaction pattern	Change to the financial transaction pattern compared to the past pattern	Transaction amount, frequency and target account		Detection of change in financial transaction pattern

Detecting the prior financial transaction behavior should be performed on those activities of prior information change for the transaction among user activities, and it is necessary to set the range of conditions out of the range of normal prior behavior. Likewise, detecting the financial transaction pattern information also requires information clearly identifiable of fraud transaction patterns by comparing with the existing transaction pattern information. For setting the normal range to achieve it, moving average based on the statistics theory will be applied. Figure 1 illustrates the detection and response method of fraud financial transactions that the proposed methodologies are applied.

When the user accesses the website with HTTP first and accesses the open electric financial transaction service, it will be accessed in the TLS/SSL based environment. At this time if a security program is not installed in the user terminal, the security program installation step will be processed. Next, derived items by the analysis result of the detection results of the user prior behavior and the transaction pattern will be reported to the electronic financial transaction detection server. Then the server determines if the result is anomaly transaction if the result values exceed the critical values, it will be determined as the anomaly electronic financial transaction and it will perform processes such as ‘additional authentication’ or ‘blocking transaction’. If within the normal range, it will perform the ‘transaction authorization’ process.

4 The Anomaly Pattern Recognition Model

4.1 Prior Financial Transaction Behavior Recognition

The prior behavior within electronic financial transaction means the behavior that does not perform transaction directly but may affect it. To set the normal range criteria of past prior behaviors of a user for each factor in (Table 2), the moving

Table 2 Pattern recognition logic of prior behavior

User prior behavior (example in parenthesis)	Detection logic	
	Moving average	Pattern matching
Condition (1): user log-in time significantly longer (30 min) or shorter (30 s) than used to be	Applied	Applied
Condition (2): exceeded allowed number of log-in attempts (default 5 times)	Applied	Applied
Condition (3): changed log-in location (location based authentication by access IP or GPS)	Not applied	Applied
Condition (4): use of the authentication method other than used to be (on the first authentication of the different authentication mean)	Not applied	Applied
Condition (5): if to renew the Certificate even though the valid date of the Certificate remains more than (6 months)	Not applied	Applied

average¹ [7], one of statistical trend analysis methods, and the pattern matching logic are applied in this paper.

In the detection logic, as shown in (Fig. 1), the prior behavior information of financial transaction sets the normal prior behavior range according to moving average calculation result, after the ‘log collector’ program in the user system receives past prior behavior information from the fraud transaction detection server when the user accesses the financial transaction server first. Then the user prior behaviors are detected by the IF-THEN-ELSE pattern matching method in the steps of (1)–(5). For example, after past records of the log-in time required of the user are received from the server, if to set the range of the normal prior behavior according to change in the time required, it can be calculated as shown in (Table 3).

In conclusion, by the pattern recognition result of user prior behaviors, if a factor is out of the scope of normal prior transaction behaviors, the financial company may apply ‘additional authentication’ and if two or more factors are out of them, the company may apply ‘transaction blocking’.

4.2 Transaction Pattern Information Recognition

Statistically, users of the electronic financial transaction show certain transaction patterns. The fraud transaction may be detected by using such patterns information

¹ Moving average is one of the methods to determine the trend value. For time series of X_1, X_2, \dots, X_t , and moving average \bar{X}_m in the period range of m at the time t is derived as follows. $\bar{X}_m = (X_t + X_{t+1} + \dots + X_{t+(m-1)})/m$, ($t = 1, 2, \dots, (t - m)$). When new series of $\bar{X}_{m+1}, \bar{X}_{m+2}$ are made in this way, the change in current time series represents an even trend.

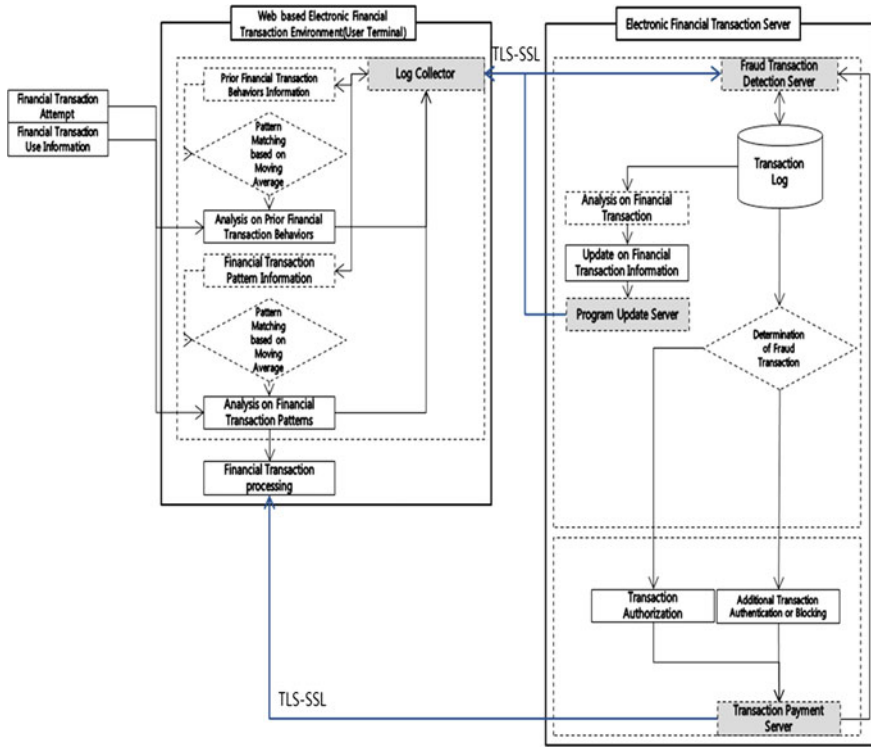


Fig. 1 Anomaly pattern recognition model in EFT

Table 3 Normal range calculation of prior behavior in ‘Log-in Time Required’

Repeat	1	2	3	4	5	6	7	8	9	...
Time(seconds)	10	15	20	14	16	12	10	20	15	...
$\bar{X}_m = 15, \bar{X}_{m+1} = 16.3, \bar{X}_{m+2} = 16.67, \bar{X}_{m+4} = 14, \bar{X}_{m+5} = 12.67 \dots$ \therefore The scope of normal prior behavior on the $m + 2$ th log-in is $\bar{X}_m - \sigma \leq \bar{X}_{m+2} \leq \bar{X}_{m+1} + \sigma$										

of individual users. If the transaction patterns such as sharp increase in transaction amount increase in the number of transactions or repeated transactions in certain amount are out of the range of normal transaction patterns, it may be subject to the fraud financial transaction.

For information recognizable of user transaction patterns, ‘transaction amounts, frequency of transactions and transfer target accounts’ may be used. Based on the range of expectations required upon normal transaction, it is possible to quantify this information by increasing risk points by detecting transactions out of the normal range. To set the range of normal transaction patterns, moving average is applied for ‘transaction amount’ and ‘frequency of transaction’ in (Table 4). With past records of transaction amounts and frequency of transactions of the user, the

Table 4 Pattern recognition logic of user financial transaction

User transaction pattern information	Detection logic	
	Moving average	Pattern matching
(1) Transaction amount	Applied	Applied
(2) Frequency of transactions	Applied	Applied
(3) Transfer target account	Not applied	Applied

normal range criteria of change in transaction amounts and frequency of transactions may be set. For example, if larger amount than the normal range set is withdrew or deposited, or frequency of transactions is sharply increased within a certain period of time, they may be detected as symptoms of fraud transaction patterns.

5 Conclusions

The EFT has contributed to increase user convenience as none face-to-face transaction method. However, due to attacker techniques that become more and more intelligent and sophisticated, security threats to users of anomaly EFT continuously increase. In addition, the existing personal firewall programs using detection techniques based on signature are not enough to fundamentally solve. Therefore, in this paper, to improve existing anomaly pattern recognition methods limited to signature based detection. That is, the pattern matching method based on statistical moving average will be used to detect and block anomaly EFTs by malicious users.

Acknowledgments This work is supported by the Korea Information Security Agency (H2101-12-1001).

References

1. Rueda LG, Oommen BJ (2002) On optimal pairwise linear classifiers for normal distributions: the two-dimensional case. *IEEE Trans Pattern Anal Mach Intell* 24(2):274–280
2. Youguo P (2007) The frame of cognitive pattern recognition. In: *Proceedings of the 26th Chinese control conference, Zhangjiajie, Hunan*, pp 694–696, 26–31 Jul 2007
3. Chan K, Fu W (1999) Efficient time series matching by wavelets. In: *Proceedings of the 15th IEEE international conference on data engineering, Sydney*, pp 126–133
4. Zhang HQ, Cai QS (2003) Time series similar pattern matching based on wavelet transform. *Chinese J Comput* 26(3):373–377
5. Zhu T (2006) Suspicious financial transaction detection based on empirical mode decomposition method. In: *Proceedings of the 2006 IEEE Asia-Pacific conference on services computing (APSCC'06), Guangzhou, Guangdong*. IEEE Computer Society, pp 300–304

6. Telecommunications Technology Association (TTA) (2011) Fraud detection and response framework in electronic financial transaction system, TTAK.KO-12.0178
7. NIST/SEMATECH (2012) e-Handbook of Statistical Methods: what are moving or smoothing techniques? Average, Apr. 2012. <http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc42.htm>

An Investigation on the Research Topics in Relation to Information Systems in Supply Chain Management

Younjung Kim, Young-Ho Lee, Kyung-Yong Chung
and Kang-Dae Lee

Abstract This study is designed to investigate the trend of information systems (IS) research in relation to supply chain management (SCM). The purpose of this research is to predict how IS research is likely to evolve in the near future and to suggest to which direction it should be further conducted. We attempted to investigate the previous research topics about information systems in the fields of SCM on the basis of 96 journal articles published between the year of 2006 and 2010 including the database of ScienceDirect. The most frequently appeared keywords in the titles and abstracts were searched by two different categories: “business, management and account” and “decision science”. As a result we found that the most popular research topics appeared in the articles were “impacts of IS on the performance of SC and enterprise” “framework and model of IS in SCM”. We also suggested the future research implications of the current research trend or preference. This study academically and practically contributes to deepening our understanding of the on-going issues discussed on the current IS papers in the field of SCM by suggesting the future direction of studies.

Y. Kim · K.-D. Lee (✉)

Department of Packaging, Yonsei University, 1 Yonseidae-gil, Wonju,
Gangwon-do 220-710, South Korea

e-mail: pimeson@yonsei.ac.kr

Y. Kim

e-mail: timiotera@gmail.com

Y.-H. Lee

Department of Computer Science, Gachon University of Medicine and Science,
534-2 Yeonsu3-dong, Yeonsu-gu, Incheon, South Korea

e-mail: lyh@gachon.ac.kr

K.-Y. Chung

Department of Computer Information Engineering, Sangji University, Wonju, South Korea

e-mail: dragonhci@gmail.com

Keywords Information system · Supply chain management · Keywords · Trend · Future direction

1 Introduction

Supply chain is defined as a network of relationship between organizations such as suppliers and customers (Kelle and Akbulut 2005), and the main goal of supply chain management is the management of the interconnection of organizations which relate to each other through upstream and downstream linkages between the different processes that produce value in the form of products and services to the ultimate consumer [1]. As the stakeholders of SC are located all over the world by globalization, it is more required to integrate the activities and process both intra and internal organization [2]. This means that without implementation and integration of ISs for information sharing it is impossible to achieve the benefits of SCM. Accordingly, there were few literature studies on information technology and system in the field of logistics, supply chain management, and some industries [1–4]. The purpose of this research is to predict how IS research is likely to evolve and to suggest to which direction it should be further conducted. Previous research was mainly focused on the comprehensive studies of information technology and systems justification; however, in this paper we attempted to explore the extensive studies of information system in supply chain management.

This study is designed to investigate the trend of ISs research in relation to SCM. The remainder of this paper is organized as follows. In [Sect. 2](#), we discuss IS in SCM with literature review. The methodology of research is explained in [Sect. 3](#). The results of investigation on the previous research are being discussed in [Sect. 4](#). In the last section, the limitation and conclusion of this paper are presented.

2 Literature Review

While Menzer et al. (2010) viewed supply chain (SC) as a set of three or more organizations directly involved in the upstream and downstream flows of products, services, finances and/or information from a source to a customer [5], Kelle and Akbulut (2005) interpreted it as a network of relationship between organizations such as suppliers and customers. For a more detailed explanation they pointed out that SCM is the management of the interconnection of organizations that relate to each other through upstream and downstream linkages between the different processes that produce value in the form of products and services to the ultimate consumer [1].

Ginnarsson and Jonsson (2005) showed that increased collaboration among the partners of a supply chain contributed to improve the performance of SC. In order

to achieve significant performances in SCM [6] suggested the integration of business processes and information flows of SC partners cannot be overlooked. The benefits associated with the integration of supply chain system involve increasing competitive advantage, decreasing operational costs, and improving collaboration and coordination among the partners of supply chain. However, without information system (IS) and information technology (IT), it is unlikely to achieve the benefits of SCM since the stakeholders of SC are located all over the world. Thus, it is required to integrate the activities and process both intra and internal organization [2] to increase efficiency in SCM. This means that implementation and integration of ISs for information sharing has played a key role in promoting business competitiveness.

Without IT/IS that support SCM relatively poor services or high cost products in terms of quality would be provided to customers. Many attempts have been made to reduce manufacturing costs such as the costs of managing resources and controlling inventories. IS used within organizations for SCM can be categorized into four types: (1) transaction processing system (TPS), (2) management information system (MIS), (3) decision support system (DSS), and (4) artificial intelligence and expert system (AI/ES) (Oz 2008).

It has been known that the most fundamental system is TPS that handles the large volume of business transactions that occur daily within an organization, and MIS is operated on the basis of the information from a TPS that supports management decision making. This system yields to a variety of reports such as scheduled reports and demand reports. DSS is information and planning system, which organizationally collects people, procedures, databases, and devices used to support problem-specific decision making. DSS differs from MIS in the support given to users toward the decision emphasis, the development and approach, and system components, speed, and output. Another system considered to support management decision making is the expert system (ES) [7].

3 Methodology

3.1 Article Sampling

Literature review was conducted to capture a snapshot of the current academic research of information system in SCM field. 96 Journal articles were sourced from the ScienceDirect database which has been known as one of the major scientific databases offering journal articles and book chapters. In this research, we only sorted keywords from the topics and abstracts of the journal articles. According to the 2011 journal citation data, ScienceDirect includes many of top journals in the field of SCM such as Journal of Operations Management, Omega International Journal of Management Science, European Journal of Operational Research, and Decision Support System. The search terms that were found in the intersection of

“information system” and “supply chain management” within abstracts, titles and keywords (as of September 2012) in all journals of ScienceDirect. In order to increase accuracy, the fields of disciplines that are subjected to this search are limited to Business, Management and Accounting and Decision Science. The time horizon in this research is limited to 5 years between 2006 and 2010 in that A. Gunasekaran and E.W.T. Nagai conducted a similar analysis for the past 15 years between 1991 and 2005.

3.2 Descriptive Features of the Whole Literature

In the past decade between 2001 and 2010 the quantity of research dealt with the information system in supply chain management has steadily increased (Fig. 1). Compared to the earlier half of the last decade with its later half, the number of published journal articles in the later half was one and half ratio (Figs. 1 and 2). The types of journals dealt with these topics also became more various toward the year of the 2010, and the list of journal is described in Table 1.

4 Analysis, Results and Implications

Li et al. [8] compared the research trend of the two groups of researchers whose research topics were about IS and management. They described that IS researchers are more likely to focus on the information flow while management researchers tended to be more likely to focus on the materials and finances. For example, in the aspect of inter-organizational information sharing IS researchers analyzed the benefits of EDI or the level of information sharing. On the other hand, management researchers studied the topics regarding the decrease in the bullwhip effect due to information sharing or the decision making of inventory level with information sharing (Table 2).

Fig. 1 Journal articles by years (2001–2010)

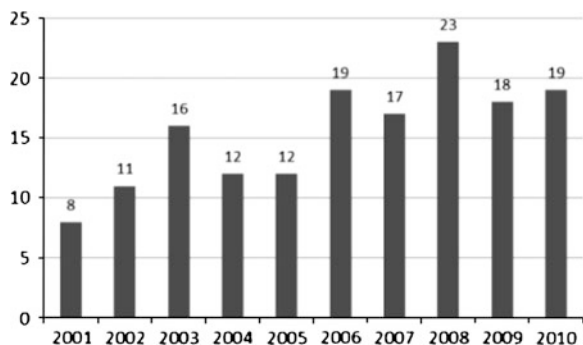


Fig. 2 Journal articles by years (2006–2010)

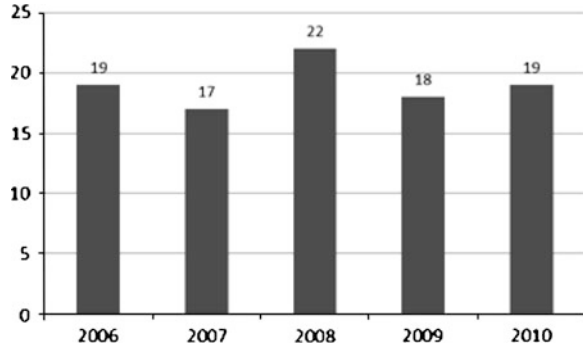


Table 1 Theresources of articles—journal

Journal	Number	Percentage	Cumulative percentage
International journal of production economics	19	20	21
European journal of operational research	19	20	41
Decision support systems	9	9	50
Computers in industry	8	8	58
Expert systems with applications	8	8	67
Computers & industrial engineering	6	6	73
Journal of operations management	5	5	78
Omega	5	5	83
International journal of information management	3	3	86
International journal of project management	2	2	89
Others	11	11	100
Total	95	100	–

Table 2 Previous research by category

Category	Reference
<i>SCM (Management, business and account)</i>	
Impact of IS on SC performance	[9, 10, 17, 18, 21, 22, 23, 25, 26, 27, 29, 30, 31]
The function of IS in SCM	[11, 12, 13, 19, 28]
Application of IS in SCM	[16, 23; Persona et al. (2007); 14, 20]
<i>Information system (decision science)</i>	
Framework and model of IS in SCM	[35, 40, 43, Wang et al. (2010), 33, 36, 37, 39, Zhang et al. 55, 56, Uçkun et al. 38, 48, Kwon et al. (2007)]
Application	[48, 52, Kurata and Yue (2008), 47]
Infrastructure	RFID: Lin [43, 44, Wang et al. (2010), 45, 50] Web: Repoussis et al. [54]
Technique	Optimization: Shukla et al. [34, 49], Policy: Lee and Wu [46, 37], Datamining: Thomassey [41, 32]

[9–56]

In this section, we analyzed the characteristics of information system (IS) in supply chain management appeared in information system papers and management papers to explore the research trend formed in accordance with the view point of IS. With the subject category of ScienceDirect, we divided sampling papers into two areas, (1) business, management and accounting and (2) decision science. The former investigated the journals: Omega, Journal of Operations Management, Decision Support Systems, etc. The latter mainly consists of the following journals: Computer in Industry, Computers and industrial Engineering, etc. However, International Journal of Production Economics and European Journal of Operational Research were distributed to the area of business, management and accounting and the area of decision science respectively depending on research subject.

In the fields of business, management and accounting, previous research paid more attention to (1) the impact of IS on the performance of enterprises or SC, (2) the present status of IS implementation and use, and (3) application of IS in a variety industry' SC. On the other hand in the field of decision science, previous research more concentrated on the classification of (1) the framework and modeling of SCM IS, (2) the application of IS in SCM, and (3) the technique of IS.

Su and Yang [9] explored what and how ERP system impacts on the performance of enterprises with structural equation model, and showed the positive relationship between ERP system and the operation, customer relationship and planning and control of enterprises. In addition, Seggie et al. [10] examined the impact of IT on enterprises' brand equity, which is one of intangible performances by adopting IT with structural equation model, and concluded high degree of IT alignment between SC firms positively affects brand equity. When it comes to intra and inter organizational information sharing system, functions and benefits of SCM are introduced and examined in these papers. For example, Fildes et al. [11] examined the use of forecasting support system, which is one of the decision support systems, and identified design features with literature review Holweg and Pil [12].

In contrast to the first research field that is related to the impacts of IS on SC or business performance, the area of decision science explored the impact of technique, information visualization, on ERP system (Parush et al. 2007).

In decision science, the main subject of researches is to develop frameworks and models for optimization system performance. In general, SC is composed with many different enterprises with complex systems so the integration and coordination of heterogeneous IS in internal SCare a major issue to achieve efficient SCM performance. For example, Li and Wang [40] developed the model of coordination for centralized SC and Chatfield et al. [35] worked for the supply chain modeling language so he can provide a generic framework (XML based format) for storing SC structure and managerial information to overcome the difference between heterogeneous systems.

Both business and decision science areas studied the application of IS in a variety of SC like a fashion industry (Kurata and Yue 2008; Lo et al. [39]).The attributes of an industry and its structure of supply chain should be taken into

consideration when adopting IS. Few research was done with regards to the IS application in terms of regions such as China (Ge and Voß [47]), and South East Europe [13].

4.1 Future Research

In this paper, we founded that the function of IS in SCM has been advanced from operational support to decision support function, and the object of information sharing has been enlarged the efficiency of individual enterprise into the efficiency of whole SC. Compared to the previous research [1–4], who studied information system in supply chain management, recent research tends to focus on decision support system instead of operational system. According to Forme et al. (2007) who classified the information sharing of SC into four categories: the downstream and upstream parts of SC and inter and cross of SC, the progress of SCM philosophy and IT leads to the change of research topic from operational system into decision support system.

Secondly, according to Williamson et al. [1] who categorized the phase of inter-organizational system development within SCM based on Shore (2001), recent main topic of research of IS in SCM become more focused on the web for easier integration and coordination of heterogeneous systems within SC. As a result, many research suggested web based SCM IS research such as privacy and security of information in open platform and the standard language as the future research.

5 Conclusion

This paper investigates the previous research topics about information systems in the fields of SCM on the basis of 96 journal articles published between the year of 2006 and 2010 including the database of ScienceDirect. The most popular research topics appeared in the articles were “impacts of IS on the performance of SC and enterprise” “framework and model of IS in SCM”. Also, many research suggested the research of web based SCM IS such as privacy and security of information in open platform and the standard language as the future research.

References

1. Williamson EA, Harrison DK, Jordan M (2004) Information systems development within supply chain management. *Int J Inf Manag* 24(5):375–385
2. Gunasekaran A, Ngai EWT (2004) Information systems in supply chain integration and management. *Eur J Oper Res* 159(2):269–295

3. Helo P, Szekely B (2005) Logistics information systems an analysis of software solutions for supply chain co-ordination. *Logist Inf Manag* 105(1):5–18
4. Gunasekaran A, Ngai EWT, McGaughey RE (2006) Information technology and systems justification: a review for research and applications. *Eur J Oper Res* 173(3):957–983
5. Mentzer JT, DeWitt W, Keebler JS, Min S, Nix NW, Smith CD, Zacharia ZG (2001) Defining supply chain management. *J Bus Logist* 22(2):1–25
6. Kalakota R, Robinson M (2005) The role of ERP tools in supply chain information sharing, cooperation, and cost optimization. *Int J Prod Econ* 93–94:41–52
7. Stair RM, Reynolds GW (1998) Principles of information systems: a managerial approach, 3rd edn. Course Technology, USA, p 672
8. Li J, Sikora R, Shaw MJ, Woo Tan G (2006) A strategic analysis of inter organizational information sharing. *Decis Support Syst* 42(1):251–266
9. Su Y, Yang C (2010) A structural equation model for analyzing the impact of ERP on SCM. *Expert Syst Appl* 37(1):456–469
10. Sегgie SH, Kim D, Cavusgil ST (2006) Do supply chain IT alignment and supply chain interfirm system integration impact upon brand equity and firm performance? *J Bus Res* 59(8):887–895
11. Fildes R, Goodwin P, Lawrence M (2006) The design features of forecasting support systems and their effectiveness. *Decis Support Syst* 42(1):351–361
12. Holweg M, Pil F (2008) Theoretical perspectives on the coordination of supply chains. *J Oper Manag* 26(3):389–406
13. Ketikidis P, Koh S, Dimitriadis N, Gunasekaran A, Kehajova M (2008) The use of information systems for logistics and supply chain management in South East Europe: current status and future direction☆. *Omega*, vol 36, no 4, pp 592–599
14. Choy KL, Chow HKH, Tan KH, Chan C-K, Mok ECM, Wang Q (2008) Leveraging the supply chain flexibility of third party logistics—hybrid knowledge-based system approach. *Expert Syst Appl* 35(4):1998–2016
15. Dehning B, Richardson VJ, Zmud RW (2007) The financial performance effects of IT-based supply chain management systems in manufacturing firms. *J Oper Manag* 25(4):806–824
16. Guo Z, Fang F, Whinston AB (2006) Supply chain information sharing in a macro prediction market. *Decis Support Syst* 42(3):1944–1958
17. Ke W, Liu H, Wei KK, Gu J, Chen H (2009) How do mediated and non-mediated power affect electronic supply chain management system adoption? The mediating effects of trust and institutional pressures. *Decis Support Syst* 46(4):839–851
18. Gunasekaran A, Lai K, Edwinc Cheng TC (2008) Responsive supply chain: a competitive strategy in a networked economy☆. *Omega* 36(4):549–564
19. Caddy IN, Helou MM (2007) Supply chains and their management: application of general systems theory. *J Retail Consumer Services* 14(5):319–327
20. Lin F, Kuo H, Lin S (2008) The enhancement of solving the distributed constraint satisfaction problem for cooperative supply chains using multi-agent systems. *Decis Support Syst* 45(4):795–810
21. Bozarth CC, Warsing DP, Flynn BB, Flynn EJ (2009) The impact of supply chain complexity on manufacturing plant performance. *J Oper Manag* 27(1):78–93
22. Theodorou P, Florou G (2008) Manufacturing strategies and financial performance—the effect of advanced information technology: CAD/CAM systems. *Omega* 36(1):107–121
23. Toivonen J, Kleemola A, Vanharanta H, Visa A (2006) Improving logistical decision making—applications for analysing qualitative and quantitative information. *J Purch Supply Manag* 12(3):123–134
24. Yao Y, Dresner M (2008) The inventory value of information sharing, continuous replenishment, and vendor-managed inventory. *Transp Res Part E Logist Transp Rev* 44(3):361–378
25. Trkman P, McCormack K, de Oliveira MPV, Ladeira MB (2010) The impact of business analytics on supply chain performance. *Decis Support Syst* 49(3):318–327

26. Hartono E, Li X, Na K-S, Simpson JT (2010) The role of the quality of shared information in interorganizational systems use. *Int J Inf Manag* 30(5):399–407
27. Klein R (2007) Customization and real time information access in integrated eBusiness supply chain relationships. *J Oper Manag* 25(6):1366–1381
28. Yao Y, Palmer J, Dresner M (2007) An interorganizational perspective on the use of electronically-enabled supply chains. *Decis Support Syst* 43(3):884–896
29. Samiee S (2008) Global marketing effectiveness via alliances and electronic commerce in business-to-business markets. *Ind Mark Manag* 37(1):3–8
30. Bayraktar E, Demirbag M, Koh SCL, Tatoglu E, Zaim H (2009) A causal analysis of the impact of information systems and supply chain management practices on operational performance: evidence from manufacturing SMEs in Turkey. *Int J Prod Econ* 122(1):133–149
31. de Haan J, Kisperska-Moroń D, Placzek E (2007) Logistics management and firm size; a survey among polish small and medium enterprises. *Int J Prod Econ* 108(1–2):119–126
32. Sheu J-B (2008) A hybrid neuro-fuzzy analytical approach to mode choice of global logistics management. *Eur J Oper Res* 189(3):971–986
33. Verdouw CN, Beulens AJM, Trienekens JH, Verwaart T (2010) Towards dynamic reference information models: readiness for ICT mass customisation. *Comput Ind* 61(9):833–844
34. Shukla SK, Tiwari MK, Wan H-D, Shankar R (2010) Optimization of the supply chain network: Simulation, Taguchi, and Psychoclonal algorithm embedded approach. *Comput Ind Eng* 58(1):29–39
35. Chatfield DC, Harrison TP, Hayya JC (2009) SCML: an information framework to support supply chain modeling. *Eur J Oper Res* 196(2):651–660
36. Gong Z (2008) An economic evaluation model of supply chain flexibility. *Eur J Oper Res* 184(2):745–758
37. Li Q, Zhou J, Peng Q-R, Li C-Q, Wang C, Wu J, Shao B-E (2010) Business processes oriented heterogeneous systems integration platform for networked enterprises. *Comput Ind* 61(2):127–144
38. Uçkun C, Karaesmen F, Savaş S (2008) Investment in improved inventory accuracy in a decentralized supply chain. *Int J Prod Econ* 113:546–566
39. Lo W-S, Hong T-P, Jeng R (2008) A framework of E-SCM multi-agent systems in the fashion industry. *Int J Prod Econ* 114(2):594–614
40. Li X, Wang Q (2007) Coordination mechanisms of supply chain systems. *Eur J Oper Res* 179(1):1–16
41. Thomassey S (2010) Sales forecasts in clothing industry: the key success factor of the supply chain management. *Int J Prod Econ* 128(2):470–483
42. Cheng C-B, Wang C (2008) Outsourcer selection and order tracking in a supply chain by mobile agents. *Comput Ind Eng* 55(2):406–422
43. Lin LC (2009) An integrated framework for the development of radio frequency identification technology in the logistics and supply chain management. *Comput Ind Eng* 57(3):832–842
44. Chang S, Klabjan D, Vossen T (2010) Optimal radio frequency identification deployment in a supply chain network. *Int J Prod Econ* 125(1):71–83
45. Meyer GG, Främling K, Holmström J (2009) Intelligent products: a survey. *Comput Ind* 60(3):137–148
46. Lee HT, Wu JC (2006) A study on inventory replenishment policies in a two-echelon supply chain system. *Comput Ind Eng* 51(2):257–263
47. Ge L, Voß S (2009) ERP application in China: an overview. *Int J Prod Econ* 122(1):501–507
48. Wang S, Sarker BR (2006) Optimal models for a multi-stage supply chain system controlled by kanban under just-in-time philosophy. *Eur J Oper Res* 172(1):179–200
49. Kurata H, Liu JJ (2007) Optimal promotion planning—depth and frequency—for a two-stage supply chain under Markov switching demand. *Eur J Oper Res* 177(2):1026–1043
50. Zhou W (2009) RFID and item-level information visibility. *Eur J Oper Res* 198(1):252–258

51. Emerson D, Zhou W, Piriathu S (2009) Goodwill, inventory penalty, and adaptive supply chain management. *Eur J Oper Res* 199(1):130–138
52. de la Fuente MV, Ros L, Cardós M (2008) Integrating forward and reverse supply chains: application to a metal-mechanic company. *Int J Prod Econ* 111(2):782–792
53. Li Y, Kramer MR, Beulens AJM, van der Vorst JGAJ (2010) A framework for early warning and proactive control systems in food supply chain networks. *Comput Ind* 61(9):852–862
54. Repoussis PP, Paraskevopoulos DC, Zobolas G, Tarantilis CD, Ioannou G (2009) A web-based decision support system for waste lube oils collection and recycling. *Eur J Oper Res* 195(3):676–700
55. Zhang T, Liang L, Yu Y, Yu Y (2007) An integrated vendor-managed inventory model for a two-echelon system with order cost reduction. *Int J Prod Econ* 109(1–2):241–253
56. Sumita T, Shimazaki M, Matsuyama K (2009) A proposal for inventory adjustment using ‘multiple-layers SEC–CIS model’. *Int J Prod Econ* 118(1):208–216

Designing a Model for Context Awareness Based Reliable Auction-Recommendation System (CARARS) by Utilizing Advanced Information

Hwa-Jin Park and Sang-Beom Kim

Abstract This thesis aims to solve a problem that, even though the existing auction recommending system provides auction property based on the conditions and contexts users prefer, users cannot rely on the recommended property, but must analyze its investment value or request experts to analyze it. To solve this problem, advanced information in the real estate auction, an analysis of rights, an analysis of commercial quarters, and a development plan, are classified into 5 levels indicating investment value, which will be applied at the recommendation phase. This reliable recommending service is designed to be incorporated in the current context awareness system under a smart mobile environment. Therefore, we call it context awareness-based reliable auction recommending system (CARARS).

Keywords Context awareness system · Recommendation system · Real estate auction system

H.-J. Park

Department of Multimedia Science, Sookmyung Women's University,
Hyochangwon-gil 52, Yongsan-gu, Seoul, South Korea
e-mail: phj2000@sm.ac.kr

S.-B. Kim (✉)

School of Real Estate Management, Sejong Cyber University, 111-1 Gunja-dong,
Gwangjin-gu, Seoul, South Korea
e-mail: dan@sjcu.ac.kr

1 Introduction

The auction information system has made a remarkable development as an essential element of a comprehensive information system for real estate, even since the early part of the 2000 s when private companies started to provide internet-based information up until the present, in which auction information in the Supreme Court is provided free of charge. In our current time, smart phones are diffused so prevalently that the information can be received at any time and in any place, and a context awareness system has been developed so thoroughly that real-time information that is constantly changing can be easily provided. Yet, most auction information systems for real properties offer personalized recommendation results on the basis of user profiles written at the time when customers joined, and most of the results are too multiple to be of any help. So, a user dissatisfied with the results has to select the region and kind of an auction property, every time he searches for the property. Although the results from content-based recommendations number fewer than 10, sometimes a user suffers a loss when he does not refer to advanced information comprising an analysis of rights and an analysis of commercial quarters. Subsequently, an ordinary person who is not an expert at auctioning has to refer to the advanced information of recommended properties in order to reduce such risk. Currently, however, advanced information is described in the form of a report and cannot be directly consulted.

Therefore, this research proposes turning the advanced information into indicators so that risks of investment may be reduced, that recommendations may be made according to the investment value of properties and that, at the same time, this system may be set up for recommending properties in compliance with current situation of a user. A recommendation system using the advanced information will be embedded in the context awareness auction information system suggested in [1] and will also be designed to acquire real-time contextual data such as a floating population when advanced information is analyzed.

2 Development of Advanced Information Analysis Indicators

2.1 *Advanced Information Analysis*

As for auction information, advanced information refers to well defined information regarding auction property derived from data which are collected and analyzed by an expert or a relating agency. The advanced information in real estate auction fields specifically means an analysis of rights to auction properties, an analysis of commercial quarters, and a city development plan. This advanced information is treated as a very valuable asset, provided by fee and mostly documented; so, it isn't reasonable to classify such an enormous amount of information into a few levels.

Nevertheless, it is important to provide a service, in which an ordinary person can have easy access to the information, at least through setting up appropriate levels to reduce risks.

In an analysis of rights, when an estimated price for bidding is entered, advance payments will be made from the first priority in the order of the rights, and the investment value will be decided. The investment value can be classified into 5 levels from level 1 (very risky) to level 5 (highly profitable), which are indicators. An analysis of commercial quarters is usually required when an auction property belongs to a business section, and there are various factors for the analysis. An analysis needs to be comprehensively made of a floating population at the business section, the number of competitors in each business type in the neighborhood, the sales, and the number of potential customers. As a result, estimated sales can be calculated and a decision shall be given on if the investment can be made or not. Conclusively, investment possibilities shall be divided into 5 levels. In addition, in the case of land or business section, a city development plan set up by the government must be consulted; if there is such a plan, auction properties shall be divided into 5 levels according to their location and approachability, and their investment value shall be classified. When information for an analysis of commercial quarters is collected, a context awareness system shall be used since information on hourly statistics of the population and sales can be collected in real-time.

The following Table 1 shows the types of real estate and corresponding input data, context awareness, and level of each type of advanced information.

3 Context Awareness-based Reliable Auction Recommending System Model

3.1 Design of a Recommending System

A context awareness based recommending service may be defined as an intelligent service which recommends information appropriate to the context of users after the profile by users is combined with the context that the system automatically recognized [2, 3]. In this thesis, the context awareness based service is applied to the auction field so that the conditions and specific context of users may be recognized, so as to automatically search in real-time for properties suitable to the context, analyze advanced information provided by experts, turn the information into indicators, and recommend highly profitable and reliable properties. In particular, in the case of an analysis of commercial quarters, which is essentially used when a business section is recommended, a context awareness-based system is utilized to collect information on the context in real-time. The service is called CARARS (context awareness-based reliable auction recommending system).

Table 1 Consideration and levels of advanced information

Advanced information documents	Types of real estate	Input data	Consideration	Real-time context awareness (times of updating)	5 levels
Analysis of rights	All types of real estate	Appraised price	Analysis of market price	Unnecessary (1 time/week)	1. Very risky 2. Risky 3. Cautious 4. Somewhat profitable 5. Highly profitable
Analysis of commercial quarters	Estimated bid price	Business section building	Times of failed bidding	Necessary (real-time)	1. Very risky 2. Risky 3. Cautious 4. Somewhat profitable 5. Highly profitable
Development plan	Specific store address	Business section building address	Statistics of the population at a business section	Unnecessary (when a development plan is announced)	1. No plan 2. No area affected nearby 3. Affected somewhat (nearby) 4. Effects are expected (nearby) 5. Much influence is expected (included)
	Land Address of the land	Weather a development plan is set up or not	Sales Number of competitors	Location of a business section	

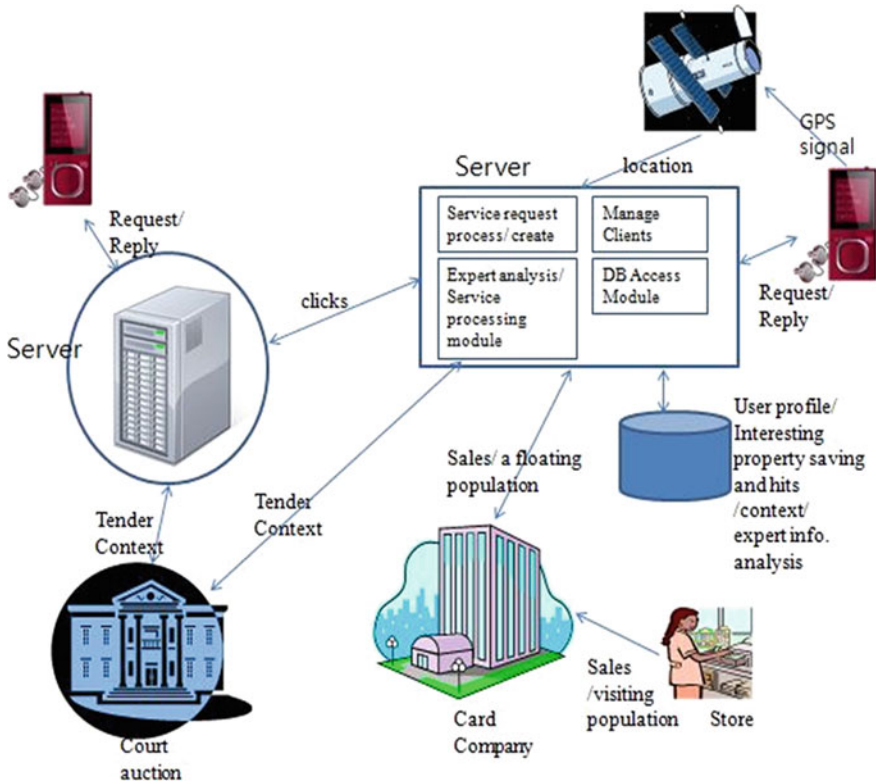


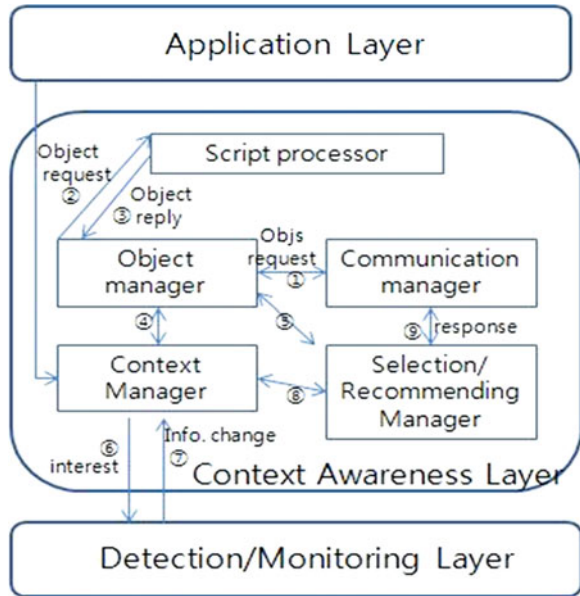
Fig. 1 Composition of CARARS

Figure 1 depicts the information necessary for context awareness, including the number of inquiries into specific properties, the number of bidders attending an open bidding offered by the court of justice, and the process to obtain advanced information necessary for the recommendation. For example, information for an analysis of commercial quarters is provided via credit card companies whenever each store sells products.

3.2 Components of the System

Basically, the components of this system are the same as those of the existing context awareness system [4]. This system is composed of 3-level conceptual classes, which includes detection & monitoring layer, context-awareness layer, and application layer. First, the detection and monitoring layer monitors computing resources or network conditions including events based on interactions with

Fig. 2 Components of CARARS



location information of remote objects or with the application program, as well as CPU use amount, memory use amount, available bandwidth, and event information on blue tooth devices and other devices. Second, the application layer functions to develop various application programs based on context awareness processing technologies for the lower part. Third, the context awareness layer plays the role of middle-wear as a core layer to process context awareness for ubiquitous computing. As is seen in Fig. 2, the context awareness layer consists of 5 managers including script processor, object manager, context manager, context selecting/recommending manager, and communication manager.

3.3 Service Flow

The following is a service flow applied to the CARARS structure in Fig. 2, in case the location of users is recognized as the context when reliable auction properties are recommended by utilizing advanced information indicators.

First, ①A user requests a recommendation of about 10 reliable auction properties that are within a certain distance from the user satisfying their preference and history, via communication manager. ②The object manager sends request message language to the script processor. ③The results of an analysis of the language made by the processor are sent to the object manager. ④Based on the information on the profile and history of the user, the object manager searches auction properties and sends context information including if the user is in a

certain distance to the context manager. ⑤The information on properties acquired based on the information on the profile and history of the user is sent to the context selecting/recommending manager. ⑥The context manager sends the context requested by the user to the detecting/monitoring object, or to the location of the user in this case. ⑦The context manager updates the location value or the most recent context requested by the user of the detection/monitoring layer in the lower part. ⑧ The 10 most profitable properties are selected among objects searched by the object manager, which satisfy the context of the user and are reliable, by utilizing advanced information prepared by experts. ⑨The results of the screening are provided to the user via the communication manager.

3.4 Structure of Database

In order to implement CARARS, DB consisting of three advanced information and DB consisting of five objects and context information are constructed. Among these, auction property DB and user profiles belong to object DB; inquiry context DB, user context DB, and open bidding DB belong to context DB and are updated in an event-like manner.

- Auction property DB (court of justice, auction case number, division, location, bidding date, appraised price, lowest price, area, times of failed bidding, photos of the surroundings)
- Profile of the user (ID, resident number, telephone number, sex, conditions of division, conditions of location, the lowest price, conditions of the area)
- History file of the user (ID, number of an interested registered property, number of inquiries)
- User context DB (ID, mobile object ID, current location)
- Inquiry context DB (auction property number, court of justice, auction case number, external server number, number of inquiries)
- Open bidding DB (auction property number, court of justice, auction case number, bidding number, bidding period, number of bidders, estimated highest bid price)
- Analysis of rights (court of justice, auction case number, ID, estimated bid price, market price, levels)
- Analysis of commercial quarters (court of justice, auction case number, floating population, sales, number of competitors, levels)
- Development plan (ID, location, scheme drawing)

4 Conclusions and Future Works

This thesis is aimed at solving a problem that users cannot rely on the recommended property even though the existing auction recommending system provides auction property that are based on the conditions and contexts users prefer. To solve this problem, three advanced information in the real estate auction—an analysis of rights, an analysis of commercial quarters, and a development plan—are turned into one of 5 levels indicating investment value to select high profits. This reliable recommending service is embedded in a context awareness system under smart mobile environment. Therefore, CARARS enables users carrying smart phones to have profitable auction properties recommended on their mobile device if they are located nearby, and to receive advanced information including market price and estimated highest bid price if they desire. Also, it is prospected that the users can be provided with advanced information including the number of bidders for properties which are on period bidding; therefore, the value of information will increase and possibly will be provided by fee in the future.

For future works, this recommending system should be implemented in the near future and more a concrete ranking procedure among several analyses should be developed to improve this research.

References

1. Park H (2011) Study on the context awareness-based real estate auction information system under the smart phone environment. *J Digit Contents Soc* 12(4):585–592
2. Woerndl W, Schueller C, Wojtech R (2007) A hybrid recommender system for context-aware recommendations of mobile applications. In: *Proceedings of the 2007 IEEE 23rd international conference on data engineering workshop*, pp 871–878
3. Kim S, Oh B, Kim M, Yang J (2012) A movie recommendation algorithm combining collaborative filtering and content information. *J Korea Inform Sci Soc Softw Appl* 39(4):261–268
4. Shim C, Tae B, Chang J, Kim J, Park S (2006) Implementation of an Application System using Middleware and Context Server for Handling Context-Awareness. *Journal of the Korea Information Science Society: Computing Practices*, 12(1):31–42

Part III
Security Protocols and Applications

A Secure Self-Encryption Scheme for Resource Limited Mobile Devices

Yongjoo Shin, Seungjae Shin, Minsoo Kim and Hyunsoo Yoon

Abstract Recently, IT and mobile technology are developed rapidly. Mobile devices such as a smartphone and tablet PC have become widely popular, and many people store their private information into the devices. However, the portability and mobility of devices take risks such as being lost or stolen. These devices are the one of main cause to leak the sensitive and confidential information. To protect the information leakage from devices, the encryption algorithm is required. The existing encryption algorithms take long delay time and heavy battery consumption in mobile devices with restricted resources. Previously, the Self-Encryption (SE) which is a lightweight encryption algorithm for mobile devices was proposed, which had critical weaknesses. This paper proposes a novel SE scheme with a random permutation and bit-flipping process. Our proposed scheme derives the keystream from the plaintext itself, but the statistical correlations are effectively removed from the novel randomization process. It gives a solution to overcome weaknesses of the original SE scheme and the complexity to make adversaries it difficult to launch a brute force attack, and satisfies a 0/1 uniformity of key and cipher stream, which is an important property of the secure stream cipher. Moreover, it is faster and more energy-efficient than other ciphers such as AES and RC4. The evaluation is performed by the Android platform.

Y. Shin · S. Shin · M. Kim · H. Yoon (✉)
Department of Computer Science, KAIST, Guseong-dong, Yuseonggu,
Daejeon, South Korea
e-mail: hyoon@nslab.kaist.ac.kr

Y. Shin
e-mail: yj_shin@nslab.kaist.ac.kr

S. Shin
e-mail: sjshin@nslab.kaist.ac.kr

M. Kim
e-mail: mskim@nslab.kaist.ac.kr

The delay time and the battery consumption are measured and analyzed, which show that the proposed scheme is secure and best suited to the mobile platform.

Keywords Data security · Lightweight encryption · Mobile device

1 Introduction

The evolution and popularization of mobile devices have dramatically changed the communication paradigm. More and more people tend to store the private and sensitive information into their devices, and these devices are likely to be lost or stolen. Generally, the data without encryption can be used in other crimes such as the identity theft by an adversary obtaining the devices illegally. Although the manufactures of devices provide a remote service to protect the mobile device, they can be easily crippled by disabling network services through removing a battery before operating such programs. To protect the information leakage from lost or stolen devices, the data encryption is an essential process. However, Since existing encryption algorithms require sufficient computing resources, they are not suitable for the mobile device with restricted resources.

The Self-Encryption (SE) is proposed by Chen and Ku [1]. This algorithm is lightweight for mobile devices, which generates the keystream from the plaintext itself. However, the weaknesses of the SE scheme are discovered by Gasti and Chen [2]. To overcome the weaknesses, this paper proposes a novel SE scheme with a random permutation and bit-flipping process. Using the novel process, our proposed scheme is able to generate a desirable pseudorandom keystream. In addition, our experimental results show that it has better performance than other conventional encryption algorithms such as AES and RC4 in terms of the delay time and the energy consumption in the Android platform.

The rest of this paper is organized as follows: the concept of self-encryption is introduced in Sect. 2. We proposed a novel SE scheme with a random permutation and bit-flipping to overcome weaknesses in Sect. 3 and show the security analysis and performance efficiency of our proposed scheme in Sect. 4. In Sect. 5, we summarize our work and make a conclusion.

2 The Overview: the Self-Encryption Scheme

The main ideas of the SE scheme [1] are extracting the keystream from the plaintext itself and managing it in a secure remote server at the Internet Service Provider (ISP) side. The randomly extracted keystream is treated as the secret information and stored into the remote server. Authors insisted that this scheme provides a short PIN to generate a high entropy keystream bit sequence and that

the simple computational operation makes it easy to be implemented on most mobile platforms. Unfortunately, the weaknesses of Chan's SE scheme are discovered by [2].

First weakness is that Chen's SE scheme is not secure against Indistinguishability under Chosen Plaintext Attack (IND-CPA). IND-CPA is a mathematical challenge-response game to guess which one of two messages is encrypted when an adversary sends two messages into a challenger to encrypt. Generally, the adversary has a $1/2$ probability to guess correctly. However, if the adversary has more than $1/2$ probability, the encryption algorithm is not secure. In Chen's SE scheme, the adversary sends two messages, m_0 and m_1 to challenger, where m_0 is a random bit string, and m_1 is a string consisted of all 1s bits. If the challenger encrypts the m_0 , then the ciphertext is a random string, and if encrypting the m_1 , then the ciphertext is all 0s string. Because the adversary can predict the result, she has more $1/2$ chance to guess correctly.

Another weakness is the statistical bias problem. In Chen's SE scheme, the keystream is uniformly selected from the plaintext itself. That is, if the plaintext has some statistical patterns of the string, then the keystream and ciphertext must have the same patterns. Such a bias problem makes it easy to mount a ciphertext-only attack. An adversary can obtain the meaningful information about the sensitive message by observing the ciphertext statistically.

3 Proposed Scheme: A Secure Self-Encryption

In this section, we introduce our proposed scheme with three subsections. In 3.1, a protocol framework for the mobile telecommunication infrastructure is presented. The encryption and decryption procedures are explained in 3.2 and 3.3, respectively.

3.1 Basic Framework

Figure 1 shows how our proposed self-encryption system is operated in a cellular mobile environment. As seen in the figure, the encryption procedure transforms the plaintext file stream into ciphertext stream by using the master secret value set by the user (ex: device PIN or file password). Next, the mobile device transfers the secret information, which is required for decryption, to the Authentication Center (AuC) of the Internet Service Provider through a secure channel [3]. After that, the secret information is removed from the device. Whenever the mobile user wants to access the encrypted file, the device should request the secret information to the AuC. If the device is lost or stolen by someone, the AuC denies providing the secret information in order to prevent the illegal decryption of the protected file.

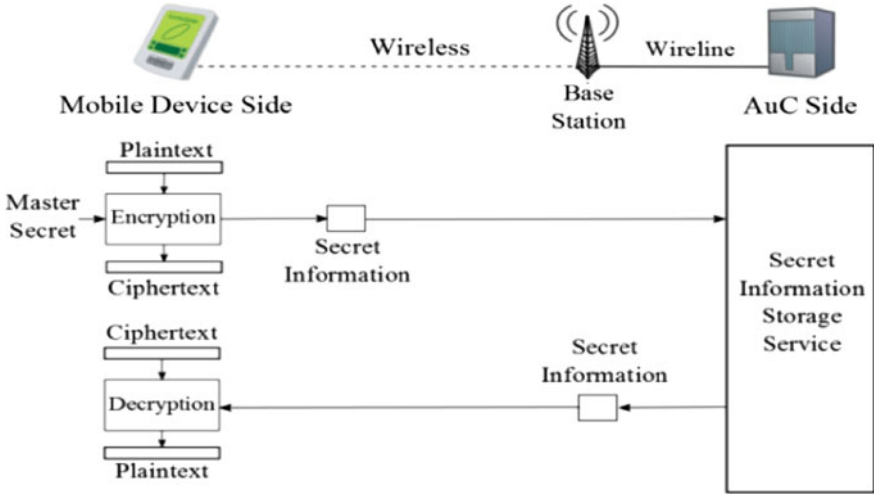


Fig. 1 System model and basic framework

3.2 Encryption Phase

In our proposed scheme, we use pseudo-random permutation function P_r as mathematical constructs. It is defined as

$$P_r = P(S, s, r), \quad (S \in 2^{\mathbb{N}} \text{ and } s, r \in \mathbb{N})$$

where \mathbb{N} is a set of natural number(positive integer), $2^{\mathbb{N}}$ is its power set. In equation, P_r represents a randomly selected r -permutation from S . s is a seed number for pseudorandom permutation.

Figure 2 (a) describes the encryption phase as a block diagram. The encryption procedure takes the master secret $s \in \mathbb{N}$, security strength factor $\Delta \in [0, 1]$ and plaintext bit stream \mathcal{P} as an input data, and performs following steps:

- Step 1 Let us denote $m = |\mathcal{P}|$ as a bit-length of \mathcal{P} . Then, determine the length of the key stream as $j = \Delta^1 \cdot m^1$
- Step 2 Let \mathbb{S}_m denote $\{1, 2, \dots, m\}$. And derive a random permutation of it, which is denoted as $P_j = P(\mathbb{S}_m, s, j)$ ²

¹ If Δ is too small, the security strength is weakened. Although higher Δ means more strong security, computational overhead also increases. If $\Delta = 1$, our self-encryption becomes equivalent to the *one-time pad* which provides unconditional secrecy [4].

² $P(\mathbb{S}_m, s, j)$ can be implemented by random shuffling algorithms like *Fisher-Yates* [5] and *Sattolo's* [6] shuffle.

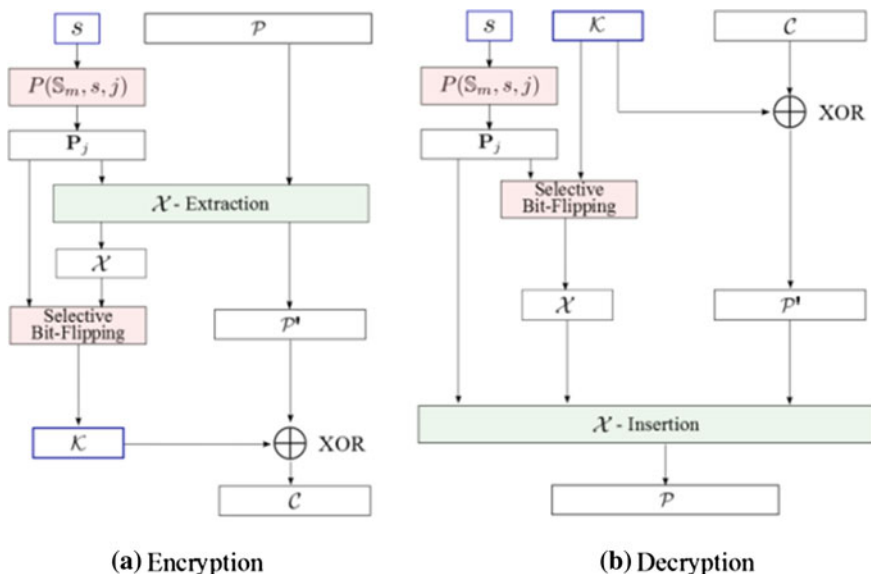


Fig. 2 Encryption and decryption process in a Novel SE scheme **a** encryption **b** decryption

- Step 3 Suppose P_j is presented as (p_1, p_2, \dots, p_j) , ($p_k \in \mathbb{S}_m$ where $k = 1, 2, \dots, j$). Then, extract p_k -th bit for every $k = 1, 2, \dots, j$ from \mathcal{P} , and make j -bit string by concatenating them. We denote this j -bit string as \mathcal{X} . After extracting, flip the k -th bit in the concatenated string if $p_k \bmod 2 = 0$. We denote this flipped string as \mathcal{K} and name it *the secret stream*. For example, when $\mathcal{P} = 0110001010\dots$ and $P_j = (7, 2, 5, \dots)$, \mathcal{K} has 7-th bit, the negation of 2-nd bit, and 5-th bit of \mathcal{P} as its starting 3 bits respectively. Therefore, the first 3-bit suffix of \mathcal{K} is 100
- Step 4 For every p_k -th ($k = 1, 2, \dots, j$) bit in \mathcal{P} , replace them with empty symbol λ . In other words, extracted bits are removed from the original plaintext. This results in the shortened plaintext \mathcal{P}' of which length is m' ($= m - j$)
- Step 5 Now, perform XOR operation between \mathcal{P}' and \mathcal{K} . If $m' > j$, XOR the x -th bit of \mathcal{P}' with $((x - 1) \bmod j) + 1$ -th bit of \mathcal{K} . By this way, the ciphertext stream \mathcal{C} is eventually obtained

After the encryption is completed, the device transfers the master secret s and the secret stream \mathcal{K} to the AuC through a secure channel. They are the secret information described in 3.1. The mobile device stores only the final ciphertext \mathcal{C} .

3.3 Decryption Phase

As explained in 3.1, decryption phase starts with receiving the master secret s and the secret stream \mathcal{K} . Figure 2 (b) shows the block diagram of decryption procedure in our scheme.

- Step 1 By using $j = |\mathcal{X}|$ and s , compute P_j as in the same manner as the step 2 in 3.2
- Step 2 Same as in step 3 in 3.2, derive \mathcal{X} from \mathcal{K} . We have to perform selective bit-flipping before insertion. Same as in step 3 in the encryption phase, when P_j is presented as $(p_k)_{k=1,2,\dots,j}$, and if $p_k \bmod 2 = 0$, k -th bit of \mathcal{K} must be flipped (Boolean negation) before insertion and if $p_k \bmod 2 = 1$, k -th bit of \mathcal{X} must be flipped (Boolean negation) before insertion
- Step 3 Obtain \mathcal{P}' by XOR-ing \mathcal{C} with \mathcal{K}
- Step 4 Now, for all bits in \mathcal{X} , insert each of them into \mathcal{P}' . At this step, thanks to P_j , we can determine the correct insertion point for each bit from \mathcal{X}

After step 4, we obtain the plaintext \mathcal{P} which is merged from \mathcal{X} and \mathcal{P}' .

4 Performance Evaluation

4.1 Security Analysis

Just as the Chen's SE scheme, the variable length of the keystream makes it difficult to launch a brute force attack, where the complexity is bounded to $O(2^m)$. And the complexity of the permutation is $O(m^j) \approx m!/(m-j)!$. Then the total complexity is $O(2^m m^j)$, which is relatively secure compared to other stream ciphers.

The major differences from Chen's SE are using the pseudo-random permutation function and the process of bit-flipping. Especially, bit-flipping mechanism changes approximately the half of all bits in random permuted bit-stream (\mathcal{K} in Sect. 3) because the keystream is generated by extracting the almost same amount of even and odd position of the plaintext through the pseudo-random permutation function ($P(\mathbb{S}_m, s, j)$). In addition, as long as the secret information s is protected, it is difficult to know which positions of the keystream are flipped. Such a process is able to resist the IND-CPA because the uniform bit string which the adversary selects is encrypted into the random bit string. In addition, due to switching the random position of the keystream, the particular pattern of the plaintext doesn't appear in the ciphertext. In short, our proposed scheme significantly outperforms Chen's one in terms of the confusion property which is an important property of information secrecy.

As one of the most important design consideration for a stream cipher, the keystream should be an approximate the properties of a true random number stream as close as possible. That is, there should be an approximately equal number of 1s and 0s. The more random-appearing the keystream is, the more randomized the ciphertext is, making cryptanalysis more difficult [4]. Our proposed scheme can make the keystream with the property of randomness using a novel combination of random permutation and bit-flipping process.

Theorem The probability to assign 0 or 1 to any bit in both the keystream and ciphertext takes 1/2 by our proposed scheme.

Proof Let us denote with $\text{str}[i]$ the i th bit of a bit string str . We have that

$$\begin{aligned} \Pr[ks[i] = 0] &= \Pr[pl[i] = 0] \cdot \Pr[i \bmod 2 = 1] + \Pr[pl[i] = 1] \cdot \Pr[i \bmod 2 = 0] \\ &= \Pr[pl[i] = 0] \cdot \left(\frac{1}{2} + \delta\right) + \Pr[pl[i] = 1] \cdot \left(\frac{1}{2} - \delta\right) \\ &= \frac{1}{2}(\Pr[pl[i] = 0] + \Pr[pl[i] = 1]) + \delta(\Pr[pl[i] = 0] - \Pr[pl[i] = 1]) \\ &= \frac{1}{2} + \delta \end{aligned}$$

$$\Pr[ks[i] = 1] = 1 - \Pr[ks[i] = 0] = \frac{1}{2} - \delta$$

$$\therefore \Pr[ks[i] = 0] \approx \Pr[ks[i] = 1] = \frac{1}{2} \pm \delta$$

$$\begin{aligned} \Pr[c[i] = 0] &= \Pr[p'[i] = 0] \cdot \Pr[ks[i] = 0] + \Pr[p'[i] = 1] \cdot \Pr[ks[i] = 1] \\ &= \Pr[p'[i] = 0] \cdot \left(\frac{1}{2} + \delta\right) + \Pr[p'[i] = 1] \cdot \left(\frac{1}{2} - \delta\right) \\ &= \frac{1}{2}(\Pr[p'[i] = 0] + \Pr[p'[i] = 1]) + \delta(\Pr[p'[i] = 0] - \Pr[p'[i] = 1]) \\ &= \frac{1}{2} + \delta \end{aligned}$$

$$\Pr[c[i] = 1] = 1 - \Pr[c[i] = 0] = \frac{1}{2} - \delta$$

$$\therefore \Pr[c[i] = 0] \approx \Pr[c[i] = 1] = \frac{1}{2} \pm \delta \square$$

Therefore, our proposed scheme has strong properties of diffusion and confusion [7]. Moreover, it has a uniformly distributed keystream and output which are fundamental requirements of any secure stream cipher [8].

4.2 Efficiency Analysis

In order to evaluate the performance of our proposed scheme, we measured the delay time and the amount of battery consumption of RC4, AES and our proposed scheme. Table 1 shows the experimental environment to measure the performance. We implement our scheme and compare it with the AES and RC4 which are existing algorithms for file encryption. In our proposed scheme implementation, we set the security strength factor Δ ($2.4/10^5 \approx 256\text{bit}$ per 10Mbit).

The encryption process is performed with the different file size. We assume that a video file including the confidential conference is the biggest size of file to access in the mobile devices and that its size is approximately 100 MB. The delay time means the time taking in order to produce a ciphertext from a plaintext. The amount of battery consumption is measured by checking the current device battery state before and after encryption process.

Figure 3 (a) shows the delay time of encrypting a variable file size from 10 KB–100 MB. With small size files, all algorithms tend to take similar delay time. However, as the file size increases, the delay time of AES and RC4 take much longer time than our proposed scheme. In terms of energy efficiency, in Fig. 3 (b), we recognize that the battery consumption of our scheme outperforms the other algorithms. Thanks to a simple methodology to generate the keystream and XOR operating to encrypt, the performance of our proposed scheme is much faster and more efficient than other conventional block cipher and stream cipher, AES and RC4.

Table 1 The Environment to measure the performance

	Specification
Product	Samsung Galaxy Tab 10.1
OS	Google Android 4.0.4 (Ice Cream Sandwich)
CPU	1 GHz dual-core Nvidia Tegra 2 process
Memory/storage	1 GB DDR2/64 GB flash memory
Power	Lithium-Ion 7,000 mA h battery

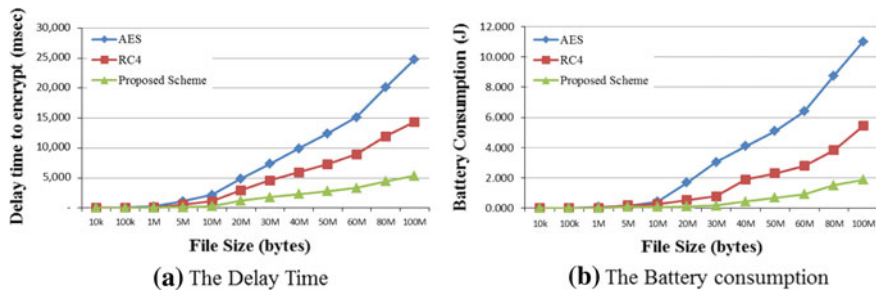


Fig. 3 The results of delay time and battery consumption **a** the delay time **b** the battery consumption

5 Conclusion

The theft and loss of mobile devices may provoke the leakage of sensitive and confidential information. However, conventional encryption schemes require considerable computational burden and energy consumption to prevent the private information leakage from the mobile devices which have restricted resources. There is a tradeoff between strong security and lightweight process.

This paper proposed a novel lightweight SE scheme with a random permutation and bit-flipping process to overcome the weaknesses of the Chen's SE scheme. Our proposed scheme provides robust complexity against IND-CPA and satisfies the property of 0/1 uniformity which is one of the most important requirements of the stream cipher. Through experimental measurement, we confirm that our proposed scheme is faster and better energy efficient than the AES and RC4 cryptosystem.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea Government (MEST). (No. 2012-0005390).

References

1. Chen Y, Ku WS (2009) Self-encryption scheme for data security in mobile devices. In: IEEE consumer communication and networking conference (CCNC)
2. Gasti P, Chen Y (2010) Breaking and fixing the self-encryption scheme for data security in mobile devices. In: Proceedings of the 18th Euromicro conference on parallel, distributed and network-based processing (PDP)
3. 3GPP Std. TS33.220 v11.3.0 (2012) Generic authentication architecture (GAA); generic bootstrapping architecture (GBA)
4. Stallings W (2003) Cryptography and network security, 3 edn. Prentice Hall, Philadelphia, pp 43–44, 66–67, 192–194
5. Knuth DE (1998) The art of computer programming. Seminumerical algorithms, vol 2. Addison-Weseley
6. Wilson MC (2005) Overview of sattolo's algorithm. Technical Report 5542, INRIA
7. Shannon CE (1949) Communication theory of secrecy systems. Bell Syst Tech J 28:656–715
8. Katz J, Lindell Y (2007) Introduction to modern cryptography. Chapman and Hall/CRC, Boca Raton

A Study on the Network Stability Measurement Based on an Articulation Nodes with Weight Value

Myung-Ki Jung and Seongjin Ahn

Abstract The concept of network stability has been introduced to ensure the easy management of the actual complex networks. The network stability is a concept that allows the network administrator to easily provide the network management. There have been relevant advanced studies, which were meaningful in that they measured the network stability, but they did not address the subject deeply enough. In this study, a new network stability calculation method was proposed using the weight of the articulation node based on the existing studies. The method proposed in this study offers more precise information about the weakness of the network, and is expected to provide more accurate stability and weakness information than those from the existing method.

Keywords Network stability · Articulation node · Unicast connection · Subgraph

1 Introduction

A network consists of edges that connect different nodes, and allows the data exchange through the cables between hosts. The initial network was relatively simple. Therefore, it was easy to manage. However, with the increase in its size, networks became more and more complex, and now need detailed and effective management.

M.-K. Jung · S. Ahn (✉)

Department of Computer Education, Sungkyunkwan University,
Myeongnyun 3-ga, Jongno-gu, Seoul, South Korea
e-mail: sjahn@skku.edu

M.-K. Jung

e-mail: uyh45@skku.edu

Accordingly, the network stability has been proposed as a concept for a higher level network management. The network stability shows the comprehensive stability of the network, and presents overall management information about the current network in a relatively simple method. This standard is meaningful in managing the network, and needs to be studied.

There have been many researches related to network stability, as network stability is considered as a crucial concept. All of them were meaningful in themselves. However, it was found from the survey on the advanced studies that these study results could be further expanded. Accordingly, a new network stability calculation method was proposed in this study with what was not considered in the previous studies.

2 Articulation Node and Stability

2.1 Basic Definition and Features of Articulation Nodes

A network can be expressed in a form of an undirected graph. In the network expressed as an undirected graph, the node that exists in the form of an articulation point is an articulation node. The articulation node is a component that connects the biconnected component (part that does not include the articulation node and has the form of a connected undirected graph, in the connected undirected graph that represents the entire network), an only route between two network sets, and an only node that is shared by two biconnected components. It normally handles a high amount of traffic. If the articulation node is attacked or it does not operate, problems occur in the connection between networks. Therefore, articulation nodes are weak points of the network.

2.2 Existing Study on the Network Stability Measurement Method Using Articulation Nodes and Adjacent Nodes

This method represents the weakness level using the number of adjacent nodes that are connected to articulation nodes in the network. The number of nodes that are connected to the articulation node is proportional to the number of disconnections when a problem occurs to the articulation node. Therefore, the number of adjacent nodes is used as a main parameter of the stability equation.

In this method, if an articulation node exists, the network stability is the reciprocal number of the ‘average number of adjacent nodes of many articulation nodes in the network’.

$$N_{\text{Stability}} \begin{cases} \frac{n}{\sum_{i=1}^n T_i} & (\text{if } n \geq 1) \\ 1 & (\text{if } n = 0) \end{cases}$$

$N_{\text{Stability}}$: Stability in Network
 n : Number of Articulation node
 T_i : Number of Adjacent node

3 Algorithm for Calculating Network Stability Based on Weight

3.1 A Formula to Calculate Network Stability

In this study, the network stability is represented using the number of articulation nodes and the importance of each articulation node. This is because the articulation nodes are weak points in the network, and the importance represents the negative effect of the damaged articulation point on the entire network.

When the network stability is represented based on these two factors, the stability should become lower with a greater number of weak points and a higher negative effect of a damaged weak point. The importance is represented by the weight, and the following stability equation is obtained.

$$N_{\text{Stability}} \begin{cases} \frac{1}{\sum_{i=1}^n W_i} & (\text{if } n \geq 1) \\ 1 & (\text{if } n = 0) \end{cases}$$

$N_{\text{Stability}}$: Stability in Network
 n : Number of Articulation node
 W_i : Weight of Articulation node

The equation shows that the weight of the articulation node in the network is a critical factor for the network stability. Because the weight plays an important role, the method of finding the weight of the articulation node must be clearly defined to calculate the accurate stability.

3.2 A Method to Weight Articulation Nodes in Network

The meaning of the weight should be examined prior to study the weight of the articulation node. Given to each node that composes the network, the weight represents the comprehensive size of damage due to the corresponding damaged node to allow easy understanding of importance of each part.

From this point of view, the weight of the articulation node should include both the importance of the corresponding node and the damage from the disconnection due to the corresponding damaged articulation node. (In addition, only the unicast connection is considered as a connection in this study.)

The weight of the articulation node is defined as follows.

$$\text{(Weight of articulation node)} = \text{(Size of damage to the corresponding damaged articulation node)} + \text{Size of damage from the unicast disconnection due to the corresponding damaged articulation node}$$

The weight must be given to all nodes in the network to calculate the weight of the articulation node in the network as defined above. This process is required to calculate the size of the damage due to disconnection. In this case, the weight must be independently given to all nodes considering only the node itself, regardless of the connection relationship with other adjacent nodes.

From this process, the independent weight of each node can be examined. Accordingly, the size of damage to the articulation node can be identified, without considering the connection between nodes.

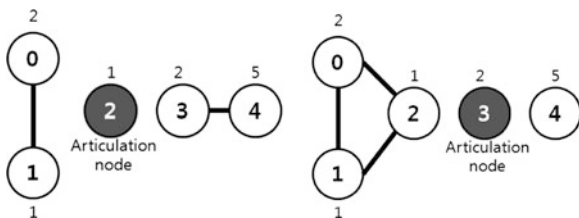
After the weight is independently given by node, all articulation nodes in the network are found, and the damage in terms of unicast connection due to each articulation node should be measured. To measure the damage in terms of connection, many disconnections due to the articulation node must be considered.

To examine the disconnections due to the articulation node, the breakdown in each articulation node must first be assumed. If the breakdown in each articulation node is assumed one by one, it can be seen that the network is divided into two or more subgraphs. (The subgraph includes the articulation node itself.)

As shown in Fig. 1, the network is divided into subgraphs differently in each articulation node cases. Then each articulation node division case must be examined one by one, as the degree of damage due to disconnection is different each other.

When we examine each articulation node, subgraphs made by damage to articulation node can be found. In this case, disconnections between arbitrary two subgraph out of all the subgraphs created as a result of breakdown in articulation node should be considered to measure the degree of damage from disconnections caused by the breakdown of each articulation node, because these disconnections

Fig. 1 Subgraphs created by an articulation node



between subgraphs indicates damage from disconnections due to breakdown in articulation node.

So, the damage due to the disconnection between subgraphs must be calculated. For the unicast connection between subgraphs, the index for representing the potential damage due to disconnection and the probability of each connection in the network are required. In this case, all the disconnections between subgraphs should be examined case by case.

The potential damage due to one disconnection can be defined as follows.

$$D = \sum_{i=1}^n P_i W_i$$

- D: Damage to the unicast connection due to the specific articulation node damage
- n: The number of unicast connections disconnected due to the specific articulation node damage

$$n = \frac{(\text{Number of subgraphs for this articulation node})!}{(\text{Number of subgraphs generated for this articulation node} - 2)! * 2!}$$

- P_i : Generation probability of a unicast connection that crosses the corresponding articulation node
- W_i : Potential damage expected from one unicast disconnection that crosses the corresponding articulation node

The potential damage expected from a disconnection that crosses the articulation node must be presented. For this purpose, the weights of two subgraphs with the corresponding connection are calculated and summed up. This is because the unicast connection is formed through the articulation node between two subgraphs, and they influence the weight of the articulation node.

$$W_i = W_{\text{sub1}} + W_{\text{sub2}}$$

- W_i : Potential damage expected from the disconnection of a unicast connection that crosses the articulation node
- W_{sub1} : Weight of a subgraph that is one end of the corresponding unicast connection
- W_{sub2} : Weight of a subgraph that is the other end of the corresponding unicast connection

To calculate the weight of the subgraph, all the initially given weights for all nodes that belong to the subgraph are summed up. This is because the weight of a subgraph is influenced by the weights of all nodes in the subgraph.

$$W_{\text{sub}} = \sum_{i=1}^n \omega_i$$

- W_{sub} : Weight of a subgraph
- n : Number of nodes in a subgraph
- ω : Weight of a node that exists in the corresponding subgraph

Next, the probability must be calculated for either of the unicast connections between two subgraphs that do not include the articulation node or the connections between a subgraph and articulation node. The probability can be expressed as follows.

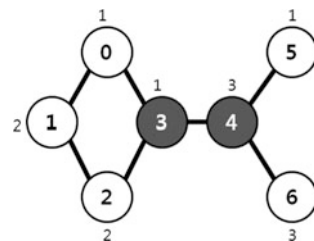
$$P_i = \frac{ab}{\frac{m!}{(m-2)! 2!}}$$

- P_i : Generation probability of either of the unicast connections between two subgraphs
- m : Number of nodes that compose the entire network
- a : Number of nodes of the subgraph that is one end of the corresponding unicast connection
- b : Number of nodes of the subgraph that is the other end of the corresponding unicast connection

4 Test and Analysis

The articulation node weight calculation method and the resulting network stability calculation method were addressed above. The understanding of the application of this method to actual examples is also required. Therefore, the method

Fig. 2 Network composed of weighted nodes



proposed in this study needs to be understood by comparing this method with those in the existing study.

The two following tests show that the proposed network stability calculation method differs from the existing method. The network stability is examined according to each method, and the characteristics of the method are presented.

A network which consists of nodes which are weighted without considering the connection relationship was formed as shown in Fig. 2.

First, the results of a prior study that used the method similar to this method are examined. In the prior study, the number of adjacent nodes of the articulation node is used. For all articulation nodes, the number of the adjacent nodes is calculated and summed up, and the reciprocal number of the result is obtained.

With this method, the network stability is calculated as follows.

$$N_{\text{Stability}} = \frac{(\text{Number of articulation nodes})}{(\text{Number of adjacent nodes of node 3})+(\text{Number of adjacent nodes of node 4})} = \frac{2}{3+3} = \frac{1}{3} = 0.3333$$

Then the network stability is calculated using the method proposed in this study. The articulation nodes in the network example are nodes 3 and 4. Therefore, the independent weights of the articulation nodes, in which only the weight of the node itself is considered, can be identified.

Next, the adjacent connection relationship of the articulation nodes must also be identified. The relationship must be separately considered for each articulation node, because disconnections caused by damage to each articulation node is different each other.

4.1 Calculation for an Articulation Node 3

Node 3 divides the whole network into three subgraphs (including the articulation node), as shown above. Accordingly, the connection between subgraphs must be considered (Fig. 3).

In this case, the connections between {0, 1, 2} and {4, 5, 6}/ {0, 1, 2} and {3}/ {4, 5, 6} and {3} must be examined (Table 1).

Fig. 3 Subgraphs created by damaging articulation node 3

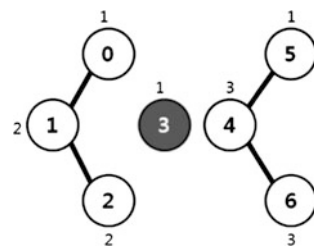


Table 1 Calculation for weight of articulation node 3

Connection	Size of damage due to disconnection of the corresponding unicast connection
{0, 1, 2} and {4, 5, 6}	$\frac{3 \times 3}{21} \{(1 + 2 + 2) + (3 + 1 + 3)\} = \frac{108}{21}$
{0, 1, 2} and {3}	$\frac{3 \times 1}{21} \{(1 + 2 + 2) + 1\} = \frac{18}{21}$
{4, 5, 6} and {3}	$\frac{1 \times 3}{21} \{1 + (3 + 1 + 3)\} = \frac{24}{21}$

Accordingly, the final weight of node 3 is $1 + \frac{108}{21} + \frac{18}{21} + \frac{24}{21} = \frac{171}{21}$.

4.2 Calculation for Articulation Node 4

Node 4 divides whole network into four subgraphs (including the articulation node), as shown above. Accordingly, the connection between subgraphs must be considered (Fig. 4).

In this case, the connections between {0, 1, 2, 3} and {4}/{0, 1, 2, 3} and {5}/{0, 1, 2, 3} and {6}/{4} and {5}/{4} and {6}/{5} and {6} must be examined.

The weight of articulation node 4 is calculated just like calculation for node 3 as follows.

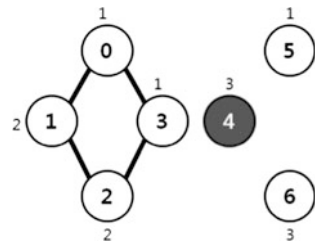
$$3 + \frac{36}{21} + \frac{28}{21} + \frac{36}{21} + \frac{4}{21} + \frac{6}{21} + \frac{4}{21} = \frac{177}{21}.$$

So, the network stability is calculated using the resulting weight of the articulation node.

$$N_{\text{Stability}} = \frac{1}{\frac{171}{21} + \frac{177}{21}} = \frac{1}{\frac{348}{21}} = \frac{21}{348} = 0.0603$$

From the test results, the prior study and this study provide clearly different methods. The two methods produce very different results. This is because the existing calculation method with the simple number of adjacent nodes does not address the damaged connections from the damaged articulation node. Meanwhile, the method in this study divides the network into subgraphs with the articulation node as the reference, and considers the connection relationship between

Fig. 4 Subgraphs created by damaging articulation node 4



subgraphs. Therefore, the method proposed in this study actually addresses the data transfer in the network.

5 Conclusions

In this study, the unicast-connection-based articulation node weight calculation method and the network stability method were proposed. In the prior study, only the adjacent nodes were considered in the network stability calculation process, and the connection between arbitrary two points in the network was not considered. To complement this, a new articulation node weight calculation method was proposed to increase the accuracy. Therefore, more accurate network stability can be identified using the proposed method.

Based on this articulation node weight calculation method, a further study will address the method of efficiently reacting to the weak points. The study is required because there is a cost limit in managing the network, and weak points must be efficiently removed using the limited cost.

Acknowledgments This work was supported by a grant from Kyonggi university advanced Industrial Security Center of Korea Ministry of Knowledge Economy.

References

1. Kim Y, Ahn S, Chung J-w (2007) Importance of network articulation node of traffic base analysis techniques. In: Proceedings of the digital contents society, pp 113–117
2. Kim Y, Kim H, Chung J-w, Cho K-j, Yu K-s (2008) Network stability assessment using the number of tree adjacent to an articulation node. In: Gervasi O, Murgante B, Laganá A, Taniar D, Mun Y, Gavrilova ML (eds) ICCSA 2008, Part II. LNCS, vol 5073, pp 1229–1241
3. Weiss MA (1999) Data structures and algorithm analysis in C++, Addison-Wesley, Reading
4. Ahn B, Chung J-w (2001) A study on the scheme for gathering network management information using biconnected component computation in internet. Thesis of Master course, Sungkyunkwan University
5. Kim Yh, Kim K, Lim CS, Ahn S, Chung J-w (2008) Technique of analysis of network's importance through analysis on the relationship between an articulation node and its adjacent nodes. In: GESTS, pp 1–8
6. Chaudhuri P (1998) An optimal distributed algorithm for finding articulation points in a network. *Comput Commun* 21(18):1707–1715
7. Kim Y, Kim K, Ahn S, Chung J-w (2008) Network stability analysis techniques using the virtual articulation node. In: Asia-Pacific network operations and management symposium, challenges for next generation network operations and service management, pp 458–461
8. Chaudhuri P (1998) An optimal distributed algorithm for finding articulation points in a network. *Comput Commun* 21(18):1707–1715

9. Huang ST (1989) A new distributed algorithm for the biconnectivity problem. In: Proceedings of international conference on parallel processing, vol 3, pp 106–113
10. Chung Y, Chung J-w (2011) Design and implementation of articulation node searching system for the improvement of network weak point. Thesis of Doctor course, Sungkyunkwan University

An Improved Forward Secure Elliptic Curve Signcryption Key Management Scheme for Wireless Sensor Networks

Suman Bala, Gaurav Sharma and Anil K. Verma

Abstract The concept of forward secrecy is extended to wireless sensor networks where it is frequent that nodes run out of energy and new nodes join the network. However it should not be able to retrieve the previous session key or some crucial information. In 2011 Hagraas et al. proposed a key management scheme for heterogeneous wireless sensor networks, which satisfies confidentiality, authentication, integrity and unforgetability but lacks forward secrecy. In this paper, the shortcomings of the victim scheme has been extricated and repaired with the help of Elliptic Curve Discrete logarithm problem (ECDLP). An elliptic curve based signcryption key management scheme has been proposed which includes forward secrecy.

Keywords Signcryption · Key management · Forward secrecy · Wireless sensor networks

1 Introduction

Sensor networks have proved its existence to verify a wide range of applications such as home automation, monitoring of critical infrastructures, environmental monitoring, forest fire detection, data acquisition in hazardous environments, and

S. Bala (✉) · G. Sharma · A. K. Verma
Computer Science and Engineering Department, Thapar University, Patiala, India
e-mail: suman.bala@thapar.edu

G. Sharma
e-mail: gaurav.sharma@thapar.edu

A. K. Verma
e-mail: akverma@thapar.edu

military operations and many more. The basic security primitives for key management schemes are confidentiality, authenticity, integrity and non-repudiation. Forward secrecy and public verifiability are two more security aspect needs to be addressed. Numerous schemes [2, 4–7] are proposed over the years to provide different level of security measures and communication/computational costs. The main objective in security is to optimize the cost of communication and computation. Elliptic Curve cryptography [16] has been widely used to attain a desired security level with smaller key size in contrast to conventional security approaches. It leads to a better utilization of memory, energy and bandwidth for the resource-constrained devices such as wireless sensor networks. Signcryption [1] can lessen the cost of communication and computation to a great extent, which can process the signature and encryption together.

Signcryption is a cryptographic process proposed by Zheng [1] to join the functionalities of encryption along with digital signature in a single logical step. The author further finds out that the signcryption costs 58 % less in average computation time and 70 % less in message expansion than does signature-then-encryption based on the discrete logarithm problem [1]. Later, the author proposed two key exchange protocols [2] using signcryption, which are based on discrete logarithm problem (DLP) called Direct Key Exchange Using a Nonce (DKEUN) and Direct Key Exchange Using Time-stamp (DKEUTS). But, the scheme fails the forward secrecy of message confidentiality when the sender's private key disclosed [7].

Moreover, Zheng and Imai [3] proposed a signcryption scheme based on elliptic curve discrete logarithm problem (ECDLP), which saves 58 % in computational cost and 40 % in communication overhead as compared with signature-then-encryption on elliptic curves but it lacks forward secrecy, public verifiability and encrypted message authentication. In the previous discussed schemes [2, 3], there is one more problem that is these schemes can't be used in such applications where third party validation is necessary using a public key as done in signature schemes. The solution is provided by Zheng [4], which introduces an independent judge. But when dispute occurs the judge can't verify the signature, as he is not having the private key of the recipient. To overcome the above problem Bao and Deng [5] enhanced Zheng's scheme [4] in such a way that verification of a signature does not need the recipient's private key but the scheme was not as efficient as Zheng's scheme. Gamage et al. [6] also modifies Zheng's [1] signcryption scheme in such a way that anyone can verify the signature of ciphertext to protect confidentiality of message in firewall application.

Jung et al. [7] proposed a signcryption scheme based on discrete logarithm problem (DLP) with forward secrecy. Later, Hwang et al. [8] proposed a signcryption scheme based on elliptic curve discrete logarithm problem (ECDLP), which provides forward secrecy for message confidentiality and public verification along with other basic security notions. When dispute occurs, the judge can verify sender's signature without the sender's private key. Kim and Youm [9] proposed two protocols named Secure Authenticated Key Exchange (SAKE) protocol and Elliptic Curve-Secure Authenticated Key Exchange (EC-SAKE) protocol. The protocols are

efficient in terms of computation complexity and communication performance as compared to DKEUN, DKEUTS and EC-DKEUN, EC-DKEUTS respectively. Zhou [10] proposed a scheme based on ECDLP with public verifiability through a trusted third party without disclosing private key. Toorani and Beheshti [11] and Elsayed and Hassan [12] proposed the schemes based on ECDLP, which provides forward secrecy for message confidentiality and public verification. Hamed and Khamy [13] proposed a scheme based on ECDLP for cluster based wireless sensor networks. Whereas, Hagrais et al. [14] proposed a scheme based on ECDLP for heterogeneous wireless sensor networks. Later, Hagrais et al. [15] proposed a scheme which is efficient [13] in terms of total number of operations, key storage, energy consumption and communication overhead as 75 %, 96 %, 23.79 mJ and 40 % respectively but lacks to provide forward secrecy. The scheme proposed by Hagrais et al. [15] satisfies all the security requirements except forward secrecy. In this paper, an improved elliptic curve based key management signcryption scheme has been proposed which provides forward secrecy along with all security requirements. In addition to confidentiality, unforgeability, integrity and non-repudiation, the proposed scheme has been proved to be more secure.

2 Problem Identification and Solution

This section covers the details regarding identification of the problem, proposed solution and parameters used for elliptic curve signcryption.

2.1 Identification of the Problem

The scheme proposed by Hagrais et al. [15] satisfies all the security requirements except forward secrecy. The condition for forward secrecy is: Even if the long-term private key of the sender is revealed, the adversary is not capable of decrypting the previously signcryptured texts.

2.2 Proposed Solution

The proposed scheme satisfies forward secrecy along with the basic security requirements. The forward secrecy of the proposed scheme will be compromised only if the attacker can solve the ECDLP that is computationally infeasible with the selected domain parameters. The proposed scheme has secure key exchange, less storage requirement, scalability and low complexity.

Table 1 Parameters public to all

P :	a large prime
E :	an elliptic curve over $GF(p^m)$, with $p \geq 2^{150}$ and $m = 1$, or $p = 2$, and $m > 150$
q :	a large prime factor of $ p^m $
G :	a point with order q , chosen randomly from the points on E
$hash$:	a one-way hash function whose output has at least 128 bits
KH :	a keyed one-way hash function
$(E; D)$:	the encryption and decryption algorithms of a private key cipher

2.3 Elliptic Curve Signcryption Parameters

In this section, we discuss various parameters and their notations, which are used throughout the paper in Table 1.

3 Proposed Scheme

This section covers the proposed scheme in detail. The proposed scheme works in three phases in the following manner.

3.1 Phase-I: Generation of Public/Private Key

This phase is responsible for creating public/private key pair for Base-Node (B), Cluster-Heads (H) and Cluster-Nodes (N). It creates the BH symmetric keys, which is used for secure communication between the cluster-heads among each other and with the base-node. Also, it creates the HN symmetric keys, which is used for secure communication between the cluster-nodes among each other in the cluster and with the corresponding cluster-head as shown in Fig. 1a.

A.1: *Base-Node generates public/private key pair.*

P_B Base-node (B) choose its private-key uniformly at a random from $[1, \dots, q - 1]$

Q_B Base-node (B) computes the public-key, $Q_B = P_B G$

A.2: *Cluster-Head generates public/private key pair.*

P_{H_i} Each cluster-head (H_i) choose its private-key uniformly at a random from $[1, \dots, q - 1]$, where $i \in 1, \dots, n_1$; n_1 : the number of cluster-heads

Q_{H_i} Each cluster-head (H_i) computes its public-key, $Q_{H_i} = P_{H_i} G$

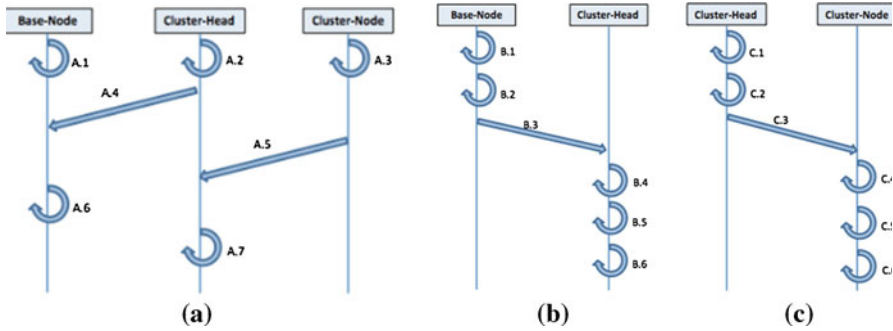


Fig. 1 a Generation of public/private keys (Phase-I), b Key Establishment of base-node cluster-head (Phase-II), c Key establishment of cluster-head cluster-node (Phase-III)

A.3: Cluster-Node generates public/private key pair.

P_{N_j} Each cluster-node (N_j) choose its private-key uniformly at a random from $[1, \dots, q - 1]$, where $j \in 1, \dots, n_2; n_2$: the number of cluster-nodes

Q_{N_j} Each cluster-node (N_j) computes its public-key, $Q_{N_j} = P_{N_j}G$

A.4: Cluster-Head sends its public key to Base-Node.

All cluster-heads (H_i) send their public-key Q_{H_i} to the Base-node.

A.5: Cluster-Node sends its public key to Cluster-Head.

All cluster-nodes (N_j) send their public-key Q_{N_j} to corresponding cluster-head (H_i).

A.6: Base-Node creates BH symmetric key.

Base-node (B) creates the symmetric key (K_{BH}), which is used for secure communication between the base-node and the cluster-heads, and among the cluster-heads.

A.7: Cluster-Heads creates HN symmetric key

Cluster-heads (H_i) create the symmetric key (K_{HN_i}), which is used for secure communication between the cluster-head and their corresponding cluster-nodes, and among the cluster-nodes within the cluster-head.

3.2 Phase-II: Base-Node Cluster-Head Key Establishment

This phase is responsible for the base-node cluster-head key establishment as shown in Fig. 1(b). The base-node generates the shared symmetric key for each cluster-head by using their public-keys; signcrypts the symmetric key (K_{BH_i}) generated in the first phase and send to the cluster-heads, which later unencrypts by the cluster-head as follows:

- B.1 *Base-Node generate a shared symmetric key for each cluster-head by using their public-key*
- B.2 *Base-Node encrypt and signature the BH symmetric key using shared symmetric key*
- B.3 *Base-Node sends the encrypted BH symmetric key and its encrypted signature*
- B.4 *Cluster-Head generates a shared symmetric key using private key of Cluster-Head and received signature*
- B.5 *Cluster-Head decrypts the BH symmetric key and its signature using shared symmetric key*
- B.6 *Cluster-Head verifies BH symmetric key signature*

3.3 Phase-III: Cluster-Head Cluster-Node Key Establishment

This phase is responsible for the cluster-head cluster-node key establishment as shown in Fig. 1(c). Each cluster-head generates the shared symmetric key for each cluster-node in the corresponding cluster by using their public-keys; signcrypts the symmetric key ($K_{HN_{i,j}}$) generated in the first phase and send to the cluster-nodes, which later unencrypts by the cluster-node as follows:

- C.1 *Cluster-Head generates a shared symmetric key using public key of Cluster-Node*
- C.2 *Cluster-Head encrypts and signature the HN symmetric key using shared symmetric key*
- C.3 *Cluster-Head sends the encrypted HN symmetric key and its encrypted signature*
- C.4 *Cluster-Node generates a shared symmetric key using private key of cluster-node and received signature*
- C.5 *Cluster-Node decrypts the HN symmetric key and its signature using shared symmetric key*
- C.6 *Cluster-Node verify HN symmetric key signature*

Algorithm 1: BH Symmetric Key Signcryption/Unsigncryption

Signcryption: The base-node signcrypts the symmetric key K_{BH_i} using its private key and sends the ciphertext (C_i, T_i, S_i) to each cluster-head

1. The base-node chooses $r_i \in_R \{1, \dots, q-1\}$
2. $(k_{i,1}, k_{i,2}) = \text{hash}(r_i Q_{H_i})$
3. $C_i = E_{k_{i,1}}(K_{BH_i})$
4. $R_i = KH_{k_{i,2}}(K_{BH_i} \parallel Q_{H_i} \parallel ID_A \parallel ID_B)$
5. $S_i = (r_i / (R_i + P_B))$
6. $T_i = R_i G$

Unsigncryption: The base-node sends the cipher text (C_i, T_i, S_i) to each cluster-head and each cluster-head unsigncrypts the symmetric key

7. $U_i = S_i P_{H_i}$
 8. $(k_{i,1}, k_{i,2}) = \text{hash}(U_i(T_i + Q_B))$
 9. $K_{BH_i} = D_{k_{i,1}}(C_i)$
 10. Accept K_{BH_i} if and only if $T_i = U_i G$
-

Algorithm 2: HN Symmetric Key Signcryption/Unsigncryption

Signcryption: The cluster-head signcrypts the symmetric key $(K_{HN_{i,j}})$ using its private key and sends the ciphertext (C_j, T_j, S_j) to all the cluster-nodes in the corresponding cluster

1. The cluster-head (H_i) chooses $r_i \in_R \{1, \dots, q-1\}$
2. $(k_{j,1}, k_{j,2}) = \text{hash}(r_j Q_{N_j})$
3. $C_j = E_{k_{j,1}}(K_{HN_{i,j}})$
4. $R_j = KH_{k_{j,2}}(K_{HN_{i,j}} \parallel Q_{N_j} \parallel ID_A \parallel ID_B)$
5. $S_j = (r_j / (R_j + P_{H_i}))$
6. $T_j = R_j G$

Unsigncryption: The cluster-head sends the cipher text (C_j, T_j, S_j) to each cluster-node and each cluster-node unsigncrypts the symmetric key

7. $U_j = S_j P_{N_j}$
 8. $(k_{j,1}, k_{j,2}) = \text{hash}(U_j(T_j + Q_{H_i}))$
 9. $K_{HN_{i,j}} = D_{k_{j,1}}(C_j)$
 10. Accept $K_{HN_{i,j}}$ if and only if $T_j = U_j G$
-

4 Security Analysis

The concept of forward secrecy proves its importance in wireless sensor networks. If a sensor node runs out of energy and gets replaced with a new node, the new node should not be able to unsigncrypt the previous signcrypted messages. In this paper the flaw of the existing scheme has been bring into notice and repaired. The proposed key management using public key elliptic curves signcryption for WSN provides all security functions: key confidentiality, authentication, integrity and unforgetability but lacks in forward secrecy. The security proof of all the required parameters can be directly taken from the parent scheme. The security of improved scheme is based upon the elliptic curve discrete logarithm problem (ECDLP),

which is computationally infeasible to solve for the specified parameters. The new scheme does not change the storage used for sensor nodes.

5 Conclusion

In the case of wireless sensor networks, forward secrecy is a vital security requirement. In this paper, an elliptic curve based key management scheme has been improved in terms of forward secrecy. The proposed scheme satisfies all the basic security requirements of key management schemes. The security of the proposed scheme can be proved with the help of victim scheme. The forward secrecy of the proposed scheme will be compromised only if the attacker can solve the ECDLP, which is computationally infeasible with the selected domain parameters.

References

1. Zheng Y (1997) Digital signcryption or how to achieve Cost (Signature & Encryption) cost (Signature)+cost (Encryption). In: *Advances in cryptology—Crypto97 LNCS 1294*, Springer, pp 165–179
2. Zheng Y (1998) Shortened digital signature, signcryption and compact and unforgeable key agreement schemes. In: *IEEE P1363a: standard specifications for public-key cryptography: additional technique*
3. Zheng Y, Imai H (1998) How to construct efficient signcryption schemes on elliptic curves. *Inf Process Lett* 68:227–233
4. Zheng Y (1998) Signcryption and its application in efficient public key solutions. In: *Proceeding of ISW '97, LNCS vol 1396*, Springer, pp 291–312
5. Bao F, Deng RH (1998) A signcryption scheme with signature directly verifiable by public key. In: *Proceedings of PKC08LNCS 1431*, Springer pp 55–59
6. Gamage C, Leiwo J, Zheng Y (1999) Encrypted message authentication by firewalls. In: *Proceedings of 1999 international workshop on practice and theory in public key cryptography (PKC09)*, 1–3 March 1999, Kamakura, JapanLNCS 1560, Springer, pp 69–81
7. Jung HY, Chang KS, Lee DH, Lim JI (2001) Signcryption schemes with forward secrecy. In: *Proceeding of WISA 2*, pp 403–475
8. Hwang RJ, Lai CH, Su FF (2005) An efficient signcryption scheme with forward secrecy based on elliptic curve. *J Appl Math Comput (Elsevier Inc.)*, 167 (2):870–881. DOI: [10.1016/j.amc.2004.06.124](https://doi.org/10.1016/j.amc.2004.06.124)
9. Kim RH, Youm HY (2006) Secure authenticated key exchange protocol based on EC using signcryption scheme. In: *Proceeding international conference on hybrid information technology (IEEE Computer society)*, 8 June 2006
10. Zhou X (2009) Improved signcryption scheme with public verifiability. In: *Proceeding Pacific-Asia conference on knowledge engineering and software engineering (IEEE Computer Society)*, 4 Sept 2009
11. Toorani M, Beheshti AA (2009) An elliptic curve-based signcryption scheme with forward secrecy. *J Appl Sci* 9(6):1025–1035
12. Elsayed M, Hasan E (2009) Elliptic curve signcryption with encrypted message authentication and forward secrecy. *Int J Comput Sci Netw Sec* 9(1)

13. Said EK, Amr IH (2009) New low complexity key exchange and encryption protocols for wireless sensor networks clusters based on elliptic curve cryptography. In: Proceedings of the 2009 national conference on radio science. Cairo, Egypt
14. Hagra EA, Aly HH, Saied DI (2010) An efficient key management scheme based on elliptic curve signcryption for heterogeneous wireless sensor networks. UCST 1(2):459–474
15. Hagra EA, Aly HH, Saied DI (2011) Energy efficient key management scheme based on elliptic curve signcryption for wireless sensor networks. In: 28th NRSC'11 April 26–28, 2011, National Telecommunication Institute, Egypt
16. Hankerson D, Menezes AJ, Vanstone S (2004) Guide to elliptic curve cryptography. Springer, New York

An Identity-Based Ring Signcryption Scheme

Gaurav Sharma, Suman Bala and Anil K. Verma

Abstract Signcryption enables a user to perform digital signature for providing authenticity and public key encryption for providing message confidentiality simultaneously in a single logical step with a cost lesser than sign-then-encrypt approach. As the concept of ring signcryption emerged, various practical applications like electronic transaction protocol and key management protocols, felt the requirement of signer's privacy, which was lacking in normal signcryption schemes. Without revealing the users' identity of the ring signcryption can provide confidentiality and authenticity both. In this paper, we present a new ID-based ring signcryption scheme, motivated to the scheme provided by Zhu et al. [1]. Selvi et al. [2] and Wang et al. [3] found some security flaws in the Zhu's scheme [1], which is being considered and repaired in this paper. The proposed scheme is proven to be secure against adaptive chosen ciphertext ring attacks (IND-IDRSC-CCA2) and secure against an existential forgery for adaptive chosen message attacks (EF-IDRSC-ACMA).

Keywords Identity-based ring signcryption · Identity based cryptography · Ring signcryption · Confidentiality · Anonymity · Unforgeability · Bilinear pairing

G. Sharma (✉) · S. Bala · A. K. Verma
Computer Science and Engineering Department, Thapar University, Patiala, India
e-mail: gaurav.sharma@thapar.edu

S. Bala
e-mail: suman.bala@thapar.edu

A. K. Verma
e-mail: akverma@thapar.edu

1 Introduction

The idea behind Identity-based Ring Signcryption is a collaboration of different security techniques, such as Identity Based Cryptography, Ring Signature and Signcryption. Identity based cryptography provides a variant to Certificate based public key cryptography; ring signature provides anonymity along with the authenticity in such a way that even verifier does not know who has signed the message but he can verify that one of the ring member has signed it, while signcryption provides the encryption and signature in a single logical step to obtain confidentiality, integrity, authentication and non-repudiation. The concept of identity-based cryptography was introduced by Shamir [4] in 1984, to remove the need of certification of the public keys, which is required in the conventional public key cryptography setting. But, Shamir only proposed ID-based signature and left the ID-based encryption as an open problem. Boneh and Franklin [5] presented the first Identity Based Encryption scheme that uses bilinear maps (the Weil or Tate pairing) over super singular elliptic curves. Rivest et al. [6] introduced ring signature which is a group oriented signature with privacy concerns: a user can anonymously sign a message on behalf of a group of spontaneously conscripted users, without managers including the actual signer. The first ID-based ring signature scheme with bilinear pairings, was proposed by Zhang and Kim [7]. Yuliang Zheng [8] introduced the concept of public key signcryption which fulfils both the functions of digital signature and public key encryption in a logically single step, and with a cost lower than that required by the sign-then-encrypt approach. However, Zheng didn't prove any security notions which was further proposed by Baek et al. [9], described a formal security model in a multi-user setting.

Xinyi Huang [10] combined the concepts of ID-based ring signature and signcryption together as identity-based ring signcryption. They provided a formal proof of their scheme with the chosen ciphertext security (IND-IDRSC-CCA) under the Decisional Bilinear Diffie-Hellman assumption. However, Huang et al.'s [10] scheme is computationally inefficient, since the number of pairing computations grows linearly with the group size. Huang et al.'s scheme needs $n + 4$ pairing computations, where n denotes the size of the group. The scheme lacks anonymity and had a key escrow problem as the scheme was based on ID-PKC. Wang et al. [11] eliminated the key escrow problem in [10] by proposing a verifiable certificate-less ring signcryption scheme and gave a formal security proof of the scheme in random oracle model. But this scheme also needs $n + 4$ pairing computations. The problem of ID-based ring signcryption schemes is that they are derived from bilinear pairings, and the number of pairing computations grows linearly with the group size. Zhu [1] solved the above problem; they proposed an efficient ID-based ring signcryption scheme, which only takes four pairing operations for any group size. Zhu [12] proposed an ID-based ring signcryption scheme, which offers savings in the ciphertext length and the computational cost.

The other schemes include Li et al. [13], Yong et al. [14] and Zhang [15]. Selvi et al. [2] proved that Li et al. [16] and Zhu et al. scheme [1] are not secure against

adaptive chosen ciphertext attack while Zhu’s [12] scheme and Yong’s [14] scheme are not secure against chosen plaintext attack. Qi’s [17] proved that their scheme has the shortest ciphertext and is much more efficient than Huang’s [10] and Selvi’s [2] scheme. Selvi et al. [18] proved that Zhang et al. [19] scheme is insecure against confidentiality, existential unforgeability and anonymity attacks. Zhou [20] presented an efficient identity-based ring signcryption scheme in the standard model.

Roadmap: The remaining paper is organized as follows: Sect. 2 gives some preliminaries and basic definitions of Bilinear Pairing. The formal model has been discussed in Sect. 3. In Sect. 4, we propose our ID-based ring signcryption scheme; security analysis of the proposed scheme is discussed in Sect. 5. In Sect. 6, we concluded the remarks about the paper.

2 Preliminaries

2.1 Notations Used

The following notations have been made in common for all the existing schemes and Table 1 defines the description of the notations that have been used throughout the paper.

2.2 Basic Concepts on Bilinear Pairing

Let G_1 be a cyclic additive group generated by P of prime order q , and G_2 be a cyclic multiplicative group of the same order q . Let a and b be elements of Z_q^* . Assume that the discrete logarithm problem (DLP) in both G_1 and G_2 is hard. Let $\hat{e} : G_1 \times G_1 \rightarrow G_2$ be a bilinear pairing with the following properties shown in Table 2.

Table 1 Notations used

k : security parameter	$\{0, 1\}^l$: string with length l .
$params$: systems’ public parameter generated by PKG	$\{0, 1\}^*$: string with arbitrary length.
t : secret key generated by PKG	$m \in_R M$: message, M : message space
G_1 : cyclic additive group generated by P of prime order $q > 2^k$	$\hat{e} : G_1 \times G_1 \rightarrow G_2$ is a bilinear pairing
G_2 : cyclic multiplicative group generated by P of prime order $q > 2^k$	ID_i : user identity
$P \in G_1$: random generator	S_i : private key of user i
P_{pub} public key of PKG	Q_i : public key of user i
Z_q^* : multiplicative group modulo q .	S : sender, R : receiver
	\mathcal{L} : group of ring members
	\mathcal{C} : signcrypted ciphertext
	\mathcal{A} : Adversary, \mathcal{C} : Challenger

Table 2 Properties of bilinear mapping

Bilinearity	$\forall P, Q, R \in_R G_1, \hat{e}(P + Q, R) = \hat{e}(P, R)\hat{e}(Q, R),$ $\hat{e}(P, Q + R) = \hat{e}(P, Q)\hat{e}(P, R)$. In particular, for any $a, b \in Z_q^*$ $\hat{e}(aP, bP) = \hat{e}(P, P)^{ab} = \hat{e}(P, abP) = \hat{e}(abP, P)$
Non-degeneracy	$\exists P, Q, \in G_1 \ni \hat{e}(P, Q) \neq I_{G_2}$, where I_{G_2} is the identity of G_2
Computability	$\forall P, Q \in G_1$, there is an efficient algorithm to compute $\hat{e}(P, Q)$.

3 Formal Model of Identity Based Ring Signcryption

A generic ID-based ring signcryption scheme consists of five algorithms Setup, Keygen, Signcrypt, Unsigncrypt and Consistency. The description of these algorithms has been provided in Table 3.

4 Proposed Scheme

In this section, we propose our new Identity-Based Ring signcryption Scheme. Our scheme has four following algorithms:

1. *Setup* (k): Given a security parameter k , a trusted private key generator (PKG) generates the system's public parameters $params$ and the corresponding master secret key t that is kept secret by PKG.
 - a. The trusted authority randomly chooses $t \in_R Z_q^*$ keeps it as a master key and computes the corresponding public key $P_{pub} = tP$.
 - b. Let $(G_1, +)$ and $(G_2, *)$ be two cyclic groups of prime order $q > 2^k$ and a random generator $P \in G_1$.
 - c. $e : G_1 \times G_1 \rightarrow G_2$ is a bilinear pairing.
 - d. Choose Hash Functions

$$H_1 : \{0, 1\}^* \rightarrow G_1, H_2 : G_2 \rightarrow \{0, 1\}^l, H_3 : \{0, 1\}^* \rightarrow Z_q^*, H_4 : \{0, 1\}^* \rightarrow \{0, 1\}^l$$

Table 3 Generic identity based ring signcryption scheme

Setup	For a given parameter k , a trusted private key generator generates system's public parameters $params$ and its corresponding master secret key t , which is kept secret.
Keygen	For a given user identity ID_i , PKG computes private key S_i by using $params$ and t and transmits S_i to ID_i via secure channel.
Signcrypt	For sending a message m from sender to a receiver with identity ID_R , sender's private key S_S , and a group of ring members $\{U_i\}_{i=1 to n}$ with identities $\mathcal{L} = \{ID_1, \dots, ID_n\}$, sender computes a ciphertext.
Unsigncrypt	For retrieving a message m , if \mathbb{C} is a valid ring signcryption of m from the ring \mathcal{L} to ID_R or 'invalid', if \mathbb{C} is an invalid ring signcryption.
Consistency	An identity based ring signcryption scheme is said to be consistent if $\Pr[\mathbb{C} \leftarrow \text{signcrypt}(m, \mathcal{L}, S_S, ID_R), m \leftarrow \text{unsigncrypt}(\mathbb{C}, \mathcal{L}, S_R)] = 1$

- e. The public parameters are: $params = \{G_1, G_2, e, q, P, P_{pub}, H_1, H_2, H_3, H_4\}$.
2. *Keygen* (ID_i): Given a user identity ID_i of user U_i , the PKG, using the public key computes the parameters $params$ and the master secret key t , computes the corresponding private key S_i , and transmits it to ID_i in a secure way as follows.
 - a. The public key is computed as $Q_i = H_1(ID_i)$.
 - b. The corresponding private key $S_i = tQ_i$.
 - c. PKG sends S_i to user U_i via a secure channel.
 3. *Signcrypt*: Let $\mathcal{L} = \{ID_1, \dots, ID_n\}$ be a set of n ring members, such that $ID_S \in \mathcal{L}$. ID_R may or may not be in \mathcal{L} . The sender runs this algorithm to send a message $m \in M$, where M is a message space, to a receiver with identity ID_R , the senders private key S_S , outputs a ring signcryption \mathbb{C} as follows:
 - a. Choose a random number $r \in_R Z_q^*$ and $m^* \in_R M$. And compute $R_0 = rP$, $R = e(rP_{pub}, Q_R)$, $k = H_2(R)$, $\mathbb{C}_1 = m^* \oplus k$
 - b. Choose $R_i \in G_1 \forall i = \{1, 2, \dots, n\} \setminus \{S\}$ and compute $h_i = H_3(m || \mathcal{L} || R_i || R_0)$.
 - c. Choose $r_S \in_R Z_q^* \forall i = S$ Compute $R_S = r_S Q_S - \sum_{i \neq S} (R_i + h_i Q_i)$, $h_S = H_3(m || \mathcal{L} || R_S || R_0)$, $V = (h_S + r_S)S_S$, $\mathbb{C}_2 = (m || r_S || V) \oplus H_4(m^* || R_0)$.
 - d. Finally the sender outputs the ciphertext as $\sigma = (\mathcal{L}, R_0, R_1, \dots, R_n, \mathbb{C}_1, \mathbb{C}_2)$ to the receiver.
 4. *Unsigncrypt*: This algorithm is executed by a receiver ID_R . This algorithm takes the ring signcryption σ , the ring members \mathcal{L} and the private key S_R , as input and produces the plaintext m , if σ is a valid ring signcryption of m from the ring \mathcal{L} to ID_R or 'invalid', if σ is an invalid ring signcryption as follows:
 - a. Compute $R' = e(R_0, S_R)$, $k' = H_2(R')$, $m'^* = \mathbb{C}_1 \oplus k'$
 - b. Recover m' , V' as $(m' || r_S || V') = \mathbb{C}_2 \oplus H_4(m'^* || R_0)$.
 - c. Compute $h'_i = H_3(m' || \mathcal{L} || R_i || R') \forall i = \{1, 2, \dots, n\}$
 - d. Checks if $e(P, V') \stackrel{?}{=} e\left(P_{pub}, \sum_{i=1}^n (R_i + h_i Q_i)\right)$. If the check succeeds accept m , else return \perp .

5 Security Analyses of the Proposed Scheme

5.1 Correctness

In this section, a proof of correctness has been shown, that if the ciphertext \mathbb{C} has been correctly generated, the verification equations will hold.

$$\text{If } e(P, V') \stackrel{?}{=} e\left(P_{pub}, \sum_{i=1}^n (R_i + h_i Q_i)\right) \text{ holds.}$$

$$\text{Proof: } e(P, V) = e(P, (h_S + r_S)S_S) = e(P, (h_S + r_S)tQ_S) = e(tP, h_S Q_S + R_S + \sum_{i=1, i \neq s}^n (R_i + h_i Q_i)) = e\left(P_{pub}, \sum_{i=1}^n (R_i + h_i Q_i)\right)$$

5.2 Security Analyses

5.2.1 Confidentiality

Theorem: If an IND-IRSC-CCA2 adversary \mathcal{A} has an advantage ε against IRSC scheme, asking hash queries to random oracles \mathcal{O}_{H_i} ($i = 1, 2, 3, 4$), q_e extract queries ($q_e = q_{e_1} + q_{e_2}$, where q_{e_1} and q_{e_2} are the number of extract queries in the first phase and second phase respectively), q_{sc} signcryption queries and q_{us} un-signcryption queries, then there exist an algorithm \mathcal{C} that solves the CBDH problem with advantage $\varepsilon\left(\frac{1}{q_{H_1} q_{H_2}}\right)$.

5.2.2 Unforgeability

Theorem: An identity based ring signcryption scheme (IRSC) is said to be existentially unforgeable against adaptive chosen message attack (EUF-IRSC-CMA), against any polynomially bounded adversary \mathcal{A} under the random oracle model if CDHP is hard.

6 Conclusion

Wang et al. [25] proved that the Zhu et al. scheme [1] to be insecure against anonymity and also does not satisfy the property of unforgeability. Selvi et al [2] also attacked and proved the scheme prone to confidentiality attack. Till now, a very few ID-based ring signcryption schemes have been proposed and most of them have been proved insecure. In this paper an efficient ID based ring signcryption scheme has been presented which has been proven secure against the primitive properties of signcryption: confidentiality, unforgeability and anonymity. The future work may include ring signcryption schemes in combination with ID-based threshold signcryption, ID-based proxy signcryption and id based hybrid signcryption schemes and certificate-less schemes in the standard model. Also, to reduce communication overhead, constant ciphertext size ring signcryption schemes can be improved.

References

1. Zhu Z, Zhang Y, Wang F (2008) An efficient and provable secure identity based ring signcryption scheme. *Computer standards & interfaces*, pp 649–654
2. Selvi SSD, Vivek SS, Rangan CP (2009) On the security of identity based ring signcryption schemes. In: *Proceedings of 12th International Conference on ISC 2009, Pisa, Italy, Sept 7–9, 2009*, Proceedings of LNCS 5735, Springer, Berlin, pp 310–325
3. Wang H, Yu H (2008) Cryptanalysis of two ring signcryption schemes. In: *Inscrypt 2008*, LNCS-5487, Springer, Berlin, pp 41–46
4. Shamir A (1984) Identity-based cryptosystems and signature schemes. In: *Proceedings of CRYPTO '84*, LNCS 196, Springer, Berlin, pp 47–53
5. Boneh D, Franklin M (2001) Identity-based encryption from the weil pairing. In: *Proceedings of CRYPTO '01*, LNCS 2139, Springer, Berlin, pp 213–229
6. Rivest RL, Shamir A, Tauman Y (2001) How to leak a secret. In: *Proceedings of advances in cryptology in asiacrypt 2001*, LNCS 2248, Springer, Berlin, pp 552–565
7. Zheng F, Kim K (2002) Id-based blind signature and ring signature from pairings. In: *Proceedings of Asiacrypt 02*, LNCS 2501, Springer, Berlin, pp 533–547
8. Zheng Y (1997) Digital signcryption or how to achieve cost (signature and encryption) cost (signature) + cost(encryption)'. In: *Proceedings of CRYPTO-97*, pp 165–179
9. Baek J, Steinfeld R, Zheng Y (2002) Formal proofs for the security of signcryption. In: *Proceedings of PKC—02*, LNCS 2274, pp 81–98
10. Huang X, Susilo W, Mu Y, Zhang F (2005) Identity-based ring signcryption schemes: cryptographic primitives for preserving privacy and authenticity in the ubiquitous world. In: *Proceedings of AINA 05*, Taipei, Taiwan, pp 649–654
11. Wang L, Zhang G, Ma C (2007) A secure ring signcryption scheme for private and anonymous communication. In: *Proceedings of international conference NPC workshops, 2007*
12. Zhu L, Zhang F (2008) Efficient identity based ring signature and ring signcryption schemes. In: *Proceedings of international conference on CIS'08, vol 2*, pp 303–307
13. Li F, Xiong H, Yu Y (2008) An efficient ID-based ring signcryption scheme. In: *Proceedings of ICCAS 2008*, Xiamen, pp 542–546
14. Yu Y, Li F, Xu C, Sun Y (2008) An efficient identity-based anonymous signcryption scheme. *Wuhan Univ J Nat Sci* 13(6):670–674
15. Zhang J, Gao S, Chen H, Geng Q (2009) A novel id-based anonymous signcryption scheme. In: *Proceedings of APWeb/WAIM*, LNCS 5446, Springer, Berlin, pp 604–610
16. Li F, Shirase M, Takagi T (2008) Analysis and improvement of authenticatable ring signcryption scheme. *J Shanghai Jiaotong Univ (Sci)* 13(6):679–683
17. Qi ZH, Yang G, Ren XY, Li YW (2010) An ID-based ring signcryption scheme for wireless sensor networks. In: *Proceedings of IET International of Conference WSN, China*, pp 368–373
18. Selvi SSD, Vivek SS, Rangan CP (2010) Identity based ring signcryption with public verifiability. In: *Proceedings of SECRYPT—10*, LNCS 2010
19. Zhang M, Zhong Y, Yang B, Zhang W (2009) Analysis and improvement of an id-based anonymous signcryption model. In: *Proceedings of ICIC (1)*, LNCS 5754
20. Zhou J (2011) An efficient identity-based ring signcryption scheme without random oracles. In: *Proceedings of international conference on computer and electrical engineering 4th (ICCEE—11)*, 2011
21. Huang XY, Zhang FT, Wu W (2006) Identity-based ring signcryption scheme. *Proc Tien Tzu Hsueh Pao/Acta Electronica Sinica* 34(2):263–266
22. Malone-Lee J (2002) Identity based signcryption. *J Cryptol* 2002/098
23. Chow SSM, Yiu SM, Hui LCK (2005) Efficient identity based ring signature. In: *Proceedings of ACNS 2005*, LNCS 3531, Springer, Berlin, pp 499–512

A Secure DS-CDMA Technique with Capacity Enhancement for Ad Hoc Wireless Networks

Muhammad Zeeshan, Shoab Ahmed Khan
and Muhammad Yasir Malik

Abstract In wireless ad hoc networks, the Direct Sequence Code Division Multiple Access (DS-CDMA) is the most promising candidate for wideband data access. The reasons are the high security, high throughput and soft limit on the number of active mobile devices. Soft limit on the network capacity means degradation of the performance by increasing the number of mobile devices. In many ad hoc networks, it is required to maintain the Bit Error Rate (BER) within some tolerable limits for a given SNR. This results in restriction on the number of mobile devices in the network. The number of active mobile devices or network capacity is further reduced by the mutual interference among the users caused due to multipath fading effects and synchronization problems. This paper presents an effective method of capacity enhancement of CDMA based wireless ad hoc networks by using a multiuser detection based power management framework. The system is simulated in the multipath channel models for fixed wireless applications. The proposed scheme increases the capacity two times as compared to the conventional CDMA based networks. Simulation results have been presented to demonstrate the effectiveness of the proposed scheme.

Keywords Multiuser CDMA · BER · Power management · Capacity · SUI channel

M. Zeeshan (✉) · S. A. Khan
College of Electrical and Mechanical Engineering, National University
of Sciences and Technology, Rawalpindi, Pakistan
e-mail: ranamz@live.com

S. A. Khan
e-mail: kshoab@yahoo.com

M. Y. Malik
Samsung Electronics, Suwon, Korea
e-mail: yasir_alf@yahoo.com

1 Introduction

In many capacity limited wireless communication systems, Code Division Multiple Access (CDMA) is one of the best channel access methods [1]. The reason behind is that it offers a soft limit on the capacity. There is a concept of Universal Frequency Reuse (UFR) in every CDMA system which means that all the users simultaneously share the entire spectrum [1]. Each user is distinguished by the other user by a unique code assigned to it. The code assigned to each user is orthogonal to that of the others. But due to the Multipath Fading Environment (MFE) and some other unavoidable channel effects, there is always mutual interference between the users which is called Multiple Access Interference (MAI). Network capacity (or the number of active users) which is a very fundamental attribute and performance metric of wireless networks [2] is limited due to the simultaneous transmissions between the mobile devices and Multiple Access Interference (MAI).

CDMA places a soft upper bound on the capacity of the system. It means that the number of users for a particular system can be increased by compromising the BER performance of the system. But many CDMA based Mobile Ad hoc Networks (MANETs) demand a strict upper bound on the BER performance of the system which makes the capacity of the system hard as in TDMA and FDMA. This paper presents a scheme for improving the system capacity by using a distributed power management algorithm. This is achieved by using the joint technique of power management and multiuser detection in order to increase the number of users within permissible BER bound. As mentioned earlier that Multipath Fading Environment (MFE) also restricts the capacity of a wireless system, so the system with proposed technique is also simulated in Stanford University Interim (SUI) multipath fading channels.

Section 2 presents the system model of CDMA based MANETs and network capacity. The proposed capacity enhancement algorithm based on power management and multiuser detection is presented in Sect. 3. Simulation results are given in Sect. 4 followed by conclusion in Sect. 5.

2 System Model and Network Capacity

An ad hoc communication system belongs to the class of self-configurable networks. In this type of networks there is no pre-existing central infrastructure. These networks can be as simple as point-to-point networks or as complex as wireless grid based networks. Such networks can be of many types as classified in [2]. The network we have used in this paper is a high bandwidth, high data rate and distributed (infrastructureless) network.

Let the total number of nodes in an ad hoc network be K . Assume $K/2$ of them are communicating at one time instant. If we assume AWGN and perfect frequency offset synchronization, then the received data by any node will be,

$$r(t) = \sum_{k=1}^{K/2} A_k(t)c_k(t)b_k(t) + n(t) \tag{1}$$

where $A_k(t)$, $c_k(t)$ and $b_k(t)$ are the amplitude, spreading sequence and complex constellation point values of the k th node. The term $n(t)$ shows the AWGN with zero mean and σ^2 variance. If a receiving node wants to communicate with j th transmitting node, Eq. 1 can be rewritten as

$$r(t) = A_j b_j c_j(t) + \sum_{\substack{k=1 \\ k \neq j}}^{K/2} A_k b_k c_k(t) + n(t) \tag{2}$$

Equation 2 shows that the 2nd term on right side contains the composite data of all the transmitting nodes other than the desired transmitting node. If any receiving node of the network wants to receive the data from j th transmitting node, then the soft output of conventional correlative detector is obtained by correlating the received signal with spreading code of j th node. So by using Eqs. (1)–(2), we can write the received signal from the j th node as follows,

$$y_j = A_j b_j + \sum_{\substack{k=1 \\ k \neq j}}^{K/2} \rho_{k,j} A_k b_k + \frac{1}{T_b} \int_0^{T_b} n(t) c_j(t) dt \tag{3}$$

where $\rho_{k,j}$ represents the correlation between the spreading codes of k th and j th nodes. Equation 3 shows that the correlation of the data of j th node with its spreading code gives rise to recovered data term whereas correlation with the data of all other nodes gives rise to Multiple Access Interference (MAI).

If we define \mathbf{R} as the code correlation matrix of dimension $K/2 \times K/2$, \mathbf{A} is the $K/2 \times K/2$ diagonal matrix having amplitudes of each transmitted signal on the diagonal, and \mathbf{b} is $K/2 \times 1$ complex constellation point vector, then from Eq. 3, soft output of Conventional detector can be expressed in matrix form as

$$\mathbf{y}_{conventional} = \mathbf{RAB} + \mathbf{n} \tag{4}$$

Equation 4 shows that if non-desired nodes transmit with high power then MAI goes on increasing. Conventional CDMA detector does not incorporate this mutual interference in the detection process. Due to this reason, DS-CDMA suffers from Near-Far problem which either degrades the performance of network or reduces the overall capacity of the network. It can be shown mathematically as follows.

Let P_r be the received power of each node in this system. Signals from $[(K/2)-1]$ nodes will act as interference. Then Energy-to-Interference Ratio (EIR) of the desired node will be given by Eq. 5 as,

$$\frac{E_b}{I_0} = \frac{P_r T_b}{N_0 + [(K/2) - 1] P_r T_c} \quad (5)$$

Here, T_b is the bit duration and T_c is the chip duration, N_0 is the two-sided Power Spectral Density (PSD) of the Gaussian noise. If one receiving node is near to its corresponding transmitting node, then it's received power will be aP_r , where $a > 1$. EIR then becomes,

$$\frac{E_b}{I_0} = \frac{P_r T_b}{N_0 + aP_r T_c + [(K/2) - 2] P_r T_c} \quad (6)$$

So it is clear from Eq. 6 that EIR degrades drastically by increasing a , which will happen if any receiving node is very close to the corresponding transmitting node. If we want to maintain the same SIR, the factor $[(K/2) - 2]$ is to be decreased which means the loss of capacity. Hence we conclude that Near-far effect results in the decrease of EIR or alternatively in the decrease of system's capacity.

3 Proposed Algorithm

This section describes the proposed algorithm for the capacity enhancement including the power management at the transmitter side and the subsequent detection at the receiver side by the use of Minimum Mean Square Error (MMSE) detector. The proposed algorithm is a modification of an existing CDMA cellular power control algorithm.

We have selected the Distance Based Power Allocation (DBPA) algorithm given in [3], which calculates the allocated transmitted power by using the distance between the base station and mobile station. In DBPA algorithm of [3], the base station is responsible for allocating the power to each of its served mobile. Since ad hoc networks are an infrastructureless in which the nodes have to control all the communication protocols themselves. So, for ad hoc networks, this algorithm has been modified as follows.

In ad hoc networks, the nodes naturally need to exploit location information. Location awareness means that each node in the network has a positioning device to determine its current physical location [4]. This is achieved by location aware systems, also called Indoor Positioning Systems (IPS) [5]. Using this location awareness, each node maintains a physical location matrix in its memory. This location matrix contains the distance values between all the possible pairs of transmitting/receiving nodes of the network. The location matrix may be constant or variable depending upon whether the network is fixed or mobile. Using this location matrix, the allocated transmitting power of the k th device can be given by the proposed algorithm as

$$p_k = \begin{cases} \beta \left(\frac{d_{min}}{d_0}\right)^{\alpha'}, & \text{if } d_{m,k} \leq d_{min} \\ \beta \left(\frac{d_{m,k}}{d_0}\right)^{\alpha'}, & \text{if } d_{m,k} > d_{min} \end{cases} \quad (7)$$

Where β and α' are positive constants, d_0 is the distance at which the received power is known and d_{min} is the threshold distance to avoid very small transmitted powers and $d_{m,k}$ is the distance between transmitting node (k) and receiving node (m). There are two main differences between the cellular DBPA and the proposed DBPA algorithms;

- i. In cellular DBPA algorithm, the relative base station-to-mobile distances are measured by using measurement maps, whereas in the modified DBPA algorithm, the physical locations are found using Indoor Positioning System (IPS).
- ii. In cellular DBPA algorithm, the normalization factor for power calculation is the cell radius R , whereas in the modified DBPA algorithm, this factor is d_0 , which is any distance where the received power is known a priori.

Once the transmitting power is adjusted, the transmitted signal of the k th node can be written using Eq. 1 as,

$$x_k(t) = \sqrt{p_k T_b} b_k(i) c_k(t - iT_b), \quad iT_b \leq t \leq (i + 1)T_b \quad (8)$$

T_b is the bit duration and E_k is the bit energy of the k th node. Now the composite signal received by the m th node ($k \neq m$) is given by the following equation,

$$r_m(t) = \sum_{k=1}^{K/2} \alpha_{m,k} x_k(t) + n(t) \quad (9)$$

where $\alpha_{m,k}$ is the attenuation factor and it is related to the path loss of SUI channels by the following relation given in [6],

$$PL(d)dB = 20\log_{10}(4\pi d_0/\lambda) + 10\alpha \log_{10}\left(\frac{d}{d_0}\right) + s \quad (10)$$

$$\alpha_{m,k} = \sqrt{PL_{m,k}} = \sqrt{\left(\frac{4\pi d_0}{\lambda}\right)^2 \left(\frac{d_0}{d}\right)^\alpha} \quad (11)$$

Here, $PL_{m,k}$ is the path loss when the devices m and k are communicating, d is their distance, s is the shadowing factor, λ is the wavelength of the passband signal passing through the channel. If m th node wants to receive the data of j th node, then using Eq. 9, the soft output of conventional detector corresponding to j th node can be written as,

$$y_j = \frac{1}{T_b} \int_0^{T_b} \left(\sum_{k=1}^{K/2} \alpha_{m,k} \sqrt{p_k T_b} c_k(t) b_k(t) + n(t) \right) c_j(t) dt \quad (12)$$

Assume path loss exponent $\alpha = \alpha'$, then using Eqs. 7 and 11, we can write Eq. 13 as,

$$y_j = \frac{1}{T_b} \int_0^{T_b} \left(\sum_{k=1}^{K/2} \kappa c_k(t) b_k(t) + n(t) \right) c_j(t) dt \quad (13)$$

where,

$$\kappa = \sqrt{\beta} \left(\frac{4\pi d_0}{\lambda} \right)$$

As a result of MAI, we can write Eq. 13 as,

$$y_j = \kappa b_j + \sum_{\substack{k=1 \\ k \neq j}}^{K/2} \int_0^{T_b} c_k(t) c_j(t) dt \cdot \kappa b_k + \int_0^{T_b} n(t) c_j(t) dt$$

$$y_j = \kappa b_j + \sum_{\substack{k=1 \\ k \neq j}}^{K/2} \rho_{k,j} \cdot \kappa b_k + \int_0^{T_b} n(t) c_j(t) dt$$

$$y_j = \kappa b_j + MAI_j + z_j$$

where $\rho_{k,j}$ is defined in Sect. 3. We can write Eq. 13 in matrix form as,

$$\mathbf{y}_{proposed} = \mathbf{R} \mathbf{A}_{new} \mathbf{b} + \mathbf{n} \quad (14)$$

\mathbf{R} , \mathbf{y} and \mathbf{b} are the same as mentioned in Sect. 3. The matrix \mathbf{A}_{new} has been changed and is given by,

$$\mathbf{A}_{new} = \text{diag}\{\kappa, \kappa, \dots, \kappa\}_{K/2 \times K/2} \quad (15)$$

Equations (14)–(15) show the same MAI has been introduced from all the nodes, irrespective of their locations relative to the receiving device. This MAI has been completely mitigated by the use of Multiuser Detection (MUD) technique, such as MMSE detector [7]. The MMSE detector is a linear multiuser detector which applies a linear transformation on the soft output of the conventional detector. For MMSE detector, data at the output of MMSE detector will then be

$$\mathbf{y}_{MMSE} = (\mathbf{R} + \sigma^2 \mathbf{A}^{-2})^{-1} (\mathbf{R} \mathbf{A} \mathbf{b} + \mathbf{n}) \quad (16)$$

To simplify the above equation, the Matrix Inversion Lemma [8] has been used. It is given as,

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1} + \mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1} \quad (17)$$

Taking,

$$\mathbf{A} = \mathbf{R}, \mathbf{C} = \mathbf{A}^{-2} = \left(\frac{1}{\kappa}\right)\mathbf{I}, \mathbf{U} = \sigma^2\mathbf{I}, \mathbf{V} = \mathbf{I}$$

Then, the MMSE detector output can be written as

$$\mathbf{y}_{MMSE} = \kappa \left[\mathbf{I} + \left(\left(\frac{\kappa}{\sigma} \right)^2 \mathbf{I} + \mathbf{R}^{-1} \right)^{-1} \mathbf{R}^{-1} \right] \mathbf{b} + \mathbf{n}' \quad (18)$$

The term \mathbf{n}' shows the enhanced component of white noise. Equation 18 shows that the effect of MAI has been reduced to a large extent since the term (κ/σ) is small. A special case of the MMSE detector occurs when the term $(\kappa/\sigma) \lll 1$. In this case, all the devices are decoupled and noise term is enhanced very much. This is referred to as Decorrelating detector. The output in this case will be,

$$\mathbf{y}_{DD} = \kappa \mathbf{b} + \tilde{\mathbf{n}} \quad (19)$$

The term $\tilde{\mathbf{n}}$ shows the enhanced noise term. It is clear from Eq. 19 that users are decoupled and MAI has been completely mitigated. Hence, the proposed algorithm removes the problem of MAI and Near-Far effect in DS-CDMA based wireless ad hoc Network or decreases them to a large extent (for MMSE detector).

4 Simulation Results

The proposed CDMA technique offers improved network capacity. In many wireless ad hoc networks, we have an upper bound on SNR and BER. If we go on increasing number of nodes in the network, those bounds will not be satisfied. For high bandwidth communication like CDMA, reliable communication can occur for $\text{SNR} < 0$, according to Shannon's capacity theorem. So we can set an upper bound of 10^{-3} on BER and -5 dB on SNR which results in $E_b/N_0 = 10$ dB for a spreading gain of 31. Figure 1a shows that for these bounds, the number of active nodes or the capacity can be doubled from 4 to 8 nodes by using the proposed method. The proposed method has been compared with another power control enhancement algorithm given in [9]. Figure 1a also shows that BER of the network using our proposed algorithm with 8 nodes is almost equal to the one proposed in [9] with 4 users. Hence, the network capacity is almost doubled.

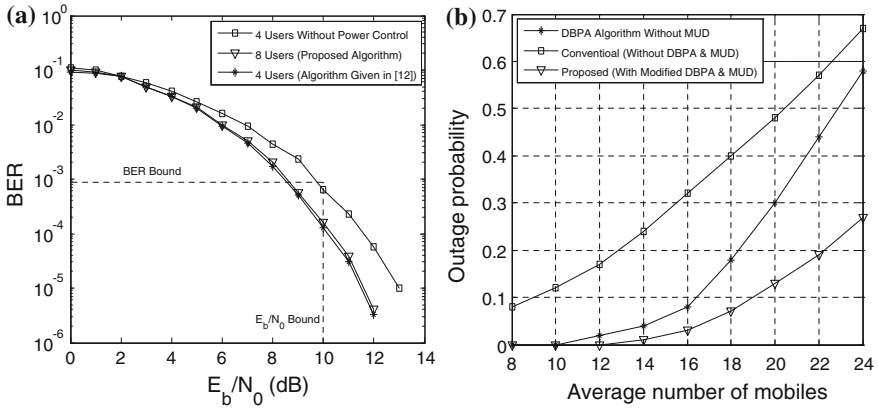


Fig. 1 Capacity enhancement in CDMA based ad hoc networks in SUI multipath channel **a** BER performance. **b** Outage probability vs number of mobile nodes

Another method of investigating the capacity of the network is the outage percentage. The outage percentage is defined as the percentage of nodes whose performance/signal reception falls below a certain specified level. Figure 1b shows the outage percentage vs number of nodes for three different techniques. The outage percentage has been calculated by using E_b/N_0 upper bound of 10 dB and BER upper bound of 10^{-3} . It can be seen that by using the proposed method, outage percentage is less than half as compared to the conventional method without MUD and DBPA, even with large number of nodes. It shows that by using MMSE-MUD and DBPA jointly as proposed in this paper, outage percentage can be decreased sufficiently.

5 Conclusions

In this paper, a secure DS-CDMA based technique for ad hoc networks has been proposed. The most prominent features of the proposed technique are twofold. Firstly, it offers improved capacity in terms of the number of devices and secondly, its performance degradation in multipath fading environment is less than that of conventional technique. The proposed technique is based on jointly using MMSE multiuser detector and the power management algorithm. The presented BER and outage probability results show a two times enhancement in the network capacity. The simulation results show that the proposed DS-CDMA technique is very effective for mobile ad hoc networks.

References

1. Kiong TS, Eng KSS, Ismail M (2003) Capacity Improvement through adaptive power control in CDMA system. In: Proceedings of 4th national conference on telecommunication technology, pp 137–140
2. Zhang J, Liew SH (2006) Capacity improvement of wireless ad hoc networks with directional antenna. In: Proceedings of vehicular technology conference, pp 911–915
3. Nuaymi L, Godlewski P, Lagrange X (2011) Power allocation and control for the downlink in cellular CDMA networks. In: Proceedings of 12th IEEE international symposium on personal, indoor and mobile radio communications, vol 1, pp C29–C31
4. Tseng Y, Wu S, Liao W, Chao C (2001) Location awareness in mobile wireless ad hoc networks. *IEEE Trans Comput* 34(6):46–52
5. Shang J, Yu S, Zhu L (2010) Location-aware systems for short-range wireless networks. In: Proceedings of international symposium on computer network and multimedia technology, pp 1–5
6. Jain R (2007) Channel models; a tutorial. Washington University, St. Louis
7. Garg M (2005) Multi-user Signal Processing Techniques for DS-CDMA Communication Systems. Master's thesis, Indian Institute of Technology, Bombay
8. Max A (1950) Woodbury, Inverting modified matrices, Memorandum Rept. 42, Statistical Research Group, Princeton University, Princeton, NJ
9. Saeed R, Khatun S, Ali B, Abdullah M (2009) Performance enhancement of UWB power control using ranging and narrowband interference mitigation technique. *Int Arab J Inf Technol* 6(2):169–178

A Study on Traceback of Illegal Users Using Anonymity Technology in BitTorrent

Geon Il Heo, Nam Hun Kim, Ah Ra Jo, Sae In Choi
and Won Hyung Park

Abstract As sharing illegal digital contents using torrent increases, the need for technology which traces the first-seeder and the seeder changed from originally a leecher is magnified to eradicate that. But it is even more difficult tracing them as the proxy software such as Easy-Tor, etc. Based Tor using anonymity network technology becomes. This paper analyzes structure and communication process of BitTorrent protocol and Tor. And based on that, this suggests the methodology for tracing the first-seeder and the seeder changed from originally a leecher in packet level and countermeasures for blocking sharing illegal digital contents.

Keywords Tracing seeder · Pirates · Illegal digital contents · BitTorrent · Tor · Easy-Tor

G. I. Heo · N. H. Kim · A. R. Jo · S. I. Choi

Global Convergent Industrial Engineering, Seoul National University of Science and Technology, 232 Gongneung-ro, Nowon-gu, Seoul 139-743, South Korea
e-mail: aza837@naver.com

N. H. Kim
e-mail: nhk1215@gmail.com

A. R. Jo
e-mail: jarjhn@nate.com

S. I. Choi
e-mail: hwa484848@naver.com

W. H. Park (✉)
Department of Information Management, Far East University, 76-32 Daehakgil, Gamgok-myeon, Eumseong-gun, Chungbuk, South Korea
e-mail: whpark@kdu.ac.kr

1 Introduction

A variety of changes of scene sprang from digital technology and develop of Internet have brought many benefits in contents industry circles. But the technological advancement didn't always bring positive aspects. This made digital contents pirates produce pirated digital contents equal to genuine thing using digital technology much easier without any restriction. And new contents distribution channel appeared, there appeared many ways that contents piracy is committed.

The torrent has taken users who use other distribution channel such as Webhard, P2P, etc. This originates in balloon effect of an advance registration system and the spread of the contents established an alliance, the rigid enforcement of the regulations of the portal sites, etc [1].

Like this, as sharing illegal digital contents using torrent increases, the need for technology which traces the first seeder and the seeder changed from originally a leecher is magnified to eradicate that. But it is even more difficult tracing them as the proxy software such as Easy-Tor, etc. based Tor using anonymity network technology becomes.

In this paper, we analyze structure and communication process of BitTorrent protocol and Tor. And based on that, we suggest the methodology for tracing the first seeder and the seeder changed from originally a leecher in packet level and countermeasures for blocking illegal digital contents sharing.

2 Related Work

2.1 *BitTorrent Protocol*

BitTorrent protocol is the representative protocol using P2P and uses several for one file sharing technique: client–client model. One client generates multitudinous sessions with a number of clients.

The operating of BitTorrent protocol is divided into four steps.

First, a user downloads a torrent file (.torrent) from a sharing site and executes the torrent file in a torrent client program. Next, the user sends Tracker Request message to a tracker, at the same time, requests IP list about peers with delivering parameter. After that, the tracker delivers the IP list through Tracker Response message. Finally, the peer who got the IP list exchanges the file with other peers [2, 3].

2.2 Structure and Operating Principles of Tor

Tor is a distributed overlay network designed to anonymize TCP-based applications like web browsing, secure shell, and instant messaging and client choose a path through the network and build a circuit. Message is located in cells and unwrapped at each node or onion router with a symmetric key. The onion routers only know the successor or predecessor but not any other onion router [4, 5].

The operating principles of Tor are divided into 3 steps as shown in Fig. 1.

First, Tor client obtains a list of Tor nodes from a directory server. Next, Tor client picks a random path based on the list to the destination server. In Fig. 1, green links are encrypted and red links are in the clear. Finally, If the destination is changed and a new session is created, a Tor client creates a new random path [7, 8].

2.3 Easy-Tor

Easy-Tor is a domestic software which employs proxy function in not only the Mozilla Firefox Browser but also other web browser, even torrent client. Seeders sharing illegal digital contents are easily able to conceal their IP through this program [9].

3 Illegal Digital Contents Trace Technology

This chapter describes the technology which traces the first seeder and heavy seeder.

3.1 A Study of Distinguishing Between the First Seeder and the Seeder Changed From Originally a Leecher, in General Environment

3.1.1 Analysis of Major Parameters of Tracker Request Message

In general environment, we analyzed about 'downloaded' and 'left' among major parameters of tracker request message by file sharing started. We assumed that the first seeder had value 'downloaded = 0' and 'left = 0' then the seeder changed

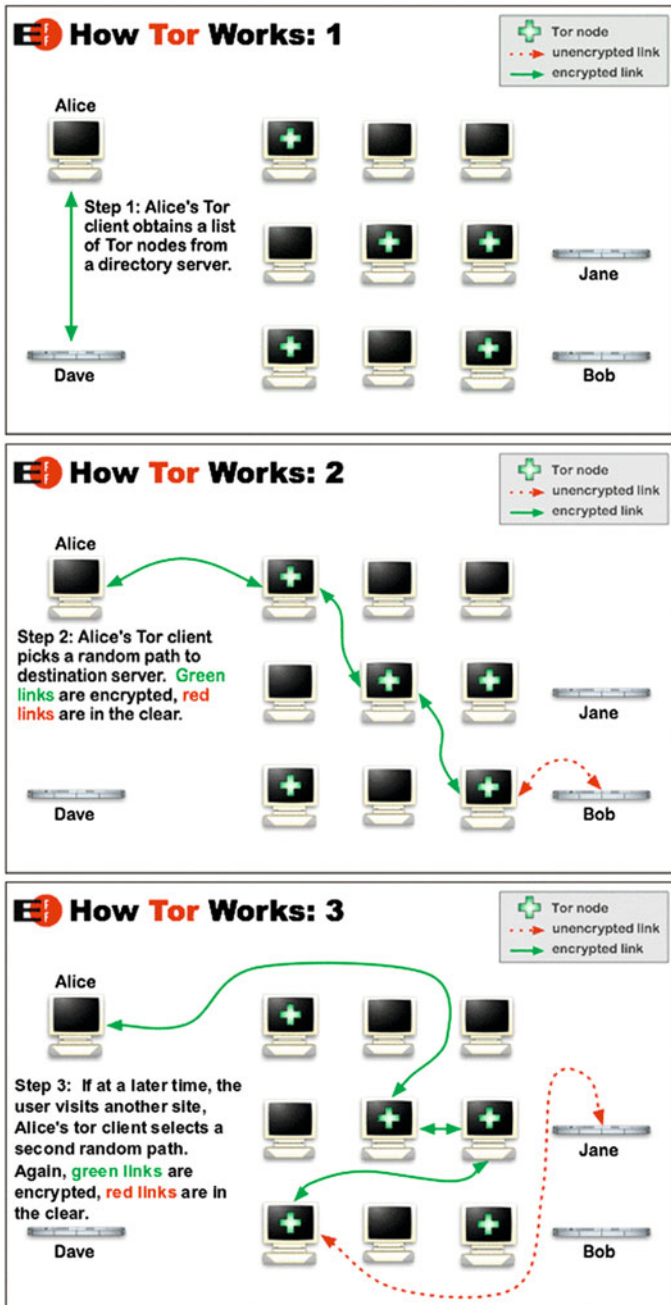


Fig. 1

from originally a leecher had value ‘downloaded = random number’ and ‘left = 0’. The analysis showed all two cases have value ‘downloaded = 0’ and ‘left = 0’. In other words these two parameters cannot distinguish between the first seeder and the seeder changed from originally a leecher. But in the case of ‘left = 0’, we found that the value ‘left = 0’ points to a seeder.

3.1.2 Analysis of Major Parameters Concerning Tracker Response Message

In general environment, we analyzed about ‘complete’ and ‘downloaded’ among major parameters of tracker request message by file sharing started.

We assumed that the first seeder had value ‘complete = 1’ and ‘downloaded = 0’ and then the seeder changed from originally a leecher had both complete and downloaded parameter that equal one or more. The analysis showed that our assumption is right. That is to say, if ‘complete = 1’ and ‘downloaded = 0’ as the [Sect. 3.1.2](#) and ‘left = 0’ as [Sect. 3.1.1](#), then there is high probability of being the first seeder.

3.2 A Study of Identifying and Tracing Client which has Real IP to be Changed Periodically, in Proxy Environment

This section presents whether we identify and trace the client which has real IP to be changed periodically.

3.2.1 Analysis of Major Parameters of Tracker Request Message

In proxy environment, we analyzed about peer_id among major parameters of tracker request message. When real IP is changed periodically, we analyze a change of peer_id. The analysis showed that even if real IP had changed, peer_id had been steady. Thus we can trace and identify the client which has real IP to be changed periodically using the peer_id in the environment.

In this experimental environment, the seeder’s IP is verified as private IP. But it will be solved if the location capturing packet is changed.

4 Countermeasures

In this chapter, Countermeasures are proposed for interrupting share behaviors with illegal digital contents.

4.1 Block Sites which Share Seed Files

In [Chap. 3](#), it suggested some plans for tracing a seeder. If we consider effectiveness in time/monetary aspects, to interrupt torrent site that shared seed files originally is more efficient than to trace.

4.2 Block Trackers' IP Address

Currently, most of the torrent users are sharing file through tracker. For that reason, after utilizing torrent site which provides tracker lists or collecting packet in backbone, we can stop sharing of illegal digital contents by blocking the high frequency of trackers.

However, we must be careful when blocking trackers because anti-virus companies or game companies often use torrents for purpose of engine updates of their products.

4.3 Block Exit Node's IP address

As mentioned in [Chap. 2.2](#), Exit Node's IP address can be distinguished because the link between final Tor node(Exit node) and destination is unencrypted. Thus, if we obtain and block Exit Node's IP [10] address in advance, it is possible to block users who share illegal digital contents through Tor.

5 Conclusion

To realize the methodology for tracing the first seeder and heavy seeder and countermeasures suggested in this study, institutional support must be preceded prior to action. Although a security administrator requires the actual users' information from ISP companies or insists about the block for certain IP obtained through collected packet, ISP companies have no responsibility for accepting it. Departments related to government have to try not to be at odds on technology and system. For next study, we will enhance basic tracing methodology continuously through update of BitTorrent protocol specification and research for tracker by protocol.

References

1. 2012 Copyright Protection Report. Korea Copyright Protection Center, May 2012
2. NMC Consulting Group (2011) Netmanias technology document: concept of BitTorrent protocol, May 2011
3. Wiki Theory BitTorrent protocol specification v1.0. <http://wiki.theory.org/BitTorrentSpecification>
4. Dingleline R, Mathewson N, Syverson P (2004) Tor the second-generation onion router. In: Proceedings of the 13th USENIX security symposium, Aug 2004
5. Syverson P (2011) A peel of onion. In: ACSAC'11, Dec 2011
6. Syverson P Theory and design of low-latency anonymity systems (Lecture 2). U.S. Naval Research Laboratory, Washington
7. Tor. <https://www.torproject.org>
8. Jung Hyung L et al (2011) A study on the countermeasure of cyber attacks using anonymous network. Korea Information Assurance Society, June 2011
9. Easypotal. <https://sites.google.com/site/essyt-or/home>
10. Tor Network Status. <http://torstatus.blutmag-ie.de/>

Internet Anonymity in Syria, Challenges and Solution

T. Eissa and Gi-hwan Cho

Abstract Surfing the web without anonymity in Syria has risked many activists and journalists' lives. Furthermore, using the available anti-censorship techniques is challenging in such country due to Internet bandwidth limitations and the extensive filtering performed by Syria's licensed Internet service providers. In this paper, we highlight the Internet censorship in Syria. We propose a solution that provides suitable anti-censorship for Syrian activists based on TOR. We shift most of the heavy computational tasks from the client to our network nodes. The cryptography tasks are chosen lightweight to suit the bandwidth limitations. We show that our solution can provide safe web surfing in Syria.

Keywords Anonymity · Censorship · TOR · Anti-censorship

1 Introduction

Internet filtering in Syria is the deepest of its kind in the political areas. Online magazines talking about human rights or political news are blocked. Syria has been reported as enemy of Internet and [1] ranked number 176 of 179 in the press freedom index [2]. The "State of Emergency" applied by Syrian government allows authorities to arrest media workers, activists or any person who criticizes the ruling regime. Ten journalists have been killed and 13000 activists have been

T. Eissa (✉) · G. Cho

Division of CSE, Chonbuk National University, Jeonbuk, Jeonju, Republic of Korea
e-mail: Tamnet83@hotmail.com

G. Cho

e-mail: ghcho@chonbuk.ac.kr

arrested between March 2011 and August 2012. During the last Syrian revolutions, many activists have been tracked by Syrian Internet providers and then arrested. Some activists have been arrested after making a call to an online News channel using unprotected VOIP software while others have been arrested after hacking the email account of one of their colleagues. Furthermore, the recruited Syrian Electronic Army has been given no limits by the government to monitor the web, track the activists and hack their accounts on' Facebook, Skype or emails accounts. Information about activist, such as geographical location or his identity, can be revealed by tracking his IP address. Circumventing censorship includes: bypassing the ISP proxy to obtain access to blocked websites, providing anonymity to hide activists' IP addresses and encrypting the traffic to secure sensitive information. A lasting battle between the regime censorship systems and the anonymity techniques and tools was in full swing. However, the Internet bandwidth average in Syria is limited to 0.6 kb. Furthermore, since June 2011, Internet connections were slowed down regularly, especially on Fridays, the common day for activists to gather and protest against the regime. Using anti-censorship tools under such Internet connection is quite slow and disturbing. Furthermore, many licensed Internet service providers in Syria has blocked most VPN protocols in interval times. Table 1 shows some Internet tools status as reported by activists inside Syria on April 2012.

The Onion Router (TOR) [3] is considered the most common anti-censorship tool available in the market. However, users in Syria are complaining from delay and Internet performance problems when using TOR. This is due to either Internet bandwidth limitation or the fake bandwidth values announced by TOR relays.

In this paper, we propose an anti-censorship solution based on TOR. The relays nodes between the clients and the destinations are chosen using social trust concept. The number of encryption tunnels on the client side is decreased to minimum to reduce bottlenecks caused by bandwidth limitation. The task of finding the optimum path and running onion routing protocol is shifted from the client to a relay outside the Syrian Internet zone. The rest of this paper is structured as follows: Section 2 surveys the Internet censorship in Syria. Section 3 presents our new solution. Section 4 evaluates our solution and finally, Sect. 5 concludes the paper.

2 Internet Censorship in Syria

Internet censorship in Syria is managed and controlled by Syrian Telecommunications Establishment (STE) and Syrian Information Organization (SIO). Reporters Without Borders confirmed STE and SIO as clients of "ThunderCache", the famous filtering system [4]. Many Internet users in Syria prefer to use Cyber Cafes to access websites they cannot access in their homes or offices. However, cyber cafes owners reported to be forced to spy on their clients by the security forces using special software provided by STE. They also have been forced to

Table 1 Anonymity tools status in Syria-Damascus as surveyed in April-2012. (SAMA, TARASAUL and SYRIATEL are licensed Internet providers in Syria)

Tool	Ultra surf [16]	Security kiss [17]	Cyberghost VPN [18]	Hotspot shield [19]	Expat shield [20]	TOR [21]	Free gate [22]	Your freedom [23]
Status	Blocked	Blocked	Blocked	Blocked	Blocked	Blocked on SAMA and TARRASUL, poor connectivity on other providers(intermitently)	Blocked	Blocked on SYRIATEL 3G, poor connectivity on other providers.

register clients' names and identities to be revealed when requested. Furthermore, the administrators of websites hosted inside Syria are forced to reveal articles authors' names and identities to security forces when requested. Anonymizer websites were blocked as well especially after the 2011 Syrian uprising. VPN-PPTP [5] and VPN-L2TP [6] have been blocked in Syria after China and Iran. Only VPN-SSTP is still working and cannot be blocked unless the Internet provider decides to block HTTPS traffic. However, it remains unstable as reported by activists inside Syria.

The most common Internet filtering techniques used in Syria are listed below:

- 1 Header-based filtering [7]: in this type of filtering, the filter checks the packet header and drops all the packets with a destination matching one of the blocked website IP addresses list. This technique is commonly used by Syrian licensed Internet provider. On 2008, more than 200 websites (including YouTube, Facebook, and Hotmail) has been blocked in Syria using this technique [8].
- 2 Content-based filtering: The filter searches for keywords inside the packet payload such as political, religious or social keywords. Deep Packet Inspection (DPI) [9] is a filtering technique where the TCP/IP payload is deeply inspected to identify the type of the incoming packets and then gives priority for some packets to increase the internet performance. DPI inspects in detail the contents of the transferred data and searches for matches with specific patterns and signatures to identify the type of service being requested. Telecomix [10] reported the existence of blue coat [11] DPI filtering devices inside Syria in 2011).
- 3 DNS tampering [12]: in this filtering, the censor provides the ISP's DNS server with a list of websites' domain names to be blocked. When ISP's DNS server receives a DNS request for one of these names, he does not reply with the corresponding IP address but instead he replies with a message such as "the requested page does not exist" or "the requested page cannot be found".
- 4 HTTP proxy based filtering [13]: This filtering technique is used when the ISP uses a proxy server as a broker between clients and the Internet. The ISP can block the entire website or pages from the website by filtering the content of the proxy server cache. We analysed the network traffic from a router hosted by the Syrian Telecommunications Establishment, it has been detected the existence of http proxy which refers to the http proxy filtering. This means that the router forward the Internet traffic to a web proxy which in turn block or allow this traffic according to a black list. It has been found also that the ports 5060 (SIP [14]) and 1723 (PPTP-VPN) were closed. The port 22 (SSH) was open; however, users complain interval disconnection due to the deep packet inspection applied on this port.

3 Our Proposed Solution

3.1 Preliminaries

3.1.1 Proxy Encryption

The proxy encryption had been proposed by Blaze et al. [15]. In this encryption, Alice may ask a third party to modify an encrypted message which has been encrypted before by Alice’s key. The third party should be able to re-encrypt the message without obtaining access to the original message (plaintext). All he needs is the re-encryption key (proxy key) provided by Alice. The purpose of this encryption is to enable another user Bob to decrypt this message using his own key.

3.1.2 Thread Model

Our adversary is a censor installed in the local ISP or the international Internet gateway. His goals are to prevent access to certain political websites and Internet tools and to spy on Internet users’ traffic. To achieve these goals, he use the filtering techniques mentioned in Sect. 2. The censor is also equipped with advanced Internet filtering devices that can block or apply packet droppings on specific security protocols. This adversary is able to execute passive attacks such as traffic monitoring and analysis. He is also able to perform active attacks such as traffic modification, deletion and generation. We do not consider the situations where the censor disconnects the Internet connectivity altogether in times of uprising. We also assume that the censor does not block HTTPs protocol. The censor may slow down Internet connection regularly in certain days.

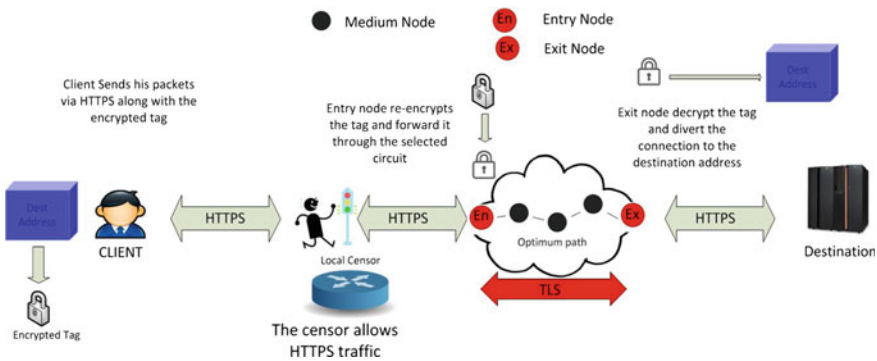


Fig. 1 Our proposed solution design

3.2 System Design

Our system approach is close to TOR in terms of network design and routing infrastructure. Fig. 1 shows the proposed solution design. Our main goal is to decrease the bottlenecks and overhead on the client device. Therefore, we shift the task of finding the optimum path from the client to the entry node to decrease the overhead towards the client side.

To reduce bottlenecks, we decrease the number of encryption tunnels to be established by the client from three to one tunnel. After establishing the circuit between the entry and exit node, the client constructs a dynamic encrypted tag called “DST” using the exit node public key: $DST = PKE - ENC (Param_{Dest}, Pk_{Exit})$. Where Pk_{Exit} represents the exit node’s public key, $Param_{Dest}$ represents information about the destination address. Then it generates a proxy key $\pi_{S \rightarrow ex}$ to enable the entry node to re-encrypt “DST” tag so that only the exit node can decrypt it:

The client then creates a hash value to be used by the exit node for verification. The client then connects to entry node via HTTPS. The entry node then re-encrypts “DEST” tag using proxy re-encryption as following: $Re - Enc(DST) = DST \oplus \pi_{S \rightarrow ex}$

It then forwards the packets to the next hop in the circuit using onion routing. The exit node removes the TLS encryption layers using the Onion Routing protocol, separates the data from the packets and decrypts the re-encrypted tag as following: $Param_{Dest} = Proxy - Dec(DST, Pr_{Exit})$. Where $Proxy - Dec$ represents the proxy decryption function, Pr_{Exit} represents the exit node’s private key.

Afterwards, the exit node extracts the destination address and diverts the connection to the destination server. Upon receiving the destination response, the exit node forwards the response through the opposite circuit from the exit node to the entry node using onion routing. The entry node then diverts the connection toward the client via HTTPS.

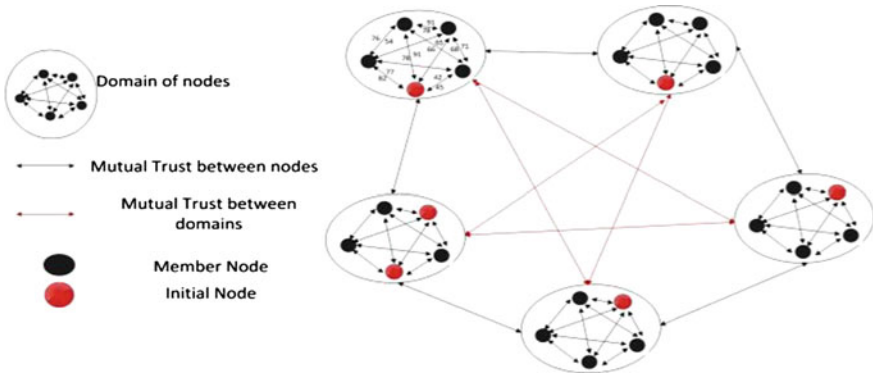


Fig. 2 An example of the mutual trust. The red nodes represent initial nodes, which are chosen by our servers. Each domain should contain at least one initial node

Because of the bandwidth limitations and the extensive censorship applied on Syrian devices, we choose the entry, exit and medium relays to be behind the Syrian Internet international gateway. Each relay is measured by three different parameters: Friendship, Bandwidth and Availability. Each relay is assigned with a score value which is calculated as following:

$$\text{Node - score}[R_i] = F(fr, bnd, av, R_i) = fr \times l_{fr} + bnd \times l_{bnd} + av \times l_{av}$$

l_{fr} : the allowed friendship latency.

l_{bnd} : the bandwidth latency.

l_{av} : the availability latency.

R_i : the relay index.

The client chooses an entry node and an exit node according to the node-scores values of each node. The entry node then chooses the optimum path to the exit node as following:

$$\text{path - score}[X_i] = \sum_{i=1}^h \frac{\text{node - score}[i]}{h}$$

where h refers to the number of nodes in the path X_i .

The entry node then chooses the path with a maximum path-score value:
 Optimum-path = max (path-score[X_i])

3.3 The Social Trust Establishment

The entry node chooses the optimum path to the exit node according to social trust value. This value represents the trustworthiness of the node as evaluated by its friends. Establishing trust between nodes is cumulative process where nodes help each other to build trust using a voting system. The mutual trust begins from the initial nodes and grows up to include new nodes. Figure 2 shows the mutual trust framework where we divide the nodes in domains based on their geographical location. Each domain should contain at least one initial node chosen and trusted by our servers. The mutual relationship can be constructed between nodes and domains.

4 Evaluation

In this section, we compare our solution with the existing TOR version. We concentrate on reducing overhead and bottlenecks towards the client side.

Table 2 - Comparison between original TOR and our modified TOR in the perspective of client

Feature	Original TOR	Our solution
Number of encryption tunnels	3	1
Optimum path finding task	Client device	Entry node device
Finding entry and exit nodes	Client device	Client device
Bandwidth consumption on client device	3 TLS tunnels	1 HTTPS connection

4.1 Computation Overhead

In original TOR, the computational overhead on the client device is generated by running cryptography operations to establish the onion routing encryption tunnels (at least 3 TLS tunnels are required). While in our solution, the tunneling cryptography operations are established by the entry node. The client only establishes HTTPS connection with the entry node. Establishing one HTTPS connection is less overhead than establishing three TLS tunnels. Furthermore, in the original TOR, the client should perform all the calculations required to find the optimum path to the destination and the circuit relays (entry, exit and medium nodes). While in our version, the optimum path calculation is shifted to the entry node. Therefore, the client should only find the optimum entry and exit nodes.

4.2 Bottlenecks Reduction

In the original TOR, the client bandwidth is consumed by three TLS encryption tunnels. In other words, the data generated by the client will be encrypted 3 times before leaving the international Internet gateway. While in our version, the client Internet bandwidth is only consumed by HTTPs connection. However, by selecting high featured entry, medium and exit nodes according to our social trust concept, the only problem that may cause a bottleneck in the network happens when the entry node or exit node has a problem in reaching each other or reaching the destination. However, it may happen that one link between the exit node and the destination, or between the entry and exit node is broken or has a bottleneck during the session. This may delay the response from the destination. To reduce this bottleneck, we propose using a Status Update Message (SUM) by each node in the circuit. This message includes the current connectivity quality between the current node, the next hop and the previous hop. To avoid a false positive SUM that may occur by modifying the content of SUM by a Man-in-Middle attacker, the reported node should attach a hash value with this message so that the other interacting nodes can verify the integrity of these messages.

Table 2 shows the comparison summary between the original TOR version and the modified TOR version. The comparison focuses on the client side only.

5 Conclusion and Future Work

In this paper, we addressed the internet censorship and discussed the challenges faced by activists and journalists surfing the web in Syria. We proposed a new solution based on TOR anonymity network to provide both lightweight and anonymous web surfing. The network relays are chosen behind the international Internet gateway with high bandwidth resources. The tasks of establishing onion routing layers is shifted to our network relays to decrease the overhead towards the client. The number of encrypted tunnels to be established by the client is reduced to one. The future work of this research is to implement this solution with cooperation with TOR network members.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (KRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2042035)

References

1. Reporters Without Borders (2012) Enemies of the Internet. <http://En.Rsf.Org/Beset-By-Online-Surveillance-and-12-03-2012,42061.html>. 19 March 2012
2. Reporters Without Borders (2011) Press freedom index 2011/2012. <http://En.Rsf.Org/Press-Freedom-Index-2011-2012,1043.Html>
3. Dingleline R, Mathewson N, Syverson P (2004) Tor: the second generation onion router. In: Proceedings on USENIX security '04, Aug 2004
4. Citizen Lab And Canada Centre For Global Security Studies (2011) The Canadian connection: an investigation of Syrian Government and Hezbollah Web hosting in Canada, 16 Nov 2011
5. Ancillotti E, Bruno R, Conti M (2011) An efficient routing protocol for point-to-point elastic traffic in wireless mesh networks. In: Proceedings on WOWMOM '11, 2011
6. Mingming H, Qin Z, Kuramoto M, Cho F, Zhang L (2011) Research and implementation of layer two tunneling protocol (L2TP) on carrier network. In: Proceedings on IC-BNMT, pp 80–83
7. Deibert R, Palfrey J, Rohozinski R, Zittrain J (2008) Access denied: the practice and policy of global Internet filtering. Information revolution and global politics. MIT Press, Cambridge
8. Freedom House Report (2012) <http://www.Freedomhouse.Org/Report/Freedom-World/2012/>
9. Liao M, Luo M, Yang C, Chen C, Wu P, Chen Y (2012) Design and evaluation of deep packet inspection system: a case study. Netw IET 1(1):2–9
10. <http://Telecomix.Org/>
11. Newswire PR (2004) Blue coat delivers high-performance Web filtering using ISS Proventia WEB Filter Technology. 26 Oct 2004, Source: Proquest Central
12. Roberts P (2010) Chinese DNS tampering a big threat to Internet security. <https://Threatpost.Com/En%5Fus/Blogs/Chinese-Dns-Tampering-Bigthreat-Internet-Security-112410>
13. Du G, Zhang Z, X. Wu. "HTTP Proxy Server Based on Real-Time Link," in Proc. on 2010 MINES, pp 169–173, 4–6 Nov 2010
14. Sinnreich H, Johnston AB (2006) Internet communications using SIP: delivering VoIP and multimedia services with session initiation protocol (networking council). Wiley, New York

15. Blaze M, Bleumer G, Strauss M (1998) Divertible protocols and atomic proxy cryptography. In: Proceedings on advances in cryptology—eurocrypt'98, 1998
16. Ultrasurf—Free Proxy-Based Internet Privacy and Security Tools. <http://Ultrasurf.U.S/Securitykiss>—Free VPN Service; <http://www.Securitykiss.Com/>
17. Surf Anonymously. <http://Cyberghostvpn.Com/>
18. Anchorfree hotspot shield | hotspotshield.com. www.Hotspotshield.com/Free-Elite-Trial/?Gclid=Ckqelqpbvbicfutt4god4ioazg
19. Expat shield. <http://www.Expatshield.Com/>
20. Tor project: anonymity online. <https://www.Torproject.org/>
21. Freegate | Global Internet freedom consortium. <http://www.Internetfreedom.Org/Freegate>
22. Your freedom—bypass firewalls and proxies, stay anonymous. <https://www.Your-Freedom.Net>

Part IV

Convergence Security

Comparison of Attacks and Security Analyses for Different RFID Protocols

Jung Tae Kim

Abstract RFID is a widely adopted in the field of identification technology these days. Radio Frequency Identification (RFID) has wide applications in many areas including manufacturing, healthcare, and transportation. Because limited resource RFID tags are used, various risks could threaten their abilities to provide essential services to users. Existing RFID protocols are able to resolve a number of security and privacy issues, but still unable to overcome other security & privacy related issues. In this paper, we analyzed the associated vulnerability and threat to the confidentiality, availability of the information assets for privacy, and mitigate the risks in RFID application. Considering this RFID security issues, we survey the security threats and open problems related to issues by means of information security and privacy. The security functions to be adopted in a system, strongly depend on the application.

Keywords Security · Privacy · RFID protocol · Ultra-weight algorithm

1 Introduction

RFID is expected to be the basic technology for ubiquitous network or computing, and to be associated with other technology such as telemetric, and wireless sensors. Recently, the wide deployment of RFID systems in a variety of applications has raised many concerns about the privacy and the security issues. RFIDs have applied widespread use in many commercial product as well as national security applications, ranging from e-passports, contactless credit cards to supply chain

J. T. Kim (✉)

Department of Electronic Engineering, Mokwon University, 800, Doan-dong,
Seo-ku, Daejeon 302-729, South Korea
e-mail: jtkim3050@hankook.ac.kr

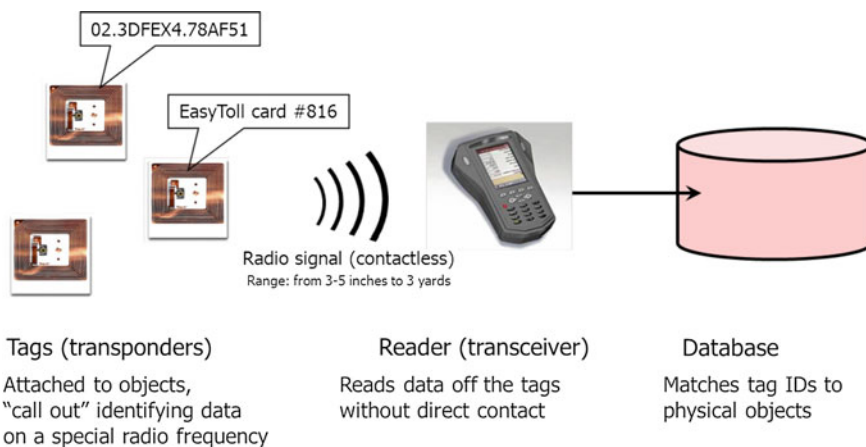


Fig. 1 Basic configuration of RFID system

management. Since RFID tags are wireless and very tiny, attached to diverse items, and often oblivious to the human user, privacy is a major concern in the design and use of RFIDs. Indeed, these tags are commonly embedded in personal devices. Another issue related to the design of RFID is the computational performance required at the tag side. These days many systems applicable to ubiquitous surroundings use wireless communications. Within such environments the majority of devices have limited computing resources, small storage and low power supply. It needs light-weight protocol to implement. As shown in Fig. 1, basic RFID system consists of six main components [1].

A basic RFID infrastructure consists of 3 major components: (1) tags, (2) a reader and its antenna, and (3) middleware application software. RFID tags are transponders that contain a chip and an antenna. The antenna receives radio signals from the readers while the chip is used to respond to the signals and to store information. Current tags are small enough to be embedded into physical objects for identification and tracking. Tags can be read only, write once/read many times or read/write capable. Because of its limited resources, two main privacy problems related to the RFID systems can be occurred: (a) leakage of information of the tags should be avoided, otherwise it would allow clandestine inventory and, (b) tracking the behavior of its user should be prevented.

2 Related Works

Privacy and security solutions can be divided into two groups: hardware solutions and software solutions. Hardware solutions are related to some controls of process variations in each integrated circuit, killing a tag or blocking a tag. As consumers use some readers to scan their own tags, a technique to protect context of tag is kill

the tag. Indeed, a kill command can be used to destroy a tag. However, a killed tag is truly dead and can never be reactivated. This is a key disadvantage for the “killing tag” technique. Different from the “killing tag” technique, the blocking tag method involves no modification to the consumer tag. Based on the “tree walking” method, a blocking-capable tag creates an RF environment to prevent unauthorized scanning of consumer items in order to “spam” misbehaving readers, i.e. there is no way to locate the protected tag’s ID. Generally, the bit used to “spam” misbehaving readers is one of 28-bit EPC management. Software solutions are based on the exchange of messages to establish authentication between two entities. Although mutual authentication without any hash function is used, most software solutions use general hash functions (like MD4 and SHA-1) to support access control and authentication. A number of security researchers have published papers on security solution. Most of the sources of security and privacy issues arise from the violation of the air interface between a tag and its reader [2]. In general, memory cost depends on digital logic gate counts. Normally, logic gates to be used for security are from 250 to 3000 in a tag. In conventional scheme, a tag store one key and two binary matrices. A tag need not store the implementation of low-cost cryptographic primitives which can be constructed with 6–13 K gates. It results in the usage of less memory than previous scheme. The conventional scheme requires a lightweight bitwise operation both in a tag and a reader. It reduces the burden on database in process of searching a tag ID as well as on a tag to operate. Communication cost can be counted number of protocol and memory size in a tag.

3 Open Issues of RFID System

Threats and privacy concerns with a low-cost RFID system should satisfy the following security requirements such as privacy, integrity, authentication, anonymity/untraceability, and even availability. Mobahat has analyzed authentication and lightweight cryptography in low cost RFID [3]. He analyzed authentication requirements not only to supply and use tags that have data authentication ability but also consider the low cost compatible cryptography protocols such as lightweight ones. The approach and comparison of protocol is based on information leakage, spoofing attack, replay attack, indistinguishability, forward security, resynchronization, no traceability, mutual authentication, DOS attack and loss of message. Security and privacy issues of RFID tags can effect both organizations and individuals. Unprotected tags may be vulnerable to eavesdropping, traffic analysis, spoofing or denial of service and many more. As we know basic RFID tags are unable to perform true cryptographic operations since it has a couple of thousand gates, which is mainly for basic operations and very few of remaining is available for security functions. Pateriya et al. has surveyed the evolution of RFID security and privacy [4]. He has described the various proposed approaches to the privacy problem. The main approaches are as follows.

- Tag killing and sleeping approach, blocking, soft blocking, and relabeling approach
- Re-encryption, minimalist cryptography and proxy approach

4 Hardware Based on Crypto Algorithm

Research and development in RFID has been focused on hardware and firmware components such as active and passive RFID tags, readers, and embedded software. The widespread deployment of RFID technology may generate new threats to security and user privacy. One of the main drawbacks of RFID technology is the weak authentication systems between a reader and a tag. In general, “weak” authentication systems that either leak the password directly over the network or leak sufficient information while performing authentication allow intruders to deduce or guess the password. A few papers explore primitive geared at the very tightly constrained environments of RFID tag. Feldhofer, Dominikus, and Wolkstorfer proposed a lightweight hardware implementation of a symmetric-key cipher, namely, a 128-bit version of the Advanced Encryption Standard (AES). Their design requires just over 3,500 gate equivalents—considerably more than appropriate for basic RFID tags, but suitable for higher cost RFID tags [5]. Juels and Weis proposed a lightweight authentication protocol called that has security reducible to a problem called Learning Parity with Noise. To implement tags needs, it only generates random bits and compute binary dot products. The key lengths required for good security are not known yet, however, and the security model is limited [6]. Huang et al. has proposed Hardware Implementation of RFID Mutual Authentication Protocol and the output waveforms from the FPGA were displayed on the 16702A logic analysis system for real-time verification [7]. Kavitha et al. has implemented RFID with secure mutual authentication protocol using FPGA (Field programmable gate array) and microcontroller [8].

We analyze the standardized cryptographic algorithms SHA-256, SHA-1, MD5, AES-128, and ECC-192 in terms of different specification. The three parameters mean are used to classify a metric of hardware implementations such as power consumption, chip area, and the number of clock cycles. The results and a comparison of the different hardware implementations are depicted in Table 1. The chip area results are based on synthesis and are given in gate equivalents (GE) [9].

Table 1 Syntheses and simulation results on 0.3 μm CMOS [9]

Algorithm	Security (bit)	I (μA)	Chip area (GE)	Clock (Cycles)
SHA-256	128	5.86	10,868	1,128
SHA-1	80	3.93	8,120	1,274
MD5	80	3.16	8,001	712
AES-128	128	3.0	3,400	10,332
ECC-192	96	18.85	23,600	502,000

Table 2 Threat analyses matrix of RFID system [10]

Threats	Affected RFID component	Risk mitigation
Rogue reader	Tag, air-interface, reader	Reader authentication
Eavesdropping	Tag, air-interface	Encryption the data, shielding the tag or limit the tag-reader distance
Reply attack	Air-interface	Using short range tags, shielding the tag or implementing the distance bounding protocol
Replay attack	Tag, air-interface	Encryption the data, shielding the tag or limit the tag-reader distance, tag authentication
Tag cloning	Tag, air-interface	Tag authentication
Tracking object	Tag, air-interface	Low range tags or shielding tags, authenticating the readers or disabling the tags
Blocking	Air-interface	Detect early and localize, take appropriate action
Jamming	Air-interface	
Physical tag damage	Tag	Use protective material

5 Threat of RFID Protocol

In order to give a better understanding of attacks, we firstly provide an overview of some possible attack types and privacy risks relevant to RFID systems. The representative attacks are included such as tag impersonation attack, tag tracking attack, Denial of service (DoS) attack, replay attack and eavesdropping. Khoo has analyzed RFID as an Enabler of the Internet of Things regarding to issues of security and privacy. He demonstrated threat analysis matrix of the RFID system. Table 1 is shown the summary of threats and risk mitigation [10]. Some well-known attacks includes such as physical attacks, Denial of Service (DoS), counterfeiting, spoofing, eavesdropping, traffic analysis, relay (man in the middle) attacks and replay attacks [11] (Table 2).

6 Vulnerability of RFID Protocol

Security design of the protocol should not impede normal operations, and should prevent a malicious adversary from getting any information. We consider the following measures:

A. Secrecy/Authentication

The cryptographic methods used (for example the keyed Hash function H) correspond to the state of the art in industry today, and reasonably guarantee the secrecy of the message. Thus, we assure the recipient that the messages originate from valid sources.

Table 3 Comparison of security analyzes for different protocols [12]

	LMAP	M2AP	EMAP	SASAI	JK	Enhanced-JK
Mutual Auth.	O	O	O	O	O	O
Eavesdropping	X	X	X	O	O	O
Reply attack	X	X	X	X	O	O
Spoofing	X	X	X	X	O	O
DOS	X	X	X	X	O	O
Position detection	X	X	X	X	^	O
Forward attack	X	X	X	X	^	O

O means satisfactory, X means partial satisfactory and ^ means unsatisfactory

Table 4 Comparison of storage and operation analyzes for different protocols [12]

		LMAP	M2AP	EMAP	SASAI	JK	Enhanced-JK
Total message		4L	5L	5L	4L	3L	3L
Tag memory		6L	6L	6L	7L	2L	2L
DB memory		6L	6L	6L	4L	3L	3L
DB (read operation)	XOR	14	13	21	5	10	6
	^	1	4	4	-	2	-
	+	9	8	1	6	4	6
	Rotate	-	-	-	4	2	4
	PRNG	2	2	2	1	2	1
Tag operation	XOR	14	13	20	5	10	6
	^	-	2	3	-	2	-
	+	7	8	-	6	3	6
	Rotate	-	-	-	4	2	4
	PRNG	-	-	-	1	-	1

L means storage factor and - means unnecessary storage

B. Indistinguishableness/Tracking/Passive Replay

Using a freshly generated random nonce with every message in the protocol, it is impossible to track the tag. Assume that an adversary pretends to be a genuine reader. He sends out a query, and receives a message back. Next time he sends a query, along with a fresh nonce, he receives a different message, so he cannot track the tag. Of course, with multiple tags in an area, tracking a specific tag without keys is extremely difficult if not impossible.

C. Forward Security

This means that the current key of a tag has been found, and can be used to extract previous messages (assuming that all its past conversations are recorded). Let's say the adversary somehow finds keys. The tag always communicates using a hash function. The adversary cannot use the key to decode any of the tag's messages because the one-way hash function H is considered computationally un-invertible. In other words, the adversary needs to have access to the hash digest table for lookups. So, he cannot decipher/recreate any past messages sent with

previously used keys. There are a number of solutions proposed so far to solve the security problems and threats associated with the use of RFID systems [12]. To realize lightweight authentication protocol for low cost RFID, Chien et al. described in detail based on ECC-based lightweight authentication protocol with untraceability [13]. Most of protocols have fundamental flaws that can be readily taken advantage by a resourceful adversary. Kapoor et al. identify and discuss these vulnerabilities and point out the characteristics of protocol that exposes it to these vulnerabilities [14–16] (Tables 3, 4).

7 Conclusions

We analyzed security issues to estimate performance, threats and performance of security related to issues by means of information security and privacy. Neither a symmetric nor an asymmetric cryptographic deployment is necessarily with light weighted algorithm. The security functions to be adopted in a system, strongly depend on the application. In future work, we will develop test bed for RFID system to estimate performance and related problems.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (grant number: 2012-0007896).

References

1. Park Y-J et al (2012) On the accuracy of RFID tag estimation functions. *J Inf Commun Convergence Eng* 10(1):33–39
2. Weis SA, Sarma SE, Rivest RL, Engels DW (2004) Security and privacy aspects of low-cost radio frequency identification systems. In: *Security in pervasive computing*, vol 2802 of LNCS, pp 201–212
3. Mobahat H (2010) Authentication and lightweight cryptography in low cost RFID. In: *2nd international conference on software technology and engineering (ICSTE)*, pp 123–129
4. Pateriya RK et al (2011) The evolution of RFID security and privacy: a research survey. In: *International conference on communication systems and network technologies*, pp 115–119
5. Feldhofer M et al (2004) Strong authentication for RFID systems using the AES algorithm. *Cryptogr Hardw Embed Syst* 31(56):357–370
6. Juels A (2006) RFID security and privacy: a research survey. *IEEE J Sel Areas Commun* 24(2):381–394
7. Huang Y-J et al (2010) Hardware implementation of RFID mutual authentication protocol. *IEEE Trans Ind Electron* 57(5):1573–1582
8. Kavitha SM et al (2012) RFID implementation with secure mutual authentication protocol. In: *International conference on computing, electronics and electrical technologies*, pp 746–751
9. Phan RC-W (2009) Cryptanalysis of a new ultra-lightweight RFID authentication protocol—SASI. *IEEE Trans Dependable Secur Comput* 6(4):316–320

10. Khoo B (2011) RFID as an enabler of the internet of things: issues of security and privacy. In: IEEE international conferences on internet of things, and cyber, physical and social computing, pp 709–712
11. Kwon S-H, Park D-W (2012) Hacking and security of encrypted access points in wireless network. *J Inf Commun Convergence Eng* 10(2):156–161
12. Jeon D et al (2011) An enhanced forward security on JK-RFID authentication protocol. *J Korea Inst Inf Secur Cryptol* 21(5):161–168
13. Chiwn H-Y et al (2009) ECC-based lightweight authentication protocol with untraceability for low-cost RFID. *J Parallel Distrib Comput* 69(10):848–853
14. Kapoor G et al (2011) Vulnerabilities in Chen and Deng’s RFID mutual authentication and privacy protection protocol. *Eng Appl Artif Intell* 24:1300–1302
15. Peris-Lopez P, Hernandez-Castro JC, Estevez-Tapiador JM, Ribagorda A (2006) EMAP: an efficient mutual-authentication protocol for low-cost RFID tags. In: Meersman R, Tari Z, Herrero P (eds) OTM 2006 workshops. LNCS, vol 427. Springer, Heidelberg, pp 352–361
16. Zuo Y (2010) Survivable RFID systems: issues, challenges, and techniques. *IEEE Trans Syst Man Cybern Part C Appl Rev* 40(4):406–418

Honeypot Using Dynamic Allocation Technique with IP Scan

Hwan-Seok Yang

Abstract Network technology is developing rapidly and attack technique using this is diverse and the damage is growing. Therefore, collection of attack information and its active coping in order to handle to diverse attack become necessary urgently. Honeypot to collect attack information of attacker must provide many of the same actual resources, but it is not easy. In this paper, multiple virtual machines which utilize resources efficiently using low cost are generated by means of dynamic allocation technology and dynamic honeypot system using this is proposed. Distributed IDS by protocols is used to detect accurate intrusion of collected information. Performance of the proposed system was confirmed by the experiments.

Keywords Honeypot · Virtual machine · Intrusion detection system

1 Introduction

Attacks of hackers are diminishing and the damage has been growing according to development of information technology. In order to prevent this, various security system such as intrusion detection system (IDS), firewall, and log server are being developed and active security system recently is being developed [1–3]. New types of attacks on this basis is detected and blocked after honeypot among these become goal of attacker and attack information such as attack method and level are

H.-S. Yang (✉)

Department of Information Security Engineering, Joongbu University,
101 Majeon-ri, Chubu-myeon, Geumsan-gun, Chungnam, South Korea
e-mail: yanghs@joongbu.ac.kr

collected. Attack information about new types of attacks is very important factor that influence to performance of security system because most of security system prepare to attacks about well-known weakness [4]. Many system same as reality must exist to prevent attackers recognize honeypot, and many honeypot have to be constructed to obtain enough attack information from attacker. But a lot of cost is used to do this and it is not easy realistically. It's problem how it must be placed to detect and identify specific movements.

In this paper, virtual machine is generated to dynamically allocation to maximize utilization of resources at low cost. Dynamic honeypot system using generated virtual machine is proposed. Distributed intrusion detection system by protocols is used to detect intrusion on the basis of collected information by consisted honeypot like this.

This paper is organized as follows. [Section 2](#) shows characteristic of honeypot and [Sect. 3](#) describes dynamic honeypot using proposed dynamic allocation. Performance of proposed system is analyzed in [Sects. 4, 5](#) concludes the paper.

2 Honeypot System

Honeypot which exist to be attacked has characteristic which can perform tracing active defense after attacker is remained long [5]. It can be form like a computer system, file or record. Requirement that this honeypot has to equip is as follows. First, it has to expose easily to attacker and look weak. Every component of system is equipped and every packet which passes through system is observed. Honeypot is classified into three stages like as follows according to degree of interaction which perform with attacker [6].

- Low interaction : permit only access of specific port and service is not provided.
- Medium interaction : provide virtual service which performs like real service.
- High interaction : provide actual OS and every service

Figure 1 shows structure of honeypot. IDS, log server and honeypot are necessary when honeypot is composed like figure. It is excellent in network protection but occurrence of invalid data become many in case of that honeypot locates in front of firewall. Effectiveness becomes high and risk of internal network is increased in case of that honeypot locates behind firewall.

Honeynet is structure that uses network which is composed with many honeypots not single system to honeypot. It is very effective when action of network level which is hard to record a single honeypot alone is monitoring. Especially, it is useful to detect diffusion path or communication between zombies used DDoS or spam [7].

In the Fig. 2, each honeypot writes all commands which attacker performs and is sent to syslog server. Much information can be collected by doing so. Honeywall must perform data control function. In other words, honeypot is infected and while it prevents other server from attack using this, attacker is not aware of this and it is able to resist bandwidth limit [8]. That is why it stays for a long time to analyze attacker.

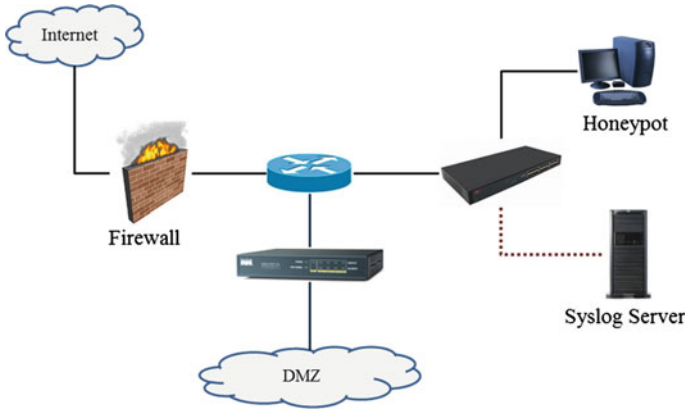


Fig. 1 Structure of honeybot

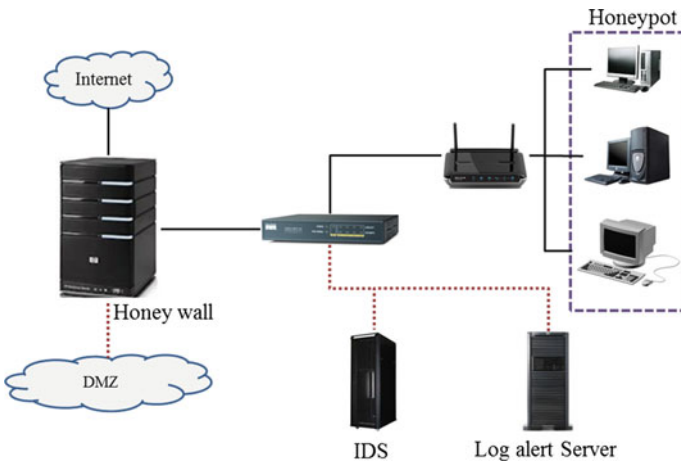


Fig. 2 Structure of honeynet

3 Dynamic Allocation Honeybot

In this section, more many resources is provided with low cost by composes dynamic honeybot using unassigned internal IP and we describe virtual machine management and honeybot operation scheme using agent.

3.1 System Architecture

In this paper, we propose dynamic honeypot system which is composed of many virtual machines in order to provide more many resources to attacker with low cost by using given resource dynamically. Proposed system in this paper is composed of honeypot control center (HCC), virtual machine honeypot farm, log server and distributed-intrusion detection system (D-IDS) to control virtual machine and manage domain. Figure 3 shows system structure.

3.2 Dynamic Allocation and Information Collect

IP address is necessary to compose virtual machine dynamically composed honeypot farm. Using address, address of sleep state and unassigned address is separated after IP of every server, client, and network device in domain do scan and stored in database. IP address scan like this is performed at regular intervals and virtual machine is generated based on this address and multiple honey farms are generated. Attacker is attempted more many intrusions by creating a lot of honeypot like this and much information is collected and information associated with the intrusion is obtained by interlocking D-IDS with this information. Intrusion detection system implements sub-IDS in order of 4 protocols as TCP, UDP, ICMP, and ARP to provide stable intrusion detection because it has a big

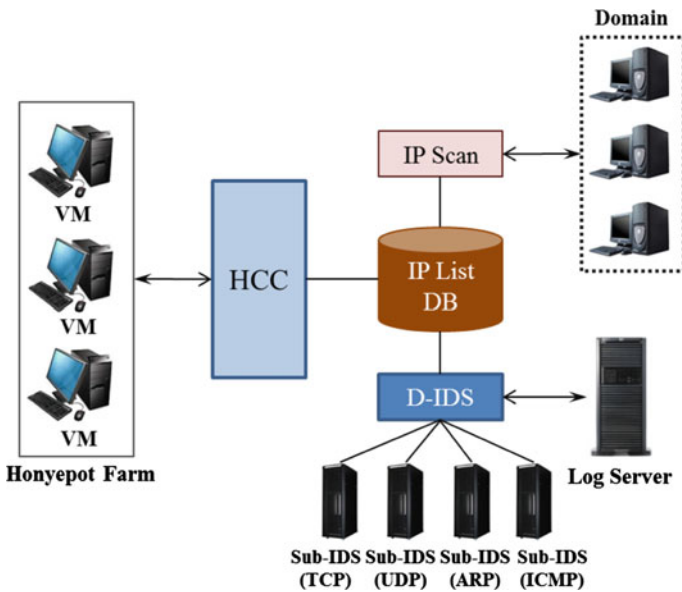


Fig. 3 Proposed system using dynamic allocation

impact depending on the amount of traffic to be processed. Rule to use in order to intrusion detection in each sub-IDS is done update from rule set DB. This system handles actively quickly about new attack by doing so. Analyzed result in distributed intrusion detection system is transmitted to HCC and performs related operation with intrusion response. Honeypot server consists of generated virtual machine dynamically by instruction of the HCC. This honeypot can be generated as IP address which is not used in the domain and intruding attack through this is processed by interworking HCC and D-IDS, and every result are stored in log server to the database form.

3.3 Honeypot Control Center

Purpose using virtualization technology to compose honeypot is because high resource utilization, low management cost, improved security, and scalability are better. HCC is composed to control of provided dynamically virtual machines and manage domain to provide more system to attacker. HCC manages domain about honeypot farm composition, management, and intrusion after virtual machine is created. It composes virtual machine dynamically as needed through cyclical scan IP address to compose honeypot farm. The result is stored in IP List DB. Change of IP list is aware happening event and created virtual machine is managed. This process is performed by dynamic management. The type of detected attacks by D-IDS is analyzed and respond about this is managed. Figure 4 shows function of HCC.

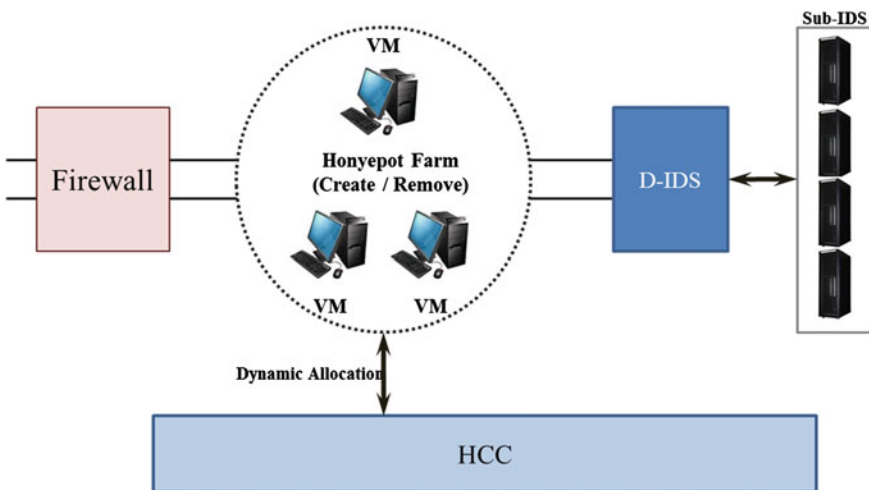


Fig. 4 The facility of HCC

Table 1 Simulation parameters

	Host Server	Honeypot
Host	203.X.Y.220	10.X.Y.50
	203.X.Y.221	10.X.Y.51
		10.X.Y.52
		10.X.Y.53
		10.X.Y.54
OS	Windows 7	CentOS 5.5
Port number	23, 80	23, 80

Table 2 Simulation result

parameter	Case #1	Case #2
Server	2	2
Honeypot	0	5
IP scan result	2	7
Telnet	0	5
Attack attempts	0	5

4 Performance Evaluation

4.1 Simulation methodology

We was assumed small and medium sized LAN using C class single domain to analyze performance of proposed system in this paper. IP address of physical server which consists of virtual machine honeypot farm is 203.X.Y.Z and 5 virtual machines are generated and honeypot farm is consisted by allocating unassigned IP address in server and network equipment. Table 1 shows used environment variables to proposed experiment of the system.

4.2 Simulation Result

We experimented to evaluate performance of the proposed system in environment generated dynamic multiple honeypot composed of virtual machine (case #1) and environment composed to general server and network equipment (case #2). Attacker scans site of experience domain using IP scan in external network and judges to server which penetration is possible, and measures whether access to honeypot. Table 2 shows IP address scan result and telnet connection attack attempt result. Proposed system as shown in the result allows access easily to attacker. Much information is collected because it is easily exposed to the attacker.

True Positive Rate (TPR) and False Positive Rate (FPR) about attack by protocol are measured to evaluate performance of D-IDS of proposed system (Fig. 5).

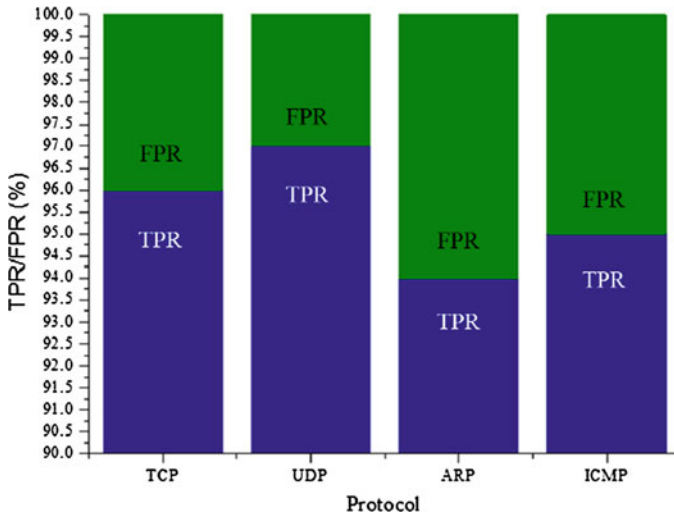


Fig. 5 Intrusion detection by protocols of D-IDS

We confirmed that D-IDS of proposed system is shown stable attack detection rate which is unaffected of traffic. We can correspond appropriately by detecting after attack information is collected using dynamic honeypot easily exposed to the attacker.

5 Conclusion

We proposed scheme which is generated multiple virtual machine to dynamic allocation and is composed of honeypot using generated virtual machine in order to provide many resources using low cost. IP address is scanned periodically and IP address information which doesn't use in domain is collected to generate virtual machine dynamically. D-IDS which performs intrusion detection by protocols is used to improve attack detection rate of attacker in generated honeypot. Proposed system in this paper is exposed easily to attacker and has high detection rate for attack. It is effective to collect and analyze information about internet, and various approach.

References

1. Niels P (2004) A Virtual honeypot framework. In: proceedings of the 13th USENIX security symposium, pp 1-14

2. Peng N, Dingbang X (2003) Learning attack strategies from intrusion alerts. In: Proceedings of the 10th ACM conference on computer and communications security, pp 200–209
3. Spitzner L (2002) Honeypots: tracking hackers. Addison-Wesley, Boston
4. Christian Plattner, Reto Baumann, White Paper: Honeypots. <http://www.inf.ethz.ch/personal/plattner/pdf/whitepaper.pdf>
5. Kreibich C, Weaver N, Kanich C, Cui W, Paxson V (2011) Practical containment for measuring modern malware system. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference, pp 397–412
6. Ali I, Thorsten H, Felix CF (2008) Monkey-spider: detecting malicious websites with low-interaction honeyclients. In: Sicherheit'08, pp 407–421
7. Corrado L, Ken M, Marc D (2005) Scriptgen: an automated script generation tool for honeyd. In: Proceedings of the 21st annual computer security application conference (ACSAC), pp 203–214
8. Nance Kara, Bishop Matt, Hay Brian (2008) Virtual machine introspection: observation or interface. IEEE Secur Priv 6(5):32–37

Attribute-Based Encryption for Commercial Content Distribution

Hyoseung Kim, Seunghwan Park, Jong Hwan Park
and Dong Hoon Lee

Abstract There is growing need for digital content distribution system in the IPTV market so that the content providers can make digital content available to users in a controlled manner. Digital right management technologies allow being able to restrict access to a content giving an access policy so content is available only to fair user. However, such a previous system has some problem in terms of management of user's access information and computational cost for server. One of solutions to overcome the abovementioned problem is attribute-based encryption scheme providing access control to the encryption system. In this paper, we propose attribute based encryption scheme which is useful in contents distribution for IPTV, DVD and access control in encrypted file systems. Our attribute based encryption scheme allows for AND, OR, threshold and wildcard operations over attribute. Furthermore, our scheme achieves $O(1)$ pairings for decryption, regardless of any number of attributes.

Keywords Attribute based encryption · Content distribution · Digital content · Access control

H. Kim (✉) · S. Park · J. H. Park · D. H. Lee
Graduate School of Information Security, Korea University, Anam-dong 5-ga,
Seongbuk-gu, Seoul, South Korea
e-mail: ki_myoo@korea.ac.kr

S. Park
e-mail: sgusa@lycos.co.kr

J. H. Park
e-mail: decartian@korea.ac.kr

D. H. Lee
e-mail: donghlee@korea.ac.kr

1 Introduction

In a broadcast system, serious breaches occur such that a malicious user copies content he has purchased and then redistributes unauthorized. Moreover, he is able to access any content without paying a charge by manipulating a billing system. Therefore, it becomes really important issues that the content providers can make digital content available to users in a controlled manner. Digital Right Management (DRM) is one of a solution for the problem with the free use and transfer of content. DRM technologies allow being able to restrict access to a content giving an access policy so content is available only to fair user. For example, according to the OMA-DRM standards [5], a hybrid encryption paradigm, where a ciphertext of content is generated using symmetric key (CEK included in RO) under the symmetric algorithm and this key (RO issued by RI) is encrypted by a user's public key, is used to efficiency of content distribution. However, In such that case, because of a separation of the roles of between the content provider and the RI, payload of managing the each DB can in no way be negligible in order to meet security such as the integrity of the access information and confidentiality of the content. In addition, whenever any user buys any new content, the RI constructs a new ciphertext of the RO using the user's public key although the ciphertext of the content is calculated and stored previously. This causes a tremendous computational cost for server. An Attribute-Based Encryption (ABE) provides access control to the encryption system as one of solutions to overcome the abovementioned problem. The concept of ABE is that allows sharing of a ciphertext and only users who have particular private key can decrypt to the ciphertext, which is different from the traditional encryption scheme in one-to-many encryption. A user's private keys and ciphertext are associated with set of descriptive attributes or an access structure, basically, made by boolean formulas such as AND, OR and threshold operations. Therefore, a broadcast system using the ABE is more efficient and simpler than the previous one since there is no independent entity, such as the RI, from content provider. The ABE furthermore does not require the certificate of a public key anymore since the ABE is an expansion of the ID-Based Encryption (IBE) [2] which has same property of certificateless. The ABE has several applications such as audit log inspection in the forensic analysis.

Attribute Based Encryption (ABE) was first introduced as a fuzzy identity based encryption by Sahai and Waters [7]. In the paper [4], the authors defined two forms of ABE: Key-Policy Attribute Based Encryption (KP-ABE) and Ciphertext Policy Attribute Based Encryption (CP-ABE). Goyal et al. [3] subsequently increased the expressibility of KP-ABE system by allowing the private key to cover any monotonic access structures over attributes. Ostrovsky et al. [6] provided an improved KP-ABE system which is allowed to express non-monotonic access structures, additionally using NOT gate associated with negative attributes. Recently, fully secure constructions were provided by Okamoto et al. [8]. However, their scheme does not support wildcards in its access policy. Zhou and Huang [9] presented a CP-ABE system which is allowed to support wildcard.

Unfortunately, their CP-ABE system is not realistic since the computational cost for decryption increases with the number of attributes.

The contribution of this paper is to construct a new ABE scheme which allows AND, OR, threshold and wildcard operation over attribute. Furthermore, our scheme achieves $O(1)$, only 3 pairings, for decryption regardless of any number of attributes. To fix the number of pairing computation to only 3, we use parallel private keys, very simple and novel method. Using wildcards makes an access structure related to a user more flexible in expressing. Furthermore, this also allows reducing to the private key size and expanding the universe of attributes. Our paper is organized as follow. We give the necessary background and our definitions of security in Sect. 2. Our proposed ABE scheme and its security are presented in Sect. 3. We then discuss the performance analysis in Sect. 4. Our paper is concluded in Sect. 5.

2 Background

2.1 Decision BDH Assumption

The decision Bilinear Diffie–Hellman (BDH) assumption is defined as follows. Let $a, b, c, z \in \mathbb{Z}_p$ be chosen at random and g be a generator of \mathbb{G} . The decisional BDH assumption is that no probabilistic polynomial-time algorithm \mathcal{B} can distinguish the tuple $(g, g^a, g^b, g^c, e(g, g)^{abc})$ from the tuple $(g, g^a, g^b, g^c, e(g, g)^z)$ with more than a negligible advantage. Note that $e: \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$ is a bilinear map. The advantage of \mathcal{B} is

$$\text{Adv}_{\mathcal{B}} = |\Pr[\mathcal{B}(g, g^a, g^b, g^c, e(g, g)^{abc}) = 0] - \Pr[\mathcal{B}(g, g^a, g^b, g^c, e(g, g)^z) = 0]|$$

Definition 1 We say that the decision (t, ϵ, q_{AS}) —BDH assumption holds in \mathbb{G} if no polynomial time algorithm that runs in time t has an advantage greater than ϵ in solving the decision BDH problem.

2.2 Access Structure

Let $P = (P_1, \dots, P_\ell)$ be a pattern of length ℓ , where each p_i would be an attribute which consists of non-empty string in $\{0, 1\}^*$, or be a wildcard denoted by $*$, or be an empty string ϕ . we let $\overline{W}(P)$ be the index set related to all attributes, and let $W(P)$ be the index set related to all wildcards. Next, we define a set $S(P)$ associated with the pattern $P = (P_1, \dots, P_\ell)$ as $S(P) := \{\mathbb{A} \cup \mathbb{B} : \mathbb{A} = \overline{W}(P) \text{ and } \mathbb{B} \in 2^{W(P)}\}$ basically, the sets in the $S(P)$ will contain the indices in $\overline{W}(P)$. In our context, we need to consider two kinds of patterns, called a Key Pattern (KP) and a

Ciphertext Pattern (CP). We require that the CP associated with the ciphertext matches the KP related to its private key.

We handle an Access Structure(AS) that can be represented by a boolean formula involving AND, OR and threshold gates over key patterns. We note that all threshold gates are easily represented by using AND, OR gates. We denote the set derived from an access structure AS by $S(AS)$, an access structure (as an element) in the set $S(AS)$ by \tilde{as} . That is, $\tilde{as} \in S(AS)$, $\tilde{as} = \cap_{\mathcal{J}(\tilde{as})} iKP_i$ where $\mathcal{J}(\tilde{as})$ is a set containing indices related to KP in \tilde{as} , and $\mathcal{J}(AS) = \cup \mathcal{J}(\tilde{as})$. We denote the set $\{1CP, \dots, mCP\}$ of ciphertext patterns by \mathbb{S} hereafter. We define the matching condition above more formally.

Definition 2 Let $P=(P_1, \dots, P_\ell)$ and $\tilde{P} = (\tilde{P}_1, \dots, \tilde{P}_k)$ be two patterns. We say that P matches \tilde{P} if (1) $P_i = \tilde{P}_k$ for all indices $i \in \overline{W}(P) \cap \overline{W}(\tilde{P})$, and (2) there exists at least one element in $S(P) \cap S(\tilde{P})$, i.e. $S(P) \cap S(\tilde{P}) \neq \phi$.

Definition 3 Let AS be an access structure and let \mathbb{S} be a set of ciphertext patterns. We say that AS matches \mathbb{S} if there is at least one access structure \tilde{as} in the set $S(AS)$ such that all key patterns in the \tilde{as} match some corresponding ciphertext patterns in \mathbb{S} , respectively.

2.3 Attribute-Based Encryption with Wildcards

We describe formal definition Attribute-Based Encryption (ABE) with Wildcards. A ABE scheme consists of four algorithms.

Setup ($1^k, \ell, \mathbf{u}$) takes as input a security parameter 1^k , a maximum length ℓ of patterns, and a maximum number \mathbf{u} of patterns corresponding \tilde{as} or ciphertext. The algorithm outputs public parameters PP and master key MK.

KeyGen(PP, MK, AS) takes as input PP, MK, and a access structure AS. The algorithm outputs a private key d_{AS} for the access structure AS.

Encrypt(PP, M, \mathbb{S}) takes as input PP, a message M, and a set \mathbb{S} of ciphertext patterns. The algorithm outputs a ciphertext CT.

Decrypt($d_{AS}, \mathbb{S}, CT, PP$) takes as input d_{AS} , \mathbb{S} , CT, and PP. The algorithm outputs a message M if AS matches $\hat{\mathbb{S}}$.

2.4 Security-Model of ABE

We define security models that will be applied to ABE scheme in selective-set model.

- **Init.** The adversary outputs an attribute set \mathbb{S}^* to be challenged.

- **Setup.** The challenger takes as input a security parameter 1^k , and then runs this algorithm. It gives the adversary the resulting system parameters PP. It keeps the MK to itself.
- **Phase 1.** The adversary performs a polynomially bounded number of queries. These queries may be asked adaptively:
 - **Extract.** The adversary submits an access structure AS does not match \mathbb{S}^* . The challenger responds by running this algorithm to generate the private key d_{AS} and sends it to the adversary.
- **Challenge.** The adversary chooses two plain texts m_0 and m_1 , on which the adversary wishes to be challenged. The adversary should not have queried for the private key corresponding to AS in phase 1. The challenger chooses randomly a bit b , and encrypts m_b under \mathbb{S}^* , and sends it to the adversary.
- **Phase 2.** It is the same as phase 1.
- **Guess.** The adversary outputs a guess b' and wins the game if $b' = b$.

As usual, the advantage of the adversary in this game is defined as $\Pr[b' = b] - 1/2$.

Definition 4 The ABE scheme is (t, ϵ, q_{AS}) —selectively secure if all polynomial time adversaries who runs in time t , make at most q_{AS} private key queries, have at most a negligible advantage ϵ selective-*sRet* game.

3 Proposed ABE Scheme

We present a ABE scheme which is secure in a selective-set security model.

3.1 Construction

Let \mathbb{G} and \mathbb{G}_T be two cyclic group of prime order p and let $e: \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{G}_T$ be the admissible bilinear map. We assume that for a pattern $P = (P_1, \dots, P_\ell)$ each attribute P_i belong to Z_p^* or $\{*\}$ or $\{\phi\}$ If necessary, number 0 can be used to represent the empty string ϕ . As in [1], we can extend our construction to handle arbitrary attributes in $\{0,1\}^*$ by first hashing each P_i using a collision resistant hash function $H: \{0,1\}^* \rightarrow \mathbb{Z}_p$ Ours construction follows:

- **Setup($1^k, \ell, \mathbf{u}$):** The setup algorithm picks a random generator $g \in \mathbb{G}$ and random $g_i, h_{i,1}, h_{i,2}, \dots, h_{i,\ell_i} \in \mathbb{G}$ where the maximum of ℓ_i is ℓ . It sets $H_{i,k} = (g_i, h_{i,1}, h_{i,2}, \dots, h_{i,\ell_i})$ for $i = 1, \dots, u$. Next, the algorithm chooses random $z_1, z_2 \in \mathbb{Z}_p$ and random $u_1, u_2 \in \mathbb{G}$. It sets, $y_1 = g^{z_1}, y_2 = g^{z_2} \in \mathbb{G}$ and $\psi = e(u_1, g^{z_1}) \cdot e(u_2, g^{z_2}) \in \mathbb{G}_T$. The public key PP (with the description of $(\mathbb{G}, \mathbb{G}_T, p, e)$ and the master key MK are given by $PP = (g, y_1, y_2, H_{i,k}, \psi) \in \mathbb{G}^{2n+3} \times \mathbb{G}_T, \quad MK = (z_1, z_2, u_1, u_2) \in \mathbb{Z}_p^2 \times \mathbb{G}^2$

- **KeyGen(PP, MK, AS):** To generate a private key for an access structure AS, the key generation algorithm first finds the set $S(AS)$ where each element $\tilde{a}_s \in S(AS)$ is represented as an AND-bound access structure. For instance, let $\tilde{a}_s = (1KP \text{ AND } 2KP)$. Then, the algorithm picks random $r_j, t_j \in \mathbb{Z}_p$ for an index $j \in \mathcal{J}(AS)$ and sets $r_j z_2 + t_j z_2 = R \in \mathbb{Z}_p$. For other indexes $i (\neq j) \in \mathcal{J}(AS)$, it selects random $r_i, t_i \in \mathbb{Z}_p$ under equation $r_i z_2 + t_i z_2 = R$. The algorithm computes

$$d_{AS} = \left(u_1 \prod_{i \in \mathcal{J}(AS)} \left(g_i \prod_{k \in \overline{w}(iKP)} h_{i,k}^{P_{i,k}} \right)^{r_i}, u_2 \prod_{i \in \mathcal{J}(AS)} \left(g_i \prod_{k \in \overline{w}(iKP)} h_{i,k}^{P_{i,k}} \right)^{t_i}, g^R, \left\{ h_{i,k}^{r_i} \right\}_{i \in \mathcal{J}(AS), k \in w(iKP)}, \left\{ h_{i,k}^{t_i} \right\}_{i \in \mathcal{J}(AS), k \in w(iKP)} \right)$$

- **Encrypt(PP, M, S):** Assume that $\mathbb{S} = \{iCP\}_{i \in \mathcal{J}(S)}$ is a set of attributes associated with a message M. The encryption algorithm chooses a random $s \in \mathbb{Z}_p$ and constructs a ciphertext as follows:

$$CT = \left(y_1^s, y_2^s, \left\{ \left(g_i \prod_{k \in \overline{w}(iCP)} h_{i,k}^{P_{i,k}} \right)^s \right\}_{i \in \mathcal{J}(S)}, \left\{ h_{i,k}^s \right\}_{i \in \mathcal{J}(S), k \in w(iCP)}, \psi^s M \right)$$

- **Decrypt(d_{AS}, \mathbb{S}, CT):** Let $T = (A, B, \{C_{i,k}\}_{i \in \mathcal{J}(S), k \in w(iCP)}, D)$. The decryption algorithm works as follows:

Check if AS matches \mathbb{S} . If not, output \perp . Otherwise, select a matching access structure $\tilde{a}_s \in S(AS)$. For instance, let $\tilde{a}_s = (1KP \text{ AND } 2KP)$ and $1CP, 2CP \in \mathbb{S}$. Let $d_{AS} = (d_1, d_2, d_3, \{\pi_{i,k}\}, \{\omega_{i,k}\})$ such that $\tilde{d}_1 = d_1 \cdot \prod_{j \in U_1} \pi_{j,K}^{P_{j,K}}$, $\tilde{d}_2 = d_2 \cdot \prod_{U_2} \omega_{j,K}^{P_{j,K}}$, where $U_m = \overline{w}(mCP) / \overline{w}(mKP)$ and output $D \cdot e(\tilde{d}_1, A) \cdot e(\tilde{d}_2, B) / e(d_3, \prod_{i \in \mathcal{J}(S)} C_i)$.

3.2 Security Proof

Theorem 1 The ABE scheme above is (t, ϵ, q_{AS}) —selectively secure assuming the decision (t', ϵ') —BDH assumption holds in \mathbb{G} , where $t' = t$ and $\epsilon' = \epsilon - q_{AS}/p$.

Proof Assume that there exists an adversary \mathcal{A} with non-negligible advantage ϵ in the selective security game. Also, assume that positive integer n is given to \mathcal{A} . We want to construct an algorithm \mathcal{B} which uses \mathcal{A} to solve the decision BDH problem in \mathbb{G} . Given a random decision BDH challenge (g, g^a, g^b, g^c, T) , \mathcal{B} outputs 1 if $T = e(g, g)^{abc}$ and 0 otherwise. \mathcal{B} interacts with \mathcal{A} as follows.

Init \mathcal{A} outputs a set S^* of attributes that it wishes to be challenged on.

Setup \mathcal{B} first finds the sets $W(mCP^*)$ and $\overline{W}(mCP^*)$ for $mCP^* \in S^*$. To generate public parameters PP, \mathcal{B} selects random $z_1, z_2, (\rho_1, \rho_2, \{\tau_1, \dots, \tau_u\}, \{\gamma_{i,1}, \dots, \gamma_{i,k_i}\}) \in Z_p$ for $i = 1, \dots, u$. Recall that k is maximum length of each k_i . It sets $y_1 = g^{z_1}, y_2 = g^{z_2}, \psi = e(g^a, g^b)^{z_1 + z_2} \cdot e(g, g)^{\rho_1 z_1 + \rho_2 z_2}$ and $g_i = g^{\tau_i} \prod_{i,k \in \overline{W}(mCP^*)} (g^b)^{-\gamma_{i,k} P_{i,k}}, \{h_{i,k} = (g^b)^{\gamma_{i,k}}\}_{i,k \in \overline{W}(mCP^*)}, \{h_{i,k} = g^{\gamma_{i,k}}\}_{i,k \notin \overline{W}(mCP^*)}$ for $i = 1, \dots, u$. In the unlikely event that $z_1 + z_2 = 0$, \mathcal{B} tries again with new random exponents. Note that $u_1 = g^{ab} g^{\rho_1}$ and $g^{ab} g^{\rho_2}$ which are unknown to \mathcal{B} . \mathcal{A} is given the public parameters, $PP = (g, y_1, y_2, H_{1,k}, \dots, H_{u,k}, \Psi)$ where $H_{i,k} = (g_i, h_{i,1}, \dots, h_{i,k_i})$. Since all the selected exponents above are random, the PP has an identical distribution to that in the actual construction.

Phase 1 \mathcal{A} issues up to q_{AS} private key queries for access structures. Let AS_i be an i th access structure chosen by \mathcal{A} . From the rule of security model, we know that there is no access structure $\tilde{as} \in S(AS_i)$ such that all key patterns in the \tilde{as} match some corresponding ciphertext patterns in S^* respectively. This means that in an access structure $\tilde{as} \in S(AS_i)$, there is at least one key pattern that does not match any ciphertext pattern in S^* . The main point is that from the key pattern, \mathcal{B} can construct a bunch of private key elements associated with the \tilde{as} .

Let $\Delta = \sum_{i,k \in \overline{W}(mCP^*)} -\gamma_{i,k} P_{i,k} + \sum_{i,k \in \overline{W}(mKP)} \gamma_{i,k} P_{i,k}$. If $j, k \in \mathcal{J}(S^*)$, Δ has not term $\gamma_{i,k} P_{i,k}$ which is not equal to zero. Otherwise, Δ has a term $\gamma_{i,k} P_{i,k}$ where $P_{i,k}$ is randomly chosen by \mathcal{B} and hidden from \mathcal{A} 's view. Thus, in both cases, Δ would be zero with negligible probability $1/p$. If $\Delta = 0$, \mathcal{B} aborts the simulation and outputs a random bit $b' \in \{0, 1\}$. Otherwise, \mathcal{B} picks random exponents $r_j, t_j \in Z_p$. Let $r_j z_1 + t_j z_2 = R \in Z_p$. Next, for $i (i \neq j) \in \mathcal{J}(KP)$, \mathcal{B} selects random $r_i, t_i \in Z_p$ under equation $r_i z_1 + t_i z_2 = R$. Define $\tilde{r}_i = r_i - a/\Delta, \tilde{t}_i = t_i - a/\Delta$ for all $i \in \mathcal{J}(KP)$. Then, \mathcal{B} has that $\tilde{r}_i z_1 + \tilde{t}_i z_2 = R - a(z_1 + z_2)/\Delta$ for all $i \in \mathcal{J}(KP)$. Now, \mathcal{B} can generate $K_{S(AS)}$ as

$$d_{AS} = \begin{pmatrix} g^{\rho_1} \prod_{i,k \in \overline{W}(mKP)} g^{\tau_i r_i (g^a)^{\frac{-\tau_i (g^b)^{\Delta r_i}}{\Delta}} g^{\gamma_{i,k} P_{i,k} r_i} \cdot (g^a)^{\frac{\sum_{i,k \in \{\overline{W}(mKP) - \overline{W}(mCP^*)\}} -\gamma_{i,k} P_{i,k}}{\Delta}} \\ g^{\rho_2} \prod_{i,k \in \overline{W}(mKP)} g^{\tau_i r_i (g^a)^{\frac{-\tau_i (g^b)^{\Delta t_i}}{\Delta}} g^{\gamma_{i,k} P_{i,k} t_i} \cdot (g^a)^{\frac{\sum_{i,k \in \{\overline{W}(mKP) - \overline{W}(mCP^*)\}} -\gamma_{i,k} P_{i,k}}{\Delta}} \\ g^R (g^a)^{\frac{-z_1 + z_2}{\Delta}} \cdot \left\{ g^{\gamma_{i,k} \gamma_i (g^a)^{-\gamma_{i,k}/\Delta}} \right\}_{i,k \in W(KP)} \end{pmatrix}$$

The validity of $K_{S(AS)}$ can be verified as in security proof of [2].

Challenge \mathcal{A} outputs two messages $m_0, m_1 \in \mathbb{G}_T$. \mathcal{B} picks a random bit $b \in \{0, 1\}$ and gives \mathcal{A} a challenge ciphertext

$$CT^* = \left((g^c)^{z_1}, (g^c)^{z_2}, \{(g^c)^{\tau_i}\}_{i \in \overline{W}(mCP^*)}, \{(g^c)^{\gamma_{i,k}}\}_{i \in W(mCP^*)}, T^{z_1 + z_2} \cdot e(g^c, g)^{\rho_1 z_1 + \rho_2 z_2} \cdot m_b \right)$$

Note that $z_1 + z_2 \neq 0$. If $T = e(g, g)^{abc}$, then CT^* is a valid challenge ciphertext under the randomness $s = c$. Otherwise, $T^{z_1+z_2} \cdot e(g^c, g)^{\rho_1 z_1 + \rho_2 z_2}$ becomes a random element in \mathbb{G}_T and thus independent of the bit b from \mathcal{A} 's view.

Phase 2 \mathcal{A} continues to ask queries not issued in Phase 1. \mathcal{B} responds as before.

Guess \mathcal{A} outputs a guess $b' \in \{0, 1\}$ in response to the challenge ciphertext. If $b' = b$, \mathcal{B} outputs 1, indicating that $T = e(g, g)^{abc}$. Otherwise, \mathcal{B} outputs 0, indicating that T is random in \mathbb{G}_T .

Let abort be the event that $\Delta = 0$ in query phase. Note that $\Pr[\text{abort}] = q_{AS}/p$. If $T = e(g, g)^{abc}$, \mathcal{A} is given the valid ciphertext. Then, \mathcal{B} 's advantage in the decision BDH game is taken from ε unless the event abort does not occur. Thus, \mathcal{B} 's advantage is $\varepsilon - q_{AS}/p$. On the other hand, if T is random in \mathbb{G}_T , then \mathcal{A} 's advantage is $1/2$. Thus, \mathcal{B} 's advantage is exactly $\varepsilon - q_{AS}/p$.

4 Performance Comparison

In Table 1, we give a performance comparison of previous ABE schemes and their properties. Table 1 presents that our ABE scheme requires only $0(1)$ pairings for decryption, regardless of $|S|$. We must note that our scheme requires the number of wildcards exponentiation computation in decryption. Recall that our scheme in Sect. 2.2 permits AND/OR operations for representing an access structure, but these are enough to express any threshold operation by using DNF formulas.

d is the pre-determined value of representing the number of minimal set overlap. t is the number of attributes in an access structure. $|S|$ is the size of attribute set. ℓ is the number of attributes appeared in an access structure. n is the maximal bound for $|S|$. φ is maximum size for repetition of attribute label per key. m is the number of patterns. ρ is the size of vector.

Table 1 Comparison of measured roughness data, machining center

	CT size	Private key size	Number of pairings	Possible operations	Security	Assumption
SW05 [7]	$\mathcal{O}(S)$	$\mathcal{O}(d)$	$\mathcal{O}(d)$	Threshold	Selective	DBDH
GPSW06 [3]	$\mathcal{O}(S)$	$\mathcal{O}(\ell)$	$\mathcal{O}(S)$	AND/OR/threshold	Selective	DBDH
OSW07 [6]	$\mathcal{O}(S)$	$\mathcal{O}(n)$	$\mathcal{O}(t)$	AND/OR/threshold/Not	Selective	DBDH
LSW10 [4]	$\mathcal{O}(S)$	$\mathcal{O}(t)$	$\mathcal{O}(t)$	AND/OR/threshold/Not	Selective	q-MEBDH
OT10 [8]	$\mathcal{O}(S \cdot \varphi)$	$\mathcal{O}(t)$	$\mathcal{O}(t \cdot \varphi)$	AND/OR/threshold/Not	Full	DLIN
ZH10 [9]	$\mathcal{O}(1)$	$\mathcal{O}(t)$	$\mathcal{O}(S)$	AND/Not/wildcard	Selective	q-DBDHE
Ours	$\mathcal{O}(m)$	$\mathcal{O}(m)$	$\mathcal{O}(1)$	AND/OR/threshold/Wildcard	Selective	DBDH

5 Conclusion

In this paper, we proposed attribute-based encryption scheme which is secure assuming the DBDH assumption holds in selective model. Our ABE construction was able to handle any access structures that can be represented by a formula consisting of wildcard, AND, OR and threshold gates. A natural question is to build a new ABE scheme that additionally provides NOT gate to represent negative constraint, with improved decryption cost. Another problem is to improve our ABE scheme to support a revocation mechanism. When considering the growing popularity of commercial broadcast systems such as IPTV, it is desirable to obtain the revocation scheme which works in the attribute-based setting.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012-0008697).

References

1. Boneh V, Boyen X (2004) Efficient selective-ID secure identity-based encryption without random oracles. In: Eurocrypt'04, Lecture notes in computer science, vol 3027, pp 223–238
2. Boneh D, Franklin M (2003) Identity-based encryption from the weil pairing. SIAM J Comput 32(3): 586–615. Earlier version in Crypto'01, Lecture notes in computer science, vol 2139, pp 213–229, 2001
3. Goyal V, Pandey O, Sahai A, Waters B (2006) Attribute-based encryption for fine-grained access control of encrypted data. In: ACM CCS'06, pp 89–98
4. Lewko A, Sahai A, Waters B (2010) Revocation systems with very small private keys. In: IEEE symposium on security and privacy (S&P)
5. OMA: DRM Specification-OMA-TS-DRM_DRM-V2_1-20081106-A (2008). http://www.openmobilealliance.org/Technical/release_program/drm_v2_1.aspx
6. Ostrovsky R, Sahai A, Waters B (2007) Attribute-based encryption with non-monotonic access structures. In: ACM CCS'07, pp 195–203
7. Sahai A, Waters B (2005) Fuzzy identity-based encryption. In: Eurocrypt'05, Lecture notes in computer science, vol 3494, pp 457–473
8. Okamoto T, Takashima K (2010) Fully secure functional encryption with general relations from the decisional linear assumption. In: CRYPTO'10, Lecture notes in computer science, vol 6223, pp 191–208
9. Zhou Z, Huang D (2010) On efficient ciphertext-policy attribute based encryption and broadcast encryption. In: ACM CCS'10

Part V
IT Convergence Applications

Sensibility Extraction for Bicycle Design Using RFID Tag-Attached Crayons

Ho-Il Jung, Seung-Jin Lee, Jeong-Hoon Kang, Min-Hyun Kim,
Jong-Wan Kim, Bo-Hyun Lee, Eun-Young Cho
and Kyung-Yong Chung

Abstract Providing sensibility design using information convergence technology is an important factor in product service strategies. It is possible to ensure future competitiveness in bicycle industries by developing and specializing highly sensibility bicycle design. The necessity arises of creating a sensibility engineering approach to develop products that appeal to a wide variety of customers by stimulating their senses and creating emotional satisfaction. In this paper, we proposed a bicycle design recommendation using RFID tag-attached crayons. The proposed method obtains visual appeal using these RFID tag-attached crayons. Associative color patterns are analyzed using data mining, which extracts conceptual information from the collected data that is not easily exposed. The association can be determined as crayon colors are presented in a specific transaction, and different crayon colors are presented in the same transaction. The association rule represents a strong relationship between color sets. Designing frames, saddles, pedals, wheel sets, tires, cranks, and other parts of a bicycle based on visual sensibility represents a final shape by advancing through an application to a virtual model. By providing bicycle design adapted to one's own design, it reduces cost and time and makes it possible to apply it to the desired styles.

H.-I. Jung · S.-J. Lee

IS Lab, School of Computer Information Engineering, Sangji University,
83 Sangjidae-gil, Wonju-si, Gangwon-do, Korea

J.-H. Kang · M.-H. Kim · J.-W. Kim · B.-H. Lee · E.-Y. Cho

Gangwon Science High School, 2242 Chiang-ro, Wonju-si, Gangwon-do, Korea

K.-Y. Chung (✉)

School of Computer Information Engineering, Sangji University,

83 Sangjidae-gil, Wonju-si, Gangwon-do, Korea

e-mail: dragonhci@hanmail.net

Keywords Sensibility engineering · Bicycle design · Recommendation · RFID tags

1 Introduction

Sensibility, diversity, and differentiation have been presented according to developments in IT convergence technologies and different life styles. This leads to highlighting visual and external appeal and motivating support in sensibility engineering design. A trend of consumption of aesthetically appealing bicycles brings a change in usability due to considering the aesthetic aspects in the efficiency, comfort, individuality, and economy of products rather than as a function of transportation. Consumers of bicycles are getting younger, so there is a new wave of considering personality in consumption. Considering personality in bicycle purchasing leads to a new trend in markets accompanying personal and particular figures [1, 2]. Design is an important factor considered in purchasing a bicycle. In the Fixie bikes of IGUANA Bike Ltd. (www.iguanabike.jp) that are popular among youngsters recently, it is possible to directly present frame colors and decals in order to demonstrate their own personalities. Types of bicycles are varied not only according to functional factors, but also according to owner taste, and environmental and cultural factors. Also, it is necessary to investigate important factors contributing to determine bicycle shapes and to consider these factors in a design process for designing purpose-devoted bicycles [3, 4].

The core elements in providing bicycle design support services using sensibility are how the sensibility is expressed and how the images expressed are specifically presented to a device. Also, the specific type of interfacing device should also be considered. As sensibility is difficult to understand and quantitatively measure, it is not easy to recognize such abstract expression and design, because it usually depends on limited adjective words. However, it is an essential process that analyzes the sensibility and recognizes its pattern and design. Thus, a relationship between humans and computers that connects design elements and patterns to preference in order to develop a system is required. In addition, a user interface that performs such sensibility as a specific shape through connective middleware is needed. For developing a sensibility bicycle design, a relationship that connects user preferences and patterns to design elements is needed. Collaborative filtering connects real and virtual spaces for presenting sensibility in virtual realities as a type of information, and provides an intellectual service centered on users based on this information [1, 5]. It is expected that the next generation of convergence technology will lead to huge changes in all economic and industrial fields. Also, it will play a role in producing demand in industrial markets as a global leader.



Fig. 1 Designing a metro bike bicycle (www.metrobike.co.kr)

2 Related Research of Bicycle Design

Importance of design in bicycle products has rapidly increased. Performance and reliability more than a specific level in bicycle products have been considered as basic conditions in entering markets, and design, usability, and subjective satisfaction have been recognized as successful factors in these products. A bicycle maker, Samchuly Ltd., introduces its Metro Bike that is produced by an order-made method in which users can select their own colors and designs. Users can select their own color for parts and bicycles can be assembled according to these selections in the factory. Colors can be selected for 11 parts like the frame, saddle, pedal, wheel set, tire, crank, and so on. Colors like red, pink, white, and dark blue are provided to these parts. It is possible to select the design and colors for such an order-made bicycle in the homepage of Samchuly at the Metro Bike section. Figure 1 shows the web site used to design a Metro Bike.

Shimano Inc. in Japan is a company that produces outdoor products and bicycle parts. They integrate all their products from low price to high end, and that makes it possible for them to establish a huge infrastructure in bicycle parts in world markets.

3 Extraction of Visual Sensibility using RFID Tag-Attached Crayons

3.1 RFID Tag-Attached Crayons

For collecting visual sensibility, SkyeModue 13.56 MHz RFID sensors [6] were used to RFID tag-attached crayons (18 colors). This consists of the SkyeModue M9CF, host interface board, PCB HF/Flex antenna, 9 V power adaptor, RS-232

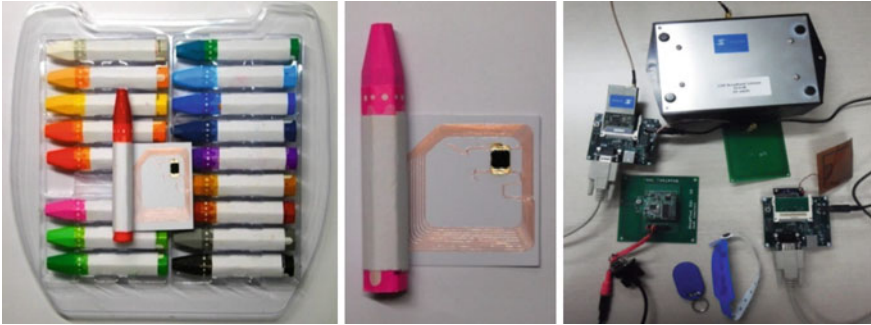


Fig. 2 RFID tag-attached crayons (18 colors), RFID leader using a UHF band antenna

cable, and RFID tag. Figure 2 shows the RFID tag-attached crayons (18 colors) for obtaining visual sensibility and RFID leader using a UHF band antenna. A cover was used to attach RFID tags to crayons, which can easily be purchased at any stationery store. Wireless communication using RFID tags makes it possible to minimize unnecessary wires and to avoid the awareness of attaching it in general use. RFID tags attached to crayons (18 colors) have identification numbers and colors that can be read and identified through an RFID reader. An extra UHF band antenna is used to maintain stable receiving conditions. As RFID in the UHF band has its own specific information, it consists of a microchip and an antenna that transmits the information. It uses a backscattering method that radiates radio waves from a reader to RFID tags, and RFID tags retransmit the specific information to the reader. This method is able to perform mid-distance, long-distance and high-speed signal transmissions.

Regarding the initialization of the RFID tag-attached crayons for an RFID reader using a UHF band antenna, a command for initializing crayons is to be executed and to wait while tags are recognized in the reader [7]. As a tag is recognized in the reader, the recognized tag is searched for in the tag list registered in a server. After searching for the ID of the recognized tag, it is considered that the same crayon color is used. Otherwise, the ID of the recognized tag is stored in the tag list and is configured as the present crayon color. Thus, it is possible to solve the problem of accumulating errors because it measures the present crayon color.

3.2 Process of Associative Crayon Colors Pattern

As an RFID tag is detected in a reader while crayons are used, the recognized color is determined as the present color and its use time is recorded. The data collected for determining crayon color and use time is transmitted to a server. Then, a preferred color list is extracted using the collected data in the server and used to find visual sensibility. The preferred color list can be recognized by the use time of a crayon

with an RFID tag. A long use time represents a preferred color and a short use time is counted as a non-preferred color. The threshold value in a time window, which is determined by the use time, was configured in 10 min increments. That is, while there are no requests to a server for 10 min, it is assumed that the time window is terminated. Table 1 shows the crayon color transaction stored in a server.

A transaction number represents an RFID tag that recognizes a crayon color and the extracted crayon color is used to configure candidate and high-frequency item sets. Crayon color transactions (18 colors) are presented as R_1, R_2, \dots, R_i ($1 \leq i \leq 18$) and include color information on these 18 colors. The color information consists of R_1 (Red), R_2 (Orange), R_3 (Yellow Orange), R_4 (Pink), R_5 (Yellow), R_6 (Gray), R_7 (Pale Orange), R_8 (Sky Blue), R_9 (Cobalt Blue), R_{10} (Blue Green), R_{11} (Green), R_{12} (Yellow Green), R_{13} (Black), R_{14} (Brown), R_{15} (Purple), R_{16} (Yellow Ochre), R_{17} (Prussian Blue), R_{18} (White). A session transaction with an item that uses the crayon color information is configured to extract associative patterns. As several users use it simultaneously, session IDs are to be allocated for each user. That is, users create sessions whenever they use crayons and then delete the sessions. The session transaction is needed to track such information. Also, the time window is used to determine session transactions. It is possible to find association patterns based on crayon colors used according to the sequence of recognizing RFID tags in the time window.

The associative colors patterns determine the relationship between crayon colors presented in a transaction. It is also used to consider the relationship between before and after presenting crayon colors in a crayon color transaction. It finds a maximum sequence in all sequences that satisfies the minimum support in the given transactions, in which the sequences are a list of transactions arranged by transaction times. High-frequency and candidate sets exhibit crayon colors presented in crayon color transactions. In this case, the mining is applied to associative crayon colors, which consist of n crayon colors, through n searches [8, 9]. In each processing step, low-frequency associative crayon colors are neglected in the mining. In the results of the mining, crayon colors are presented as a set of associative color patterns. For mining the crayon colors extracted in the crayon color transactions noted in Table 1, the process that infers association patterns using the Apriori algorithm is as follows.

Table 1 Crayon color transactions

TID	Set
T_1	$R_2, R_6, R_{10}, R_{11}, R_{13}, R_{18}$
T_2	$R_1, R_2, R_4, R_5, R_{10}$
T_3	$R_3, R_7, R_8, R_{12}, R_{14}$
T_4	$R_1, R_2, R_4, R_5, R_{17}$
T_5	$R_3, R_7, R_8, R_{12}, R_{15}$
T_6	R_1, R_3, R_7, R_9
T_7	$R_2, R_6, R_{11}, R_{13}, R_{16}$
T_8	$R_1, R_3, R_7, R_9, R_{10}$

The Apriori algorithm configures a candidate set (CC1) in the first step and can search databases to verify Support and Confidence. Also, it can configure a high-frequency set (LC1) [9]. The generated high-frequency set is a high-frequency sequence and the high-frequency sequence itself is a desired associative pattern. Here, while the crayon colors included in the associative patterns satisfy the minimum support and it represents the crayon colors used during a specific session. In the same way, the second and third steps in the Apriori algorithm configure CC2 and LC2, and CC3 and LC3 respectively [10]. The associative crayon colors set of LC3 is extracted as $\{R_1, R_2, R_4, R_5\}$, $\{R_1, R_3, R_7, R_9\}$, $\{R_2, R_6, R_{11}, R_{13}\}$, and $\{R_3, R_7, R_8, R_{12}\}$. Therefore, associative color patterns using the mining are $\{\text{Red, Orange, Pink, Yellow}\}$, $\{\text{Red, Yellow Orange, Pale Orange, Cobalt Blue}\}$, $\{\text{Orange, Gray, Green, Black}\}$, and $\{\text{Yellow Orange, Pale Orange, Green, Yellow Green}\}$. The extracted associative color patterns are applied to frames in bicycle design for recommendation.

4 Experimentation Evaluation

Data of visual sensibility was collected from July to August 2012 for two months from high school students who are in Kangwon Science High School and university students. The collected visual sensibility data was stored in a file server located at the IS Lab., Sangji university. Figure 3 shows the execution display of measuring visual sensibility using RFID tag-attached crayons and bicycle pictures drawn by experiment subjects.

For increasing attendance before measuring sensibility, the objective and process of the experiment was presented for 10 min. Experimental subjects drew bicycle pictures on a B4 size piece of paper using RFID tag-attached crayons. Also, they drew bicycle pictures using their preferred crayon colors for measuring visual sensibility in order to verify a correlation in preferred colors according to changes in the activity of an autonomic nervous system, which consists of



Fig. 3 Measuring visual sensibility using RFID tag-attached crayons

sympathetic and parasympathetic nerves. As the pose of experimental subjects affects the results of measuring visual sensibility, comfortable poses were introduced. Regarding the experimental environment of measuring visual sensibility, noise was cut off and temperature was maintained between 20–25 °C.

5 Conclusions

Recent design has been focused on new research fields including sensibility design, sports, and IT convergence technologies in context recognition in addition to design itself. In this paper, we proposed a type of bicycle design recommendation using RFID tag-attached crayons. For collecting visual sensibility, we developed 18 colors of RFID tag-attached crayons. Wireless communication using RFID tags minimizes unnecessary wires and avoids the awareness of attaching it in general uses. Regarding the find of associative color patterns from crayon color transactions using the Apriori algorithm, the associative color patterns determine the relationship between crayon colors presented in a transaction. By providing bicycle design corresponding with visual sensibility, it represents the advantages of reducing time and cost and of supporting easy bicycle design. In addition, it can be effectively used to recommend the design of bicycle frames, saddles, pedals, wheel sets, tires, and cranks based on sensibility in order to present one's own tastes, attitudes, and propensities to design. The proposed method can be used as various design knowledge information in a creative IT convergence age. Also, it informs huge changes not only in customers but also in business transactions.

Acknowledgments This work was supported by the Korea Foundation for the Advancement of Science & Creativity (KOFAC) grant funded by the Korean Government (MEST).

References

1. Lee H (2011) An research of a recently bicycle decoration design analysis and development. *J Korea Digit Des* 11(1):613–623
2. Jung HI, Jo SM, Rim KW, Lee JH, Chung KY (2012) Ergonomics automotive design recommendation using image based collaborative filtering. In: *Proceeding of the international conference on information science and applications*. IEEE Computer Society, pp 211–216
3. Shinohara A, Shimizu Y, Sakamoto K (1996) *Introduction to Kansei engineering*. Kumbuk Pub, Kyoto, pp 45–69
4. Kwan OK, Na YJ, Kim HE (2000) *Fashion and sensibility science*. Kyomunsa, Seoul, pp 25–30
5. Jung KY, Na YJ, Lee JH (2003) Creating user-adapted design recommender system through collaborative filtering and content based filtering. *EPIA'03, LNAI 2902*, Springer, pp 204–208
6. Skyetek, <http://www.skyetek.com/>

7. Chung KY (2008) Recommendation using context awareness based information filtering in smart home. *J Korea Contents Assoc* 8(7):17–25
8. Agrawal R, Srikant R (1994) Fast Algorithms for mining association rules, In: *Proceeding of the 20th VLDB conference*, Santiago, Chile, pp 487–499
9. Chung KY, Lee D, Kim KJ (2011) Categorization for grouping associative items data mining in item-based collaborative filtering. *Multimedia Tools Appl*, Published online
10. Jung KY (2006) Associative neighborhood according to representative attributes for performing collaborative filtering. *ICIC'06, LNCIS 344*, Springer, pp 839–844

Efficient Handover Schemes for Railroad Communications in 4G Mobile Networks

Ronny Yongho Kim and Baik Kim

Abstract Because of its convenience and environmental benefits, public transportation like railroad transportation is getting more attention. Railroad communication has its unique characteristics because of its unique network configuration. This paper proposes novel handover schemes for railroad communications. Railroad communication network configuration is explained in order to utilize favorable characteristics of railroad communications. In this paper, railroad handover zone is newly proposed for efficient handover and group handover is also proposed in order to reduce handover control overhead. With little modification, the proposed schemes can be applied to machine-to-machine vehicular communications where devices in a moving vehicle communicate with a server in the network.

Keywords Handover · Group handover · Railroad communications · WiMAX · IEEE 802.16 m

1 Introduction

Global warming is becoming more and more serious problem. In order to reduce carbon dioxide emission, public transportation is highly recommended as a green transportation. Because of its convenience and environmental benefits, railroad transportation is getting more attention. Railroad communication has its unique characteristics because of its unique network configuration. If such unique

R. Y. Kim · B. Kim (✉)
Department of Railroad Electrical and Electronics Engineering,
Korea National University of Transportation, 157 Cheoldo Parkmulgwan-ro Uiwang,
Gyeonggi, South Korea
e-mail: whitek@ut.ac.kr

characteristics are properly utilized for the railroad communications, very efficient communication schemes can be designed. Emerging broadband wireless air interface specification such as IEEE 802.16m [1, 2], which provides enhanced link layer amendment to the legacy IEEE 802.16 system [3, 4], is designed to meet and in many cases exceed IMT-Advanced requirements [5]. In this paper, novel group handover schemes for railroad communications in IEEE 802.16m based next-generation WiMAX [6] are presented. With little modification, the proposed schemes can be applied to machine-to-machine communications where devices in a moving vehicle communicate with a server in the network.

The remaining part of the paper is organized as follows. In Sect. 2, railroad communication network configuration is described in order to utilize favorable characteristics of railroad communications. In Sect. 3, novel railroad handover zone concept and railroad handover procedures are proposed. In Sect. 4, novel group handover schemes: group single-carrier handover, group multi-carrier handover are proposed for railroad communications. Finally, Sect. 4 provides concluding remarks.

2 Railroad Communication Network Configuration

Railroad communication network has its unique network configuration since train moves along with fixed railroad. Typical railroad communication network configuration is shown in Fig. 1. A train which consists of multiple train cars moves along a train track. Since railroad communication network is a linear network and train's movement is pre-scheduled, communication of train passengers or crews can be easily managed. Railroad communication can be categorized into two classes: railroad control communication, railroad user communication. Railroad control communication is mainly for the purpose of train operation including voice communication between trains and command center and between crews in a train. Railroad user communication is to provide railroad passengers with communication

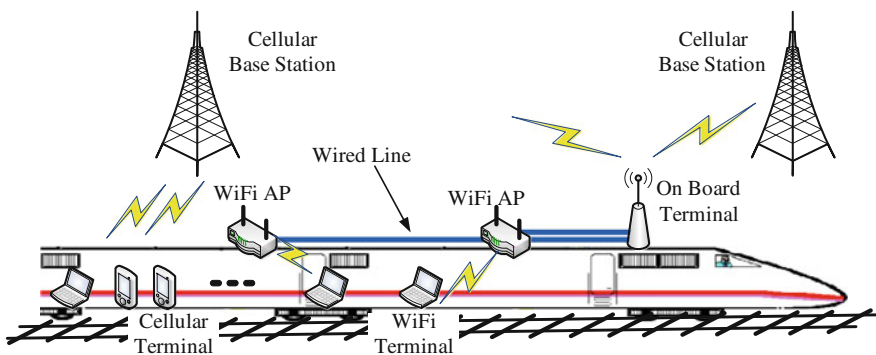


Fig. 1 Railroad communication network configuration

services. As shown in Fig. 1, railroad users are able to communicate either directly with base station or indirectly via train communication devices such as WiFi AP and on board terminal. The common way for train passengers to access data is using WiFi installed in railroad cars.

3 Proposed Handover Scheme for High Speed Train

Very fast handover scheme is required in order to use 4G mobile communication technology for a high speed train. In 4G mobile communication, such fast handover requirement for railroad communications is not fully considered. As introduced in Sect. 2, since railroad communication network has its unique network configuration due to scheduled movement of train along fixed railroad, a simple and efficient handover solution can be derived. As shown in Fig. 2, along fixed railroad, multiple BSs can be grouped to form a “railroad HO zone”. In a railroad HO zone, depending on the location of the BS, two kinds of BS are configured: *Zone Ingress BS* and *Zone Egress BS*. Upon an MS’s handover to a *Zone Egress BS*

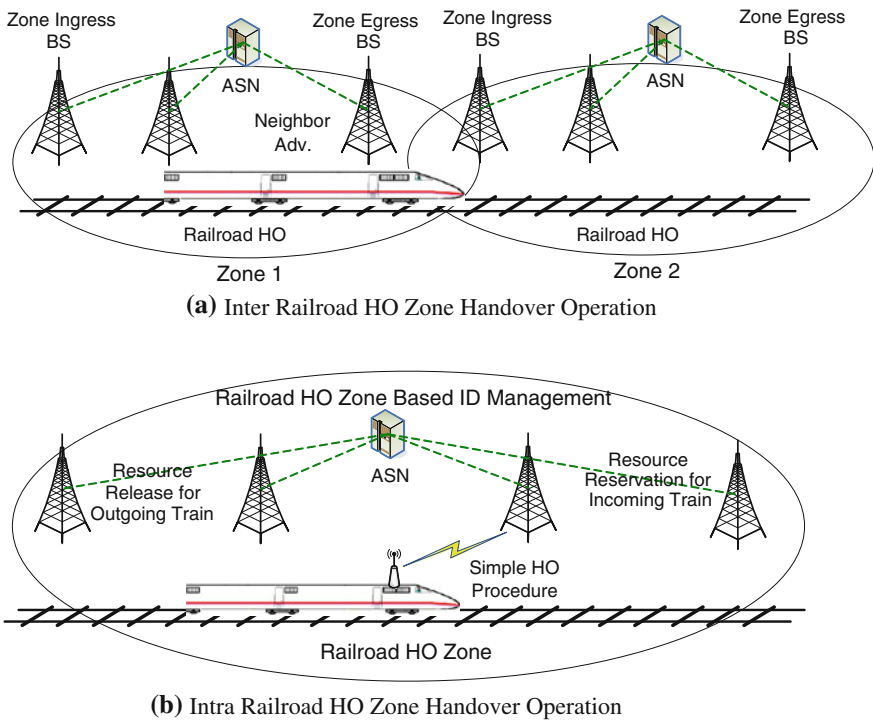


Fig. 2 A proposed railroad zone based handover system architecture. **a** Inter railroad HO zone handover operation. **b** Intra railroad HO zone handover operation

BS, the Zone Egress BS is in charge of notifying the next zone's Zone Ingress BS of the MS's handover. The next zone's Ingress BS can be prepared for the MS's handover by reserving resources. Within a railroad HO zone, since BSs is able to use same resources related to the MS, handover procedure can be simplified substantially. Proposed railroad zone based handover procedures for inter railroad HO zone handover and intra railroad HO zone handover are shown in Fig. 3. In case of inter railroad HO zone handover, BSs' information of the next railroad handover zone is delivered to the MS by the current handover zone's Egress BS. Using the received information, the MS is able to find BSs in the next handover zone and is able to perform very simple handover procedure by just performing ranging. Inter railroad HO zone handover is even simpler since all the BSs in the zone share the MS's information. By simply performing ranging procedure which setup new data path through the current BS, the MS is able to handover to the new BS within the zone.

4 Proposed Group Handover Schemes

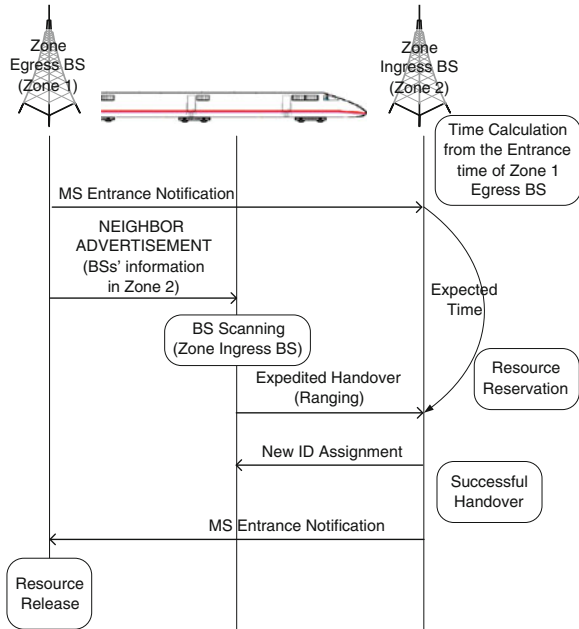
In this section, novel group handover schemes for single carrier and multi-carrier are proposed. Since MSs in a train move together along a train track, MSs can form a handover group and handover of MSs can be handled as a group which can save handover control overhead substantially. Figure 4 shows the concept of the proposed group handover. Since signal quality is not very good at cell edge and handover is typically performed due to movement across cells, resources required to transmit handover related control messages are larger than normal control message transmission. Therefore, the amount of resources saved through the proposed group handover scheme is more substantial.

4.1 Group Single Carrier Handover

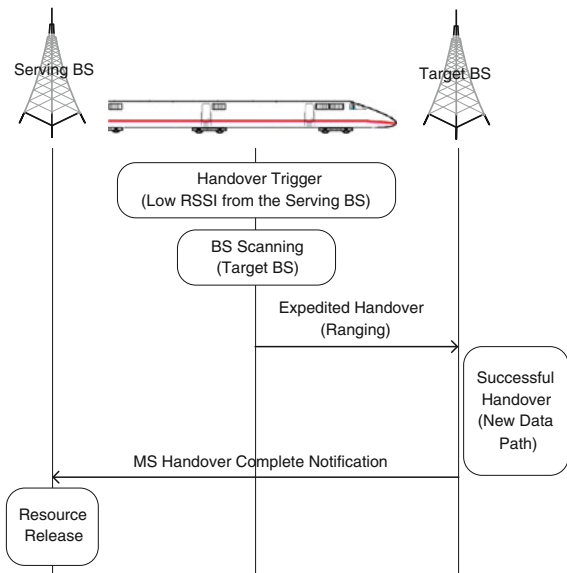
Figure 5a shows a proposed group single-carrier handover scheme. In order to perform a handover procedure, as explained in the previous section, an MS needs to store neighbor BSs' information advertised through the neighbor advertisement control message. Upon meeting the handover trigger conditions, i.e. low received signal strength (RSSI) or low signal to noise ratio (SNR), MSs starts scanning neighbor BSs using the stored information of neighbor BSs. With the proposed group handover scheme, MSs in a handover group are not required to listen to neighbor BS information and care about handover trigger conditions.

On board terminal as a handover group header performs a group handover procedures. Precondition of the proposed group handover schemes is that MSs in a train are grouped as proposed in [7]. During the group formation procedure, a group identifier (ID) is assigned to the group. After neighbor BSs scanning, the on

Fig. 3 A proposed railroad zone based handover procedures. **a** Inter railroad HO zone handover. **b** Intra railroad HO zone handover



(a) Inter Railroad HO Zone Handover



(b) Intra Railroad HO Zone Handover

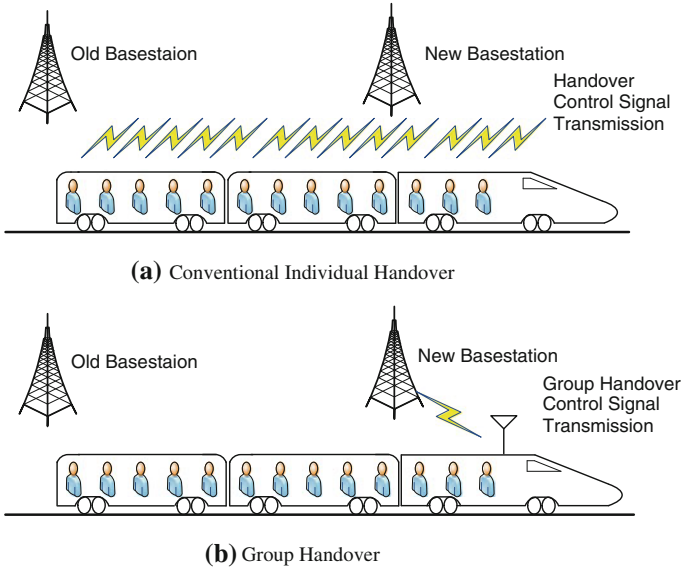


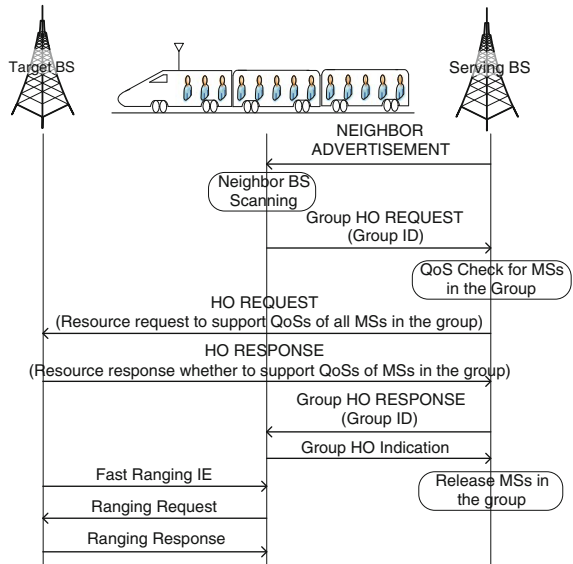
Fig. 4 A single-carrier group handover scheme. **a** Conventional individual handover. **b** Group handover

board terminal selects the candidate target BSs to handover and transmits “Group HO REQUEST” message including its group ID and BS IDs of the candidate target BSs to the serving BS. Upon receiving “Group HO REQUEST” message, the serving BS can retrieve MSs’ information in the group. Based on the retrieved information, the serving BS is able to calculate required resources for MSs in the group. The serving BS negotiates handover conditions including required resources with the candidate target BSs. After finding the best appropriate target BS among the candidate target BSs, the serving BS transmits “Group HO RESPONSE” message to the on board terminal. After receiving “Group HO Indication” message from the on board terminal, the serving BS releases MSs in the group. If the resource holding timer in the received “Group HO Indication” is set with non zero value, the serving BS will release the MSs in the group either after timer expiry or upon receiving handover success notification from the target BS. On board terminal completes the group handover procedure with the target BS with the group handover ranging procedure.

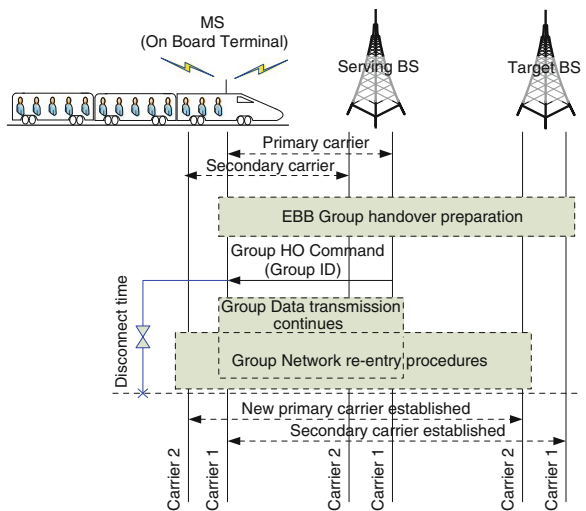
4.2 Group Multi-Carrier Handover

Figure 5b shows a proposed group multi-carrier handover scheme. If multi-carrier capability is supported by the network and the on board terminal, group handover can be performed in amore efficient way than the single-carrier group handover.

Fig. 5 A proposed group handover schemes. **a** A single-carrier group handover scheme. **b** A multi-carrier group handover scheme



(a) A single-carrier group handover scheme



(b) A multi-carrier group handover scheme

Similar to the single-carrier group handover scheme, the key in the multi-carrier group handover is Group ID in order to take care of all member MSs in the same group. Since the serving BS knows which MSs are in the same group with the on board terminal during the group formation procedure, the serving BS only need to communicate with the on board terminal for handover. Since the on board terminal

supports multi-carrier operation, the MSs communication via the on board terminal is not affected at all regardless of MSs' multi-carrier capability. Other procedures are similar to the single-carrier group handover scheme.

5 Conclusion

In this paper, novel handover schemes for railroad communications in 4G mobile networks are proposed. The proposed handover schemes are railroad handover zone based handover scheme and group handover scheme. By utilizing the proposed schemes, MSs on the train are able to perform very simple and expedited handover procedure and handover control overhead can be reduced substantially. With little modification, the proposed schemes can be applied to machine-to-machine vehicular communications where devices in a moving vehicle communicate with a server in the network.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A1014610).

References

1. IEEE 802.16m-09/0034r2 (2009) IEEE 802.16m System Description Document, Sept 2009
2. IEEE standard for local and metropolitan area networks part 16: air interface for broadband wireless access systems amendment 3: advanced air interface. IEEE Std 802.16m-2011, 12 May 2011
3. IEEE standard for local and metropolitan area networks part 16: air interface for broadband wireless access systems, amendment 2: physical and medium access control layers for combined fixed and mobile operation in licensed bands and corrigendum 1. IEEE Std 802.16e-2005, 28 Feb 2006
4. IEEE standard for local and metropolitan area networks part 16: air interface for broadband wireless access systems. IEEE Std 802.16-2009 (Revision of IEEE Std 802.16-2004), 29 May 2009
5. ITU-R M.2134 (2008), Requirements Related to Technical System Performance for IMT-Advanced Radio Interface(s) [IMT.TECH], draft new report, Nov 2008
6. WiMAX end-to-end network system architecture. Stage 3: detailed protocols and procedures, WiMAX Forum, Aug 2006
7. Kim RY (2011) Efficient wireless communications schemes for machine to machine communications. *Commun Comput Inf Sci* 181(3):313–323

Cyber Threat Prediction Model Using Security Monitoring System Event

Neo Park and Won Hyung Park

Abstract There was a large scale of DDoS(Distributed Denial of Service) attacks mostly targeted at Korean government web sites and cooperations on March 4, 2010 (3.4 DDoS attack) after 7.7 DDoS on July 7, 2009 in South Korea. To detect and respond to them, malwares must first be detected by security monitoring system. In particular, availability of a method to detect and predict such malwares in advance will lead to preventing security incidents. This study will propose a method of prediction based on security monitoring event in Security Monitoring system and a functional configuration to realize the method and will assess the prediction model based on security monitoring events proposed through a test consisting of the stages of learning, prediction and evaluation.

Keywords Cyber threat · Security monitoring event · Denial of service

1 Introduction

The cyber threat prediction technologies discussed and studied so far are mostly of predicting changes of numerical values after the unit time of time-series data used in a near future or for input by using numerical data collected on network or under

N. Park

Department of Ubiquitous IT, Far East University, Wangjang-ri, Gangok-myeon, Eumseong-gun, Chungbuk 369-700, South Korea
e-mail: neopark@kdu.ac.kr

W. H. Park (✉)

Department of Information Management, Far East University, Wangjang-ri, Gangok-myeon, Eumseong-gun, Chungbuk 369-700, South Korea
e-mail: whpark@kdu.ac.kr

individual system with a statistics-based prediction model. However, the numerical results can only be used as reference data and it is in fact almost impossible in reality to decide which handling measures must be taken against which types of threats in details. The prediction technologies based on security monitoring event are intended at predicting detailed information relating to detailed event occurrence and intrusion in advance by overcoming the problems of statistical prediction model. Due to technological difficulties and limitations, only a theoretical model has been suggested with validity evaluation through a test. Since there are yet to be cases of using the technologies in detail, this study focuses on validating and evaluating effectiveness of the prediction model based on security monitoring event.

2 Related Work

NIST (National Institute of Science and Technology) released the “Attack Graph” to assist in predicting the risk of hackers to destroy computer system security and in identifying the most vulnerable resources [1]. NIST anticipates that the attack graph will enable IT administrators to identify vulnerabilities to be handled in order to protect precious data by analyzing and designating possibilities of all routes to be used by hackers in intruding computer system. Network attack graph provides information to security analyzers so that to understand how vulnerabilities of single network service can be the cause of overall vulnerabilities and how intruders can penetrate into system by stages. A variety of approaches have been used to create attack graph so far.

An earlier approach was proposed by Sheyner [2]. It was a model checking approach. In an attack graph, nodes indicate network state and the corners indicate that the state of an attacker’s action is changing. When specific information about an attack is given, model checking technology is used to check whether it is an attack model of a system that satisfies the given characteristics. Philips and Swiler [3] developed a search engine customized to attack graph creation. In general, this approach is subject to problems in terms of scalability when state increase occurs. Ammann [4] proposed an algorithm based on graph search in order to create an attack graph. The attack graph was used in topological vulnerability analysis tools. This study assumed that an attacker’s privilege always increased while analysis was conducted. It also explains that if an attacker obtains privilege of a polynomial number, the proposed algorithm can complete an attack graph creation within the time of a polynomial expression.

Xinming Ou [5] limited approaches to create a logical attack graph of which the logical dependence between attack target and configuring information can be directly explained. The logical attack graph holds a polynomial size in relation to the analyzed network. Another approach to creating an attack graph is to create an attack graph based on vulnerability attack using attack scenario configuration techniques. Ning [6] and Cuppens [7] tested preceding conditions and results of

attacks. An attack scenario is configured by adjusting results of previous attacks considering preceding conditions of the next attack. Qin and Lee [8] proposed a warning correlation approach based on statistics in order to configure an attack scenario that recognizes a new warning relationship and is not depending on the previously acquired knowledge of an attack change pattern.

In this paper adopted an approach to create an attack graph based on data mining technology for prediction of cyber threats. The concept of using data mining at security monitoring was proposed in [9]. Thurimella [10] showed that the correlation among combinations of warnings within a security monitoring log, a result of attackers and the acts of attack, could be verified using the data mining approach.

3 Prediction Based on Security Monitoring Event

The existing time series prediction model was a frequency prediction of cyber threat items and, therefore, difficulties existed in using the prediction results. The prediction model based on security monitoring event analytically expressed the prediction results in order to improve on the weakness of time series prediction model. As a ground for analytical expression of results according to the principle of five Ws and one H, security monitoring events were used. In addition, using information extracted from security monitoring events, the source area/target area of attack can be expressed as well as visual/geographic expression of threat.

The prediction model based on security monitoring even includes the functions to collect and pre-treat security monitoring events, extract threads and sessions, create attack scenarios through correlation analysis, predict intrusions and express analytical results.

The function to collect and pre-treat security monitoring events is to collect security monitoring events and the related data, convert the collected data into a uniformed format, identify repetitive security monitoring events and contract data by integrating the events into a single event. The function to extract attack threads and sessions is to extract events to attempt single-directional and bi-directional intrusions between source area and target area. The function to create attack scenarios through correlation analysis is to identify context of security monitoring events using sequential association rules and therefore to create sequential rules and extract the time of event occurrence from the sequential rules. The function to predict intrusions is to predict intrusions by searching events on the sequential rules when security monitoring event occur and by considering the context of events. The function to express analytical results is to express prediction results in GUI environment according to the principle of five Ws and one H.

Figure 1 shows the structure and operation flow of the proposed prediction system based on security monitoring event.

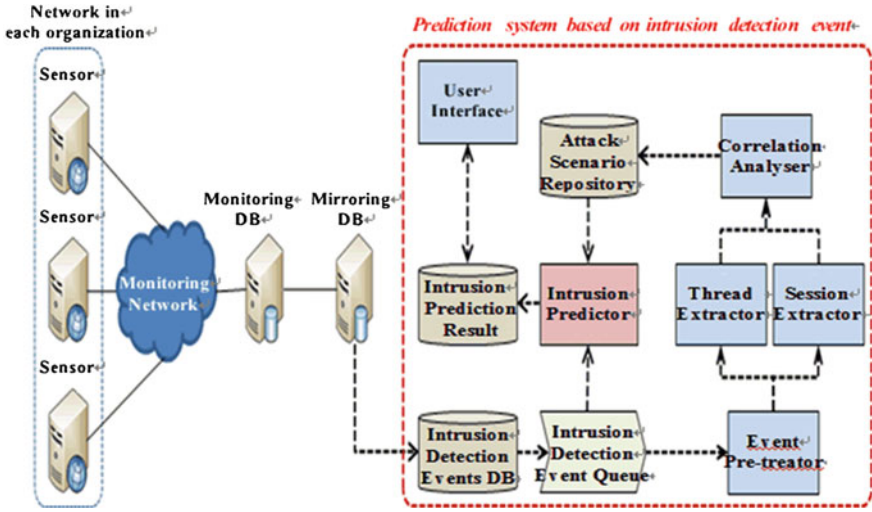


Fig. 1 Structure and operation flow of prediction system based on security monitoring event

4 Predicting Time of Security Monitoring Event

As statistical distribution characteristics of the differences in the time of occurrence between events, 90 % or more is 0 s in case of thread unit sequential rules and a distribution by max. 1 day or more is displayed. In case of session unit sequential rules, a distribution by min. 0 s and max. 2 or 3 days is displayed. As shown in (Fig. 2), a standard deviation of time differences between events is large and therefore, it is difficult to predict the time of event occurrence using average values.

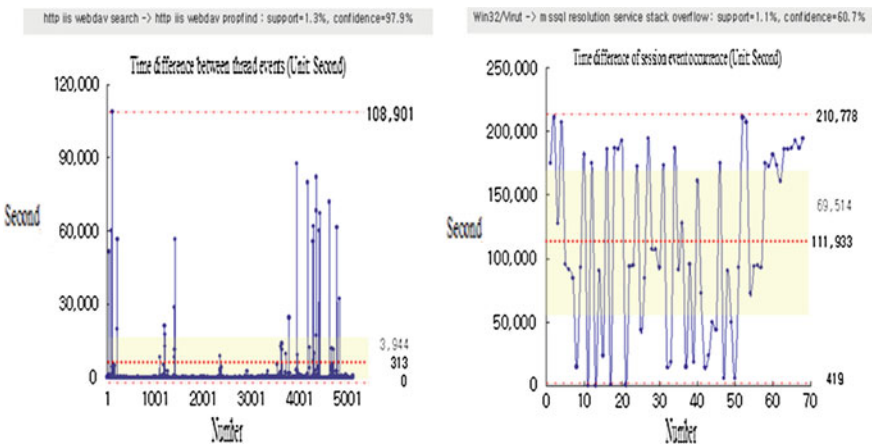


Fig. 2 Statistical distribution of differences in the time of occurrence between events

As for prediction of intrusion attempt event time, there is a method to predict an event occurrence within Tmax time based on the max. Time (Tmax).

- $$T_{\text{prediction}} \leq T_{\text{max}}$$

Another method is to predict an event occurrence with average time (Tavg) and to suggest the min. value (Tmin), max. value (Tmax) and standard deviation (Tstdev) at the same time.

- $$T_{\text{prediction}} = T_{\text{avg}}(T_{\text{min}}, T_{\text{max}}, T_{\text{stdev}})$$

Security monitoring events occurring in government and public organizations are collected into the source DB of enterprises every 5 min. Therefore, only the events of intrusion attempt of which time difference between the security monitoring events is 5 min or longer can be predicted even if the time of prediction is ignored.

4.1 Predicting of Security Monitoring

Intrusion predictor searches events on an attack scenario when an security monitoring event takes place and then predicts events to occur afterwards. When a single security monitoring event occurs, one or more events can take place afterwards. To select events among a number of possible events to present as the results of prediction. The selection criteria are necessary.

Sequential rules extracted by using sequential association rule present prediction results according to the following criteria.

- Select an event of high confidence

- $$confidence(E_n) > confidence(E_k) : \text{Select } E_n$$

- $$confidence(E_n) < confidence(E_k) : \text{Select } E_k$$

- Select an event of high support

- $$support(E_n) > support(E_k) : \text{Select } E_n$$

- $$support(E_n) < support(E_k) : \text{Select } E_k$$

- Select an event of high priority: Priority is decided according to frequency of repetitive detection of sequential rules.

– $priority(A \rightarrow E_n) > priority(A \rightarrow E_k) : \text{Select}E_n$

– $priority(A \rightarrow E_n) < priority(A \rightarrow E_k) : \text{Select}E_k$

If it is impossible to select a single event under the conditions above, both E_n (security monitoring event n) and E_k (security monitoring event k) are presented as prediction results. However, it must be implemented so that application status of the conditions above, critical values of comparison and application priorities can be changed by user.

Allocation of priorities according to frequency of repetitive detection of sequential rules is carried out as of the following.

P_{init} : Initial priority (>0)

LP : Learning period

P_{dec} : Amount of priority decrease in case sequential rule is not detected

P_{inc} : Amount of priority increase in case sequential rule is repetitively detected.

$P(A \rightarrow E_n)$: Priority of sequential rules through which E_n event occurs after A event

Initial sequential rule creation: $P(A \rightarrow E_n) = P_{init}$

The same sequential rule not detected at learning:

$P(A \rightarrow E_n) = P(A \rightarrow E_n) - P_{dec}$

The same sequential rule detected at learning:

$P(A \rightarrow E_n) = P(A \rightarrow E_n) + P_{inc}$

If $P(A \rightarrow E_n) = 0$: Deleting the sequential rule concerned $P(A \rightarrow E_n)$

※ Priority increases when the same sequential rule is detected. Priority decreases when the same sequential rule is not detected.

5 Conclusions

The prediction model based on security monitoring event analytically expressed the results of prediction in order to improve on the weakness of a time-series prediction model. In addition, this prediction model shows a possibility of threat prediction by analyzing correlation of security monitoring events.

As a result of a test on an independent prediction model based on security monitoring event, not only the problems of time required in prediction and verification of security monitoring events and of prediction effectiveness due to a narrow gap in the occurrence of security monitoring events, but also the fundamentally intrinsic problems of cyber threat prediction, such as problems concerning security monitoring data and security monitoring rules, stability and accuracy of security monitoring system and also requirements of the automated prediction and handling systems in terms of monitoring have been raised. The problems are difficult to solve under the current national environments unlike the problems raised under ideal environments proposed by the related studies. It is considered that prediction based on security monitoring event is possible only in environments where technical problems for prediction do not exist, usable data are available and the operating system is completely equipped.

References

1. 'Attack graphs' predict computer security. <http://www.eetimes.com/showArticle.jhtml?articleID=209601075>
2. Oleg Sheyner SJRL, Haines J, Wing J.M. (2002) Automated generation and analysis of attack graphs. In: Proceedings of the 2002 IEEE symposium on security and privacy (SP 2002), pp 273–284
3. Phillips C, Swiler LP (1998) A graph-based system for network-vulnerability analysis. In: Proceedings of the 1998 workshop on New security paradigms. ACM Press, New York
4. Ammann DWP, Kaushik S (2002) Scalable, graphbased network vulnerability analysis. In: Proceedings of 9th ACM conference on computer and communications security, Washington
5. Xinming Ou WFB, McQueen MA (2006) A scalable approach to attack graph generation. In: ACM conference on computer and communications security. Alexandria, Virginia
6. Ning P, Cui Y, Reeves DS, Xu D (2004) Techniques and tools for analyzing intrusion alerts. *ACM Trans Inf Syst Sec* 7(2):274
7. Cuppens F, Mieke A (2002) Alert correlation in a cooperative intrusion detection framework. Proceedings 2002 IEEE symposium on security and privacy, IEEE Computer Society, Berkeley, p 202
8. Lee W, Qin X (2003) Statistical causality analysis of infosec alert data. In: Proceedings of the international symposium on the recent advances in intrusion detection (RAID 2003), Springer, New York, pp 73–94
9. Lee SSW (1998) Data mining approaches for intrusion detection. In: Proceedings of the 7th USENIX security symposium. pp 79–94
10. Treinen JJT, Ramakrishna T A (2006) Framework for the application of association rule mining in large intrusion detection infrastructure. In: Proceedings of the international symposium on the recent advances in intrusion detection (RAID 2006), pp 1–18

A Study for Car Insurance Service Using Vehicle Real Time Information

Yong-Yoon Shin and Byung-Yun Lee

Abstract Recently the convergence of IT Services is using much information. Fusion of vehicle information technology and IT services are developing very fast. In this paper is about car insurance service used vehicle information and user information. In this paper, the car insurance service is convergence that vehicle information and user information. We propose accident service, eco-driving service, breakdown service based on the information collected from the vehicle.

Keywords Car insurance · OBD · Vehicle information

1 Introduction

Most users are used tort system for resolving accident fault and penalty. Some nations, including USA, use a no-fault system where the driver's insurance guarantees liability coverage, but the policy holder forfeits the right to seek legal recourse of the claim is settled unsatisfactorily. No fault insurance systems provide protection regardless of fault, while a tort system allows you to carry the case to a court of law if you do not receive the necessary compensation in an accident.

The insurance on your vehicle is closely tied to the information of the vehicle. If you are convicted for driving without proof of insurance, your vehicle tags and registration could be suspended, along with your driver's license.

Y.-Y. Shin (✉) · B.-Y. Lee

Future Internet Service Team, ETRI, 218 Gajeongno, Yuseong-gu, Daejeon, South Korea
e-mail: uni2u@etri.re.kr

B.-Y. Lee

e-mail: bylee@etri.re.kr

Collision coverage pays to repair your vehicle if it is damaged in an accident and you are found to be at fault. Ordinarily, property liability insurance will repair the other person's car but you have to pay for your own repairs out of pocket. Collision coverage also pays for repairs due to objects hitting your vehicle, and for damage caused by an unknown fault party such as in a hit-and-run accident.

To round out your policy for maximum protection, you can purchase insurance add-ons, to supplement your car insurance policy. Financed vehicles may be required to carry GAP coverage to protect the dealership against the car being totaled or stolen, as one example, and roadside assistance to repair a flat tire is another. Auto insurance riders can also cover such conveniences as towing or reimbursement of the cost of a rental car.

2 Driving-Information Based Car Insurance Service

Insurance companies do not have much to go on when it comes to judging a client's ability to drive. Their only source of information is your driving record. The amount you pay for car insurance will directly correlate with your driving history including both accidents and traffic violations.

The most vehicle information may be collected via the On-board diagnostics (OBD). We developed E-OBD device include OBD module, wireless internet module, GPS module, G-sensor module. Via E-OBD vehicle information, location information, G-sensor information can be gathered.

2.1 Accident Roadside Assistance

Car insurance companies collect information on your vehicle to determine how much you need to pay for insurance. Many times the information is pulled from the vehicle.

Vehicle information is collected through the E-OBD. If your car in the event of an accident, E-OBD transfer accident information (previous 15 s–after 15 s) to insurance company. Then smart insurance center analyze information related to accidents judgment (Fig. 1).

The Accidents judgment Factor is:

The insurance center received accident event from E-OBD. Then insurance center judgment accident probability calculations based factor (Table 1).

The insurance center is determined accident that vehicle speed, Engine RPM, Engine percent torque data very quickly turns to zero, or bigger than gravity sensor 5 G.

If the Break On/Off data is 'Off' and Airbag data is 'Active', insurance center determined 'Big accident'. And insurance center E-Call to rescue team. The rescue team receives vehicle location and user information from insurance center (Fig. 2).

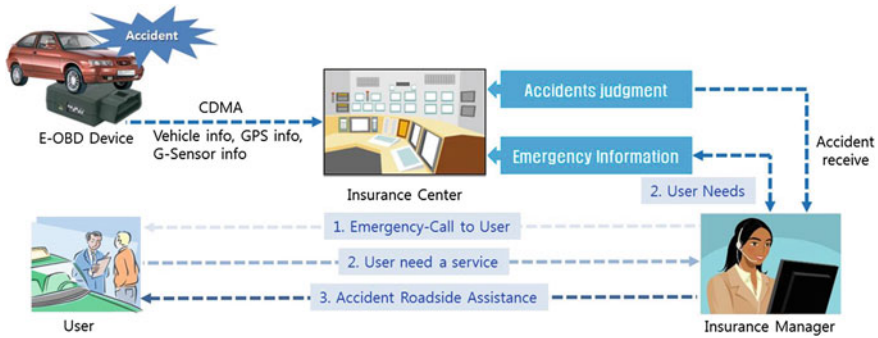


Fig. 1 Accident roadside assistance service overview

Table 1 Accident judgment main factor list (From E-OBD to insurance center)

Mode	PID	Description
00	00	User information: car model, insurance number, user name, etc....
01	0C	Engine RPM
01	0D	Vehicle speed
01	11	Throttle position
01	64	Engine percent torque data
10	00	GPS
11	01	ΔX
11	02	ΔY
11	03	ΔZ
12	01	Break on/off
12	02	Seatbelt on/off

2.2 Eco-Driving Service

Eco-Driving is the modern and smart way to save fuel and reach your destination swiftly and safely. Consuming energy/fuel costs money and causes CO2 emissions and air pollutants with negative environmental impacts. Especially driving with high engine revolutions (high RPM) raises the fuel consumption significantly. Also avoidable sequences of acceleration and braking will lower fuel efficiency.

Following the guideline, Eco-Driving enables a highly fuel-efficient, smart and relaxed driving style without any loss of time (Table 2).

The current Eco-Driving service is based on a hypothetical carbon emission. So, the actual carbon emissions and there is a difference. But E-OBD based insurance system to calculate the actual vehicle’s carbon emissions. Calculate the carbon dioxide emissions and informs users by utilizing the information of the actual vehicle.

Fig. 2 Accident roadside assistance service monitoring

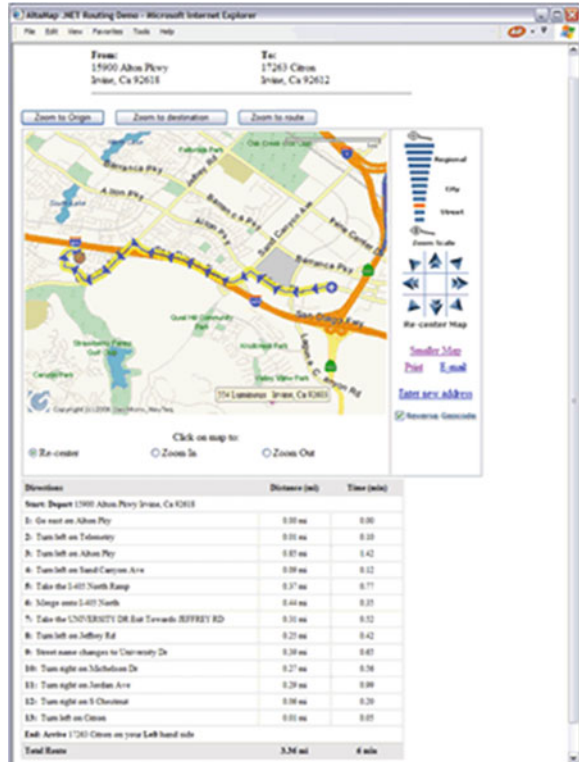


Table 2 Eco-driving main factor list (from e-obd to insurance center)

Mode	PID	Description
00	00	User information: car model, insurance number, user name, etc....
01	06–09	Long term and short term fuel % trim
01	51	Fuel type
01	0A	Fuel pressure
01	0C	Engine RPM
01	0D	Vehicle speed
10	01	Distance

2.3 Breakdown Roadside Assistance Service

Breakdown is a service that provides assistance to motorists whose vehicles have suffered a mechanical failure that is significant enough to leave them stranded at their present location. In Europe, it is popularly available via each country’s national automobile membership association, but may also be made available as part of the service of a vehicle insurance company, or other companies whose



Fig. 3 Prediction vehicle breakdown sample

primary business is to offer such assistance. Many vehicle manufacturers offer roadside assistance for their customers, sometimes for free for some period after the purchase of a new vehicle.

But this is a not automation.

We suggest cluster using technique to predict the breakdown of the vehicle. Then “Breakdown Roadside Assistance Service” can be automated (Fig. 3).

Apply a clustering technique based on the history of cars such as the type of users who have subscribed to the insurance vehicle faults that often occur. Applied based on the results of the user of the vehicle breakdown insurance should be prepared in advance.

3 Conclusions

A number of cars that can be obtained through the combination of information can make a wide variety of services. User needs to be created immediately to provide adequate services can create new services. Many car manufacturers in the car a lot of information that is collected is shown thinking of how to use. Currently, Ford collects and aggregates data from the 4 million vehicles that use in-car sensing and remote app management software to create a virtuous cycle of information [1]. Automotive telematics over coming years, an increasing number of automobiles

will be equipped with GPS and telematics (i.e., the ability to send and receive data) that can enable a range of personal safety and monitoring services [2]. Systems such as this—which are analogous to remote health monitoring in the health care system—can alert drivers to when they need repairs or software upgrades, or can locate vehicles during emergencies (e.g., when air bags have been deployed or a vehicle has been reported as stolen). Utilizing such a big-data vehicle-IT services market has enormous potential in the future.

Acknowledgments The project related to this paper is performed by the Ministry of Knowledge Economy and the Korea Evaluation Institute of Industrial Technology as part of National Platform Technology Development Project. [Development of the common component and the repository for Vehicle-IT convergence platform (10033536)].

References

1. King R, Ford gets smarter about marketing and design. CIO J
2. McKinsey Global Institute (2011) Big data: the next frontier for innovation, competition, and productivity, MGI report, June 2011

An MMU Virtualization for Embedded Systems

Sung-Hoon Son

Abstract Recently various virtualization techniques have been applied to embedded systems. Among various hardware resources, memory is one of the most important components to virtualize an embedded system. In this paper, we modify an existing virtual machine monitor (VMM) in order to support MMU functionality. The proposed VMM can support many guest operating systems, which are based on flat memory model, at the same time. The measurement study shows that the proposed scheme does not degrade performance of VMM significantly, while restrict memory access of each guest OS in order to protect the rest of the system effectively.

Keywords Virtualization · Virtual machine monitor · MMU

1 Introduction

System virtualization is recently gaining significant interest in the embedded domain [1]. In [3], a virtual machine monitor for embedded system, called VMSquare, is proposed. Although it virtualizes CPU, distributes interrupts among virtual machines, it fails to provide efficient memory protection since it does not virtualize memory.

Virtual machine monitors for embedded system ask for memory virtualization for the following reasons. First of all, since most of real-time operating systems for embedded system support virtual memory, it is required for virtual machine monitor to support virtual memory. Second, virtual machine monitor provide each guest

S.-H. Son (✉)

Department of Computer Science, Sangmyung University, 7, Hongji-dong,
Jongno-gu, Seoul, South Korea
e-mail: shson@smu.ac.kr

operating system with strict memory protection. Due to various architectural characteristics in accordance with its purpose and environment, memory management of embedded system does not have a typical form. Especially, [5] suggests four memory models for real-time operating systems in embedded system. A virtual machine monitor should be able for accommodate such a diverse memory models.

In this paper, we propose a virtual memory management scheme for embedded system virtualization. The remainder of this paper is organized as follows. In Sect. 2, the design of the proposed virtual memory scheme is described. In the following section, we present the results of detailed performance study of the virtual machine. In Sect. 4, we summarize our conclusions.

2 Design of Virtual Memory

Figure 1 depicts a new memory layout of VMsquare compared to the old one which is based on the flat memory model. As shown in Fig. 1a, VMsquare and virtual machines occupy predetermined memory areas. On the other hand, VMsquare manages each virtual machine’s memory usage using MMU in Fig. 1b. For this reason, VMsquare build a page table and every virtual memory access is translated into physical memory. The whole virtual address space is composed of three different areas; fixed address region, dynamic address region, and error region, respectively.

In fixed address region, there is one-to-one correspondence between virtual memory address and physical memory address. Usually peripheral devices, flash memory, page table, kernel reside in this region. In dynamic address region, virtual

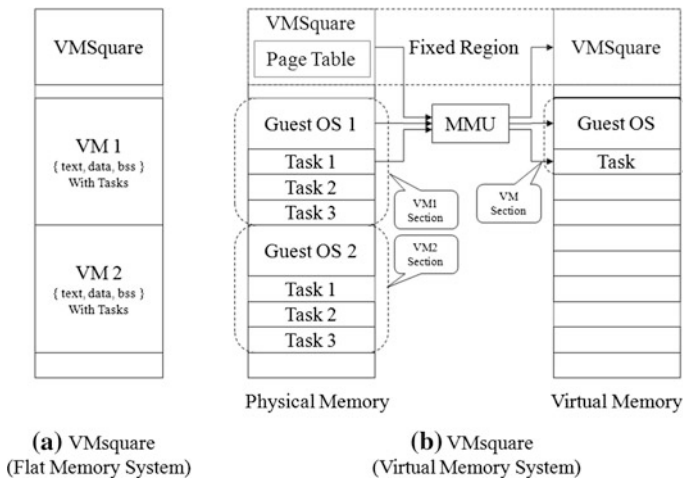
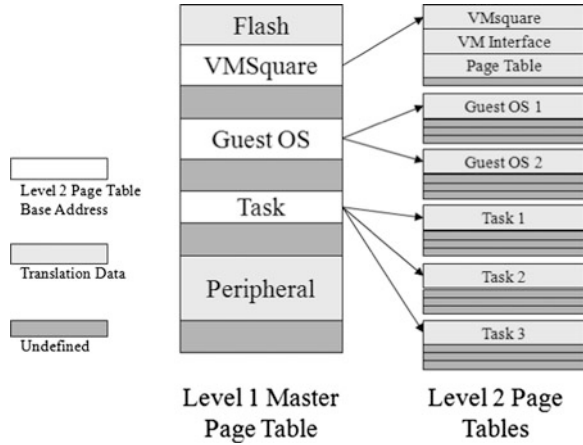


Fig. 1 VMsquare memory architecture. **a** VMsquare (flat memory system). **b** VMsquare (virtual memory system)

Fig. 2 Page table



address is dynamically mapped to physical address according to page table. Guest operating system and its tasks reside in this area. By this, virtual machine monitor can manage memory to the level of tasks. All other areas outside is undefined region, and any access to this region is regarded as an error and treated accordingly.

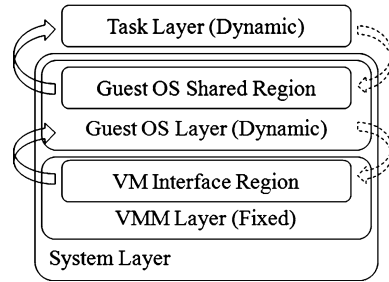
VMSquare adopts two-level page table as shown in Fig. 2. Level 1 page table is the master page table which shows the overall virtual memory. Level 1 (master) page table determines the overall virtual address space. Pages for fixed region, flash memory and peripheral devices, are allocated in this level. Level 2 page table is for dynamic area where each guest operating system and its tasks reside.

The overall system is composed of three software layers, the virtual machine monitor, guest operating systems, and user tasks. From the access privilege point of view, at least three different privilege modes are needed to accommodate these software layers. However, ARM9, which is assumed in this paper, provides only two different privilege modes, user mode and system mode. Therefore the virtual machines monitor and guest operating systems should be running on the same privilege mode on condition that they are reliable codes. Figure 3 depicts the relationship among three software layers.

As shown in Fig. 3, user tasks are running on user privilege mode. They cannot access memory areas of system privilege mode. Communication with guest operating system must be done via predefined shared memory area. The virtual machine monitor and guest operating systems exist on the system privilege mode. Although software layers exist on the same mode, communication between the layers must be accomplished via predefined virtual machine interface areas.

There are two types of context switches in virtualized system: a context switch between tasks and a context switch between guest operating systems. For task-level context switch, a guest operating system informs virtual machine monitor who is the next task to run via virtual machine interface. Then, virtual machine monitor recognize the switch-in task and place the task in virtual address space by activating corresponding page table entries. For guest OS-level context switch,

Fig. 3 Software layers on different privilege modes



virtual machine monitor saves the state of current guest operating system and its tasks, then switch to next guest operating system and its tasks by activating its page table entries.

3 Performance Evaluation

In this section, we evaluate the performance of the proposed virtual memory system. We developed VMsquare and its virtual memory on S3C2440 chip which is based on ARM920T and μ COS-II operating system which is modified to support MMU.

We first measure whether the context switch time is affected by the proposed virtual memory. Since the new context switch process includes switch of page tables, longer context switch time is expected. Table 1 shows the results of the measurement. As shown in the table, context switch time increases 40 % for task switch. For switch between guest operating systems, it increases 33 %.

Next we evaluate how CPU utilization is affected by the virtual memory. Table 2 shows the result of the experiment. As shown in Table 2, adoption of virtual memory does not affect CPU utilization immensely. For this experiment, we organize 10 guest operating systems, each of which consists of 10 user tasks. The configuration of task group is as follows:

- 96 % VMs: μ C/OS-II with nine CPU-bound tasks and one I/O-bound task
- 60 % VMs: μ C/OS-II with five CPU-bound tasks and five I/O-bound tasks
- 12 % VMs: μ C/OS-II with one CPU-bound task and nine I/O-bound tasks

Table 1 Context switch time

nVM (nTicks)	Average time (nTicks)	
	Non-MMU	MMU
Task switch	5	7
Guest OS switch	9	12

Table 2 CPU utilization (%)

nVM (nTicks)	96 %		60 %		12 %	
	Non-MMU	MMU	Non-MMU	MMU	Non-MMU	MMU
10 (50000)	85	83	72	71	32	31
20 (75000)	79	78	69	69	30	30
30 (100000)	76	74	66	65	28	28
40 (125000)	65	61	52	49	15	12
50 (150000)	54	52	40	37	8	6

2000 ticks for each VM

Table 3 Throughput

nVM (nTicks)	96 %		60 %		12 %	
	Non-MMU	MMU	Non-MMU	MMU	Non-MMU	MMU
10 (50000)	156	154	132	132	87	85
20 (75000)	264	248	232	221	175	172
30 (100000)	376	357	304	276	264	258
40 (125000)	423	397	385	364	346	321
50 (150000)	465	446	435	416	396	365

2000 ticks for each VM

We also evaluate how throughput is affected by the virtual memory. Table 3 shows the difference of throughput between the original VMSquare and VMSquare with virtual memory. The result shows that the difference is under 5 %, and adoption of virtual memory does not affect throughput immensely.

4 Conclusions

Isolation is one of the most important features of virtualization. In this paper, we develop a virtual memory system for virtual machine monitor, which provide each guest operating system memory isolation. Moreover, it makes virtual machine monitor copes with invalid memory access efficiently. Experimental results show that virtual memory does not affect performance of virtual machine monitor.

References

1. Heiser G (2008) The role of virtualization in embedded systems. In: Proceedings of the 1st workshop on isolation and integration in embedded systems, pp 11–16
2. Son S, Lee J (2009) Design and implementation of virtual machine monitor for embedded systems. J Korea Soc Comput Inf 14(1):55–64

3. Su D, Chen W (2009) SmartVisor: towards an efficient and compatible virtualization platform for embedded system. In: Proceedings of the 2nd workshop on isolation and integration in embedded systems, pp 37–41
4. Tietz L (2001) Virtual memory within embedded systems—marketing hype or engineering reality. *Dedicated Systems Magazine*, 2001
5. Zhou X, Petrov P (2007) The interval page table: virtual memory support in real-time and memory-constrained embedded systems. In: Proceedings of conference on integrated circuits and systems design, pp 294–299

The Design and Implementation of Start Stop System with Multi-Protocol in Automotive Smart Key System

Kyeong-Seob Kim, Yun-Sub Lee, In-Seong Song and Sang-Bang Choi

Abstract The researches on the start stop system, which stops the engine on idle, have been briskly carried out around the automobile makers before the appearance of the alternative energy along with the growing needs for the low energy consumption technology and the strengthening vehicle environmental regulations. In addition, the automobile makers are trying to make the start stop system general and encourage purchasing by combining the system to the popular smart key system to not only provide the convenience but reduce the energy consumption as well. In this paper, we designed and implemented the start stop system with multi-protocol that is capable of providing both an eco-friendly start stop system and a convenient smart key system on a single electronic control unit. For performance evaluation, we gathered the necessary data with a software tool for monitoring the vehicle's network, installed the proposed system on a real vehicle, measured the calculated worst case response time of the multi-protocol on the idle state control module, and compared the results with the benchmark data from the society of automotive engineers. The results show that the proposed system satisfies every time restriction that is related with the proposed system.

K.-S. Kim (✉) · Y.-S. Lee · I.-S. Song · S.-B. Choi
Department of Electronic Engineering, Inha University, 253, Yonghyeon-dong,
Nam-gu, Incheon, South Korea
e-mail: st22091108@inha.edu

Y.-S. Lee
e-mail: Skua1204@inha.edu

I.-S. Song
e-mail: nicvirus@inha.edu

S.-B. Choi
e-mail: sangbang@inha.ac.kr

Keywords Start stop system · Smart key system · Vehicle network · OBD-II · SAE benchmark

1 Introduction

The start stop system automatically stops the engine when the vehicle stops with the traffic and starts the engine again, without any inconvenience to a driver. The system is to increase the fuel efficiency, especially more with the current high oil prices. The smart key system, which is to enhance the driving convenience and the anti-theft protection, is widely being equipped to the newly released vehicles. In this paper, we designed and implemented the start stop system with multi-protocol that is capable of providing both an eco-friendly start stop system and a convenient smart key system on a single Electronic Control Unit (ECU). The OEM and the commercial bus industry have an immature market and technology than the aftermarket and the passenger car industry, as they are concentrating on the standalone start stop system. Although they have an advantage on fuel efficiency, they are not popular to general passenger car drivers due to lack of convenience. The proposed start stop system can overcome the disadvantages of the conventional start stop systems by combining the start stop system to the popular smart key system and thus, can encourage purchasing with the enhanced convenience and the enhanced fuel efficiency.

2 Design of the Proposed Start Stop System with Multi-Protocol

2.1 *The Operation Principle of the CAN Communication Module*

The Controller Area Network (CAN) communication module exchanges the ID along with some additional information with the ECU of the vehicle for further communications through the CAN bus. The operation is carried out as follows. First of all, the CAN communication module sets bitrate, timing, interrupt register, ID mask, and message filters, and then initializes itself as a normal mode. After configuring, the CAN communication module sends requests to the ECUs to obtain the information which is essential to operate the start stop system and the ECUs, which receive the requests, send the replies to the CAN communication module. At last, the start stop system analyzes the received information and starts or stops the engine according to the driver's order.

2.2 The Operation Principle of the ISO Communication Module

The International Organization for Standardization (ISO) communication module is to check and control the status of the vehicle. The ISO communication module can request and receive the information about doors, headlights, turn signals, and a transmission, and can even control them. First, the ISO communication module has to send the initialize command to wake up the ECUs as they are in sleep status for power saving. An ECU, which received the initialize command, moves to command listening status and has to receive the commands for request or control within a certain time. Else, the ECU moves to sleep status again. If an ECU successfully received the command, it returns the current status information or the result of the control operation, which is ACK, to the ISO communication module. When the ISO communication module does not receive the ACK within 50 ms, it regards the command as failed. And after 5000 ms, the ECU goes into sleep status again, so the ISO communication module has to go through the initialize process on next communication.

2.3 The Operation Principle of the Start Stop System

The start stop system starts or stops the engine after analyzing the vehicle status through the communication protocols. When the engine is on, the shift lever becomes P (Parking) or N (Neutral), and the brake is on, the start stop system requests the Revolutions Per Minute (RPM), the speed, and the temperature of the coolant through the CAN communication module. The start stop system stops the engine if the status of the vehicle indicates that the RPM is equal or less than 1500, the speed is 0, and the coolant temperature is equal or greater than 70 °C, after analyzing the response to the status request. And the start stop system starts the engine again if the brake becomes off. The coolant temperature has to be checked as the temperature has to be above certain point for a rapid ignition of the engine. Figure 1 shows the flowchart of the start stop system.

3 The Implementation and Evaluation of the Proposed Start Stop System

3.1 Implementation and Operation of the Start Stop System

The software of the proposed start stop system is developed using the MPLAB IDE 8.0. To gather the necessary data for the performance evaluation, we used a software tool for monitoring the vehicle's network. The monitoring tool can do the

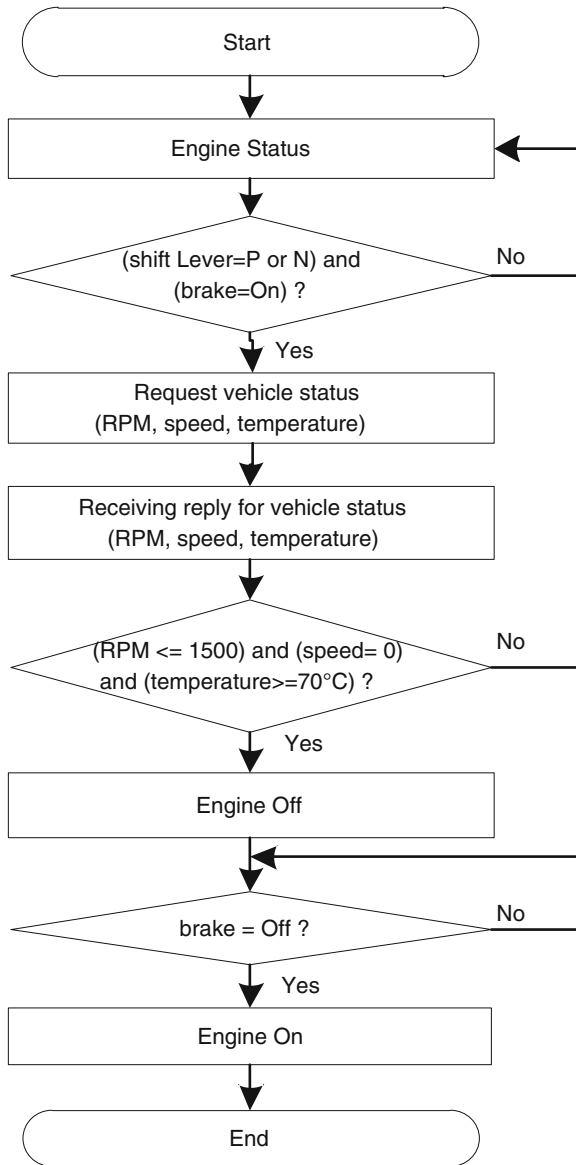


Fig. 1 The flowchart of the start stop system

diagnostics, the ECU simulation, the data collection, the automated test, and the monitoring of the vehicle's network.

3.1.1 The Implementation of the Start Stop System

The start stop system consists of the BCM of the vehicle, the main control module, and the idle state control module. In this paper, we implemented the idle state control module followed by attaching the implemented idle state control module to the conventional smart key system to build the start stop system. The PIC18F248 and the MCP2551 from the Microchip are used as the controller of the start stop system and as the transceiver. We applied 4 MHz as an operating clock frequency with the 4 MHz oscillator and applied 12 V as an input voltage.

3.1.2 The Operation of the CAN Communication Module

The CAN communication module sends the mode data and the PID data out to the CAN bus to obtain the RPM. The ID is set to default value. The ECU, which has positive match with the ID, replies with the current RPM of the vehicle. We regarded that the engine is off if there is no reply or the RPM is 0, and that the engine is on if the reply arrives or the RPM is not 0.

3.1.3 The Operation of the ISO communication module

The ISO communication module is installed on the vehicle to control the door lock. To control the door lock, first, check the door status, and second, send the door lock command to the BCM; Lock the doors if they are unlocked, and unlock the doors if they are locked. The ISO communication module communicates through the K-Line bus with the KWP-2000 protocol to obtain the door lock status.

3.2 The Installation of the Start Stop System and the Performance Evaluation

For performance evaluation of the proposed start stop system, we used the benchmark data from the Society of Automotive Engineers (SAE).

We configured the in vehicle test to evaluate the response time of the vehicle's network protocol, as follows. First, the main control module of the start stop system transmits the vehicle status request command to the idle state control module through the serial communication. Next, the vehicle communication part of the idle state control module converts and forwards the command to the vehicle

Table 1 The worst case response time (WCRT) of each command

Num.	Signal description	Size (bits)	Period (ms)	Deadline (ms)	WCRT (ms)	To
12	Vehicle speed	8	100	100	20	V/C
20	Shift lever	3	50	20	10	V/C
21	Temperature	2	1000	1000	30	V/C
22	Speed control	3	50	20	10	V/C

BCM through the OBD-II interface. Then, the BCM sends the response to the main control module, which includes the reply of the command. On arrival of the response, the idle state control module calculates the time difference between the time when the command leaves the idle state control module and the time when the reply arrives. K. Tindel et al. shows that each command satisfies its time restriction after calculating the worst case response time of the each command. On contrary, in this paper, we measured the calculated worst case response time of each command with multi-protocol on the idle state control module, compared the results with the benchmark data from the SAE, and found that the proposed system satisfies all time restrictions. Table 1 shows the worst case response time of each command.

4 Conclusions

In this paper, we proposed the start stop system with multi-protocol that is capable of providing both an eco-friendly start stop system and a convenient smart key system on a single electronic control unit. The comparison results of the worst case response times with the SAE benchmark data show that the proposed start stop system satisfies every time restriction that is related with the proposed system.

The further researches will focus on application area such as applying a human Body Area Network (BAN) to a vehicle security system, which is one of the most intuitional security systems, and on a network which can meet the requirements on class D for the fast and reliable data transmission.

References

1. Brooke L (2008) High-valve hybrids. *aei*, pp 25–27 Apr 2008
2. Navet N, Song Y, Lion FS, Wilwert C (2005) Trends in automotive communication systems. *Proc IEEE* 93(6):1204–1223
3. Bosch (1991) CAN Specification Version 2.0, Bosch

4. ISO 14230 (2000) Road vehicles diagnostic system keyword protocol 2000. International organization for standardization
5. Tindel K, Burns A (1994) Guaranteed message latencies for distributed safety critical hard real time networks. Technical report, YCS 229, Department of Computer Science, University of New York

The Design and Implementation of Improved Anti-Collision Algorithm for Vehicle User Authentication System

Kyeong-Seob Kim, Yun-Sub Lee and Sang-Bang Choi

Abstract Dazzling development of the automotive industry as a new system for the driver's convenience and security are being applied to the vehicle. In vehicle access and start-up that occurs most frequently before and after the operation of the vehicle, continually raised the needs of the customers for inconvenience related to user authentication as to compensate for the user's convenience and improve the security of the vehicle, the smart-key system for the vehicle have emerged, Because many of the FOB key, anti-collision algorithm for a seamless multi-access is applied to the smart-key system. In this paper, we have designed and implemented improved anti-collision algorithm that dramatically reduces the communication response time required in the user authentication process immediately after by dynamically changing the order in which the request of the user ID as the user ID on immediately before in the smart key system for vehicles that use many of the FOB key. In order to evaluate the performance of the system the improved anti-collision algorithm is applied, we show the behavior of the algorithm implemented in the state actually mounted on the vehicle and verify that communication response time required for many of the FOB key was reduced by about 33 % compared to existing algorithms.

Keywords Auto-ID · Smart key system · Multi-access mode · Anti-collision algorithm · RFID

K.-S. Kim (✉) · Y.-S. Lee · S.-B. Choi
Department of Electronic Engineering, Inha University, 253 Yonghyeon-dong,
Nam-gu, Incheon, South Korea
e-mail: st22091108@inha.edu

Y.-S. Lee
e-mail: Skua1204@inha.edu

S.-B. Choi
e-mail: sangbang@inha.ac.kr

1 Introduction

The anti-collision algorithm using a time division multiple access assigns a number to each of the key FOB. Also, it sends the information associated with user authentication to the key FOB from the vehicle. So, the collision avoidance algorithm has been used by the anti-collision algorithm. At this time, in case of an increase in the number of the key FOB, the problem that increases the amount of authentication time for the vehicle was occurred. Consequently, vehicle door is constantly locked in state due to the Jam phenomenon occurred at the same time an user authentication and the door opening action. Because of this, users who want to enter in the vehicle feel inconvenience caused by an increase of response time delayed between the vehicle and the key FOB.

In this paper, we designed and implemented the improved anti-collision algorithm that dramatically reduces the communication response time required in the user authentication process immediately after by dynamically changing the order in which the request of the user ID as the user ID on immediately before in the smart key system for vehicles that use many of the key FOB. As a result, communication response time required is dramatically reduced.

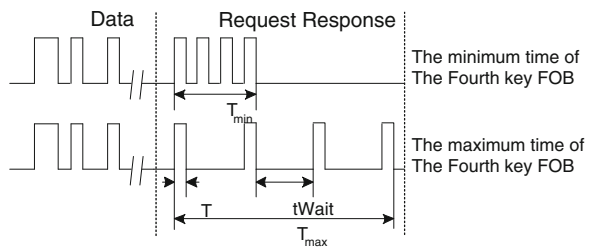
2 The Design of the Improved Anti-Collision Algorithm

2.1 The Anti-Collision Algorithm of the Smart key system

The smart key system proposed in the Texas Instruments is able to use up to four key FOB. Also, the collision is avoided by using the time division multiple access. Figure 1 shows the minimum and the maximum transmission time of the Request Response portion for the fourth key FOB as a portion of the Low Frequency (LF) data communication protocol.

T_{min} needs to keep the time of 100us and the t_{Wait} does the time of 3 ms due to the characteristics of the LF receiver chip. In addition, the information related on the user ID and the command of the key FOB is sent to the vehicle by the Ultra High Frequency (UHF) data. Then, the time of the t_{Wait} in case of the existence within the same wireless communication range and the authentication match will

Fig. 1 Transmitting time of the fourth key FOB



be a maximum. In other words, the maximum of the t_{Wait} means the UHF data is transmitted from the key FOB to the vehicle and it is taken 57.6 ms to transfer 48 bit data. So, the minimum and maximum time of the Request Response portion with the fourth key FOB is derived as follow.

$$T_{min} = 4 * 0.1 \text{ ms} + 3 * 3 \text{ ms} = 9.4 \text{ ms} \quad (1)$$

$$T_{max} = 4 * 0.1 \text{ ms} + 3 * 57.6 \text{ ms} = 173.2 \text{ ms} \quad (2)$$

2.2 The Design of the Proposed Algorithm and System

We designed the control unit for control the smart key system, the LF transmission module, and the key FOB for transmission of the UHF data. The control unit of these is connected to the Body Control Module (BCM), the external input switch for engine start, and the electronic steering lock module by wired in order to control the vehicle.

The operation of the control unit is that the control part of the control unit first delivers transmission command to the LF module in order to transmit the LF data. Then, the LF module sends data stored in the internal memory to the key FOB. In the key FOB, the UHF data is transferred to the vehicle after user authentication for LF input data. Then, the vehicle performs the operation for each command after pre-stored user ID in the memory and the UHF data received matches in the vehicle.

The composition of the LF control unit is the LF control part, the modulation generating part, and the antenna part. The LF control unit analysis the control signals received from the control unit and transfers the LF data to the modulation generating. The antenna part emits the signals modulated into the air so that the LF data is sent to the key FOB.

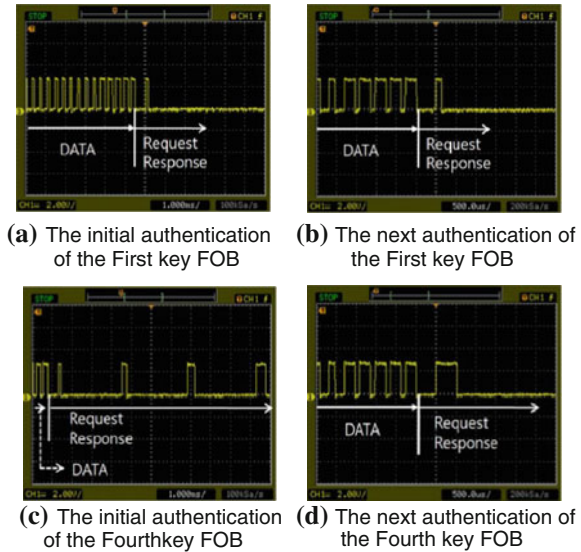
The composition of the key FOB is the receiver for the LF data, the data processing part, and the transmitter for the UHF data. The UHF data is transferred to the transmitter for the UHF data after pre-stored user ID in the memory and the LF data received matches in the key FOB.

3 The Implementation and Evaluation of the Proposed Algorithm and System

3.1 The Implementation of the System

The control unit basically sends the LF request data triggered for the external input switch to the LF module. Then, the vehicle performs the operation for each command after pre-stored user ID in the memory and the UHF data received matches in the vehicle.

Fig. 2 The signal measurement of the LF data received from the vehicle
 (a) The initial authentication of the First key FOB (b) The next authentication of the First key FOB (c) The initial authentication of the Fourth key FOB (d) The next authentication of the Fourth key FOB



3.2 The Signal Measurement of the System

Figure 2 shows the signal waves measured by a digital data received from the LF module of the vehicle.

(a) in the Fig. 2 shows the signal wave received by the initial authentication on the first key FOB. (b) shows the signal wave received by the next authentication on the first key FOB. In this case, the same Request Response signal is transmitted not changed in the transfer order. (c) shows the signal wave received by the initial authentication on the fourth key FOB. In this case, Request Response signal of the first, second, and third key FOB is transmitted because there is no response until the third key FOB from the first key FOB. (d) shows the signal wave received by the next authentication on the fourth key FOB. The point is that transfer order of the Request Response signal is changed. Therefore, the Request Response signal of the fourth key FOB is transmitted first.

3.3 The Evaluation of the System

We compared of the measured time for authentication of the conventional smart key system and the proposed system in order to evaluate the performance of the time when authenticating users.

Table 1 shows the required request response time generated by increasing the number of each key FOB authentication in the system applied to the proposed algorithm. The request response pulse width is multiples of 128us due to the

Table 1 The request response time required of the proposed algorithm

	The initial authentication	3 times	7 times	10 times
First key FOB	0.128	0.384	0.896	1.280
Second key FOB	57.984	58.496	59.520	60.288
Third key FOB	115.968	116.736	118.272	119.424
Fourth key FOB	174.080	175.104	177.152	178.688

characteristics of the LF receiver chip. The authentication time was measured with the T_{max} of the expression (2) by assuming the reception of the UHF data. In case of the fourth key FOB, even if the number of authentications increases, the authentication time is slightly increased because it changes the order of transfer right after the initial authentication. Therefore, the proposed algorithm is verify that communication response time required for many of the key FOB was reduced by about 33 % compared to existing algorithms.

4 Conclusions

In this paper, we designed and implemented the improved anti-collision algorithm that dramatically reduces the communication response time required in the user authentication process immediately after by dynamically changing the order in which the request of the user ID as the user ID on immediately before in the smart key system for vehicles that use many of the key FOB. The designed algorithm is verify that communication response time required for many of the FOB key was reduced by about 33 % compared to existing algorithms.

References

1. Finkenzeller K (2010) RFID Handbook, 3rd edn. Wiley, Chichester
2. Tang Z, He Y (2007) Research of multiaccess and an t icollsn protocols in RFID systems. In: IEEE international workshop, security, identification, Xiamen, China, pp 377–380
3. Klair DK, Chin K-W, Raad R (2010) A survey and tutorial of RFID anti-collision protocols. IEEE Commun Surv Tutor 12(3):400–421, Third Quarter 2010
4. Teas Instruments Incorporated (2001) TRIS Automotive Devices Analog Front End IC TMS37112. Reference Guide 11-09-21-046

Multi-Port Register File Design and Implementation for the SIMD Programmable Shader

Kyeong-Seob Kim, Yun-Sub Lee and Sang-Bang Choi

Abstract Characteristically, 3D graphic algorithms have to perform complex calculations on massive amount of stream data. The vertex and pixel shaders have enabled efficient execution of graphic algorithms by hardware, and these graphic processors may seem to have achieved the aim of “hardwarization of software shaders.” However, the hardware shaders have hitherto been evolving within the limits of Z-buffer based algorithms. We predict that the ultimate model for future graphic processors will be an algorithm-independent integrated shader which combines the functions of both vertex and pixel shaders. We design the register file model that supports 3-dimensional computer graphic on the programmable unified shader processor. We have verified the accurate calculated value using FPGA Virtex-4(xcvlx200) made by Xilinx for operating binary files made by the implementation progress based on synthesis results.

Keywords Graphic processor · Register file · Shader processor · HDL · FPGA

K.-S. Kim (✉) · Y.-S. Lee · S.-B. Choi
Department of Electronic Engineering, Inha University, 253 Yonghyeon-dong,
Nam-gu, Incheon, South Korea
e-mail: st22091108@inha.edu

Y.-S. Lee
e-mail: Skua1204@inha.edu

S.-B. Choi
e-mail: sangbang@inha.ac.kr

1 Introduction

High-performance graphics processor technologies that can efficiently handle massive amounts of data and intricate operation is required in order to support a realistic three-dimensional computer graphics. Existing graphics processors perform geometry and rendering operations with limited functionality. Also most of the rendering to increase a realism have been using a relatively simple algorithm receiving the support of the software. However, programmable shader receiving the support of the hardware is recently being studied so that it can be processed by the graphic processor.

In this paper, we design and implement the register file model that supports three-dimensional computer graphic on the programmable unified shader processor. we have verified the accurate calculated value using FPGA Virtex-4(xcv1x200) made by Xilinx for operating binary files made by the implementation progress based on synthesis results.

2 The Design of SIMD Programmable Shader Processor

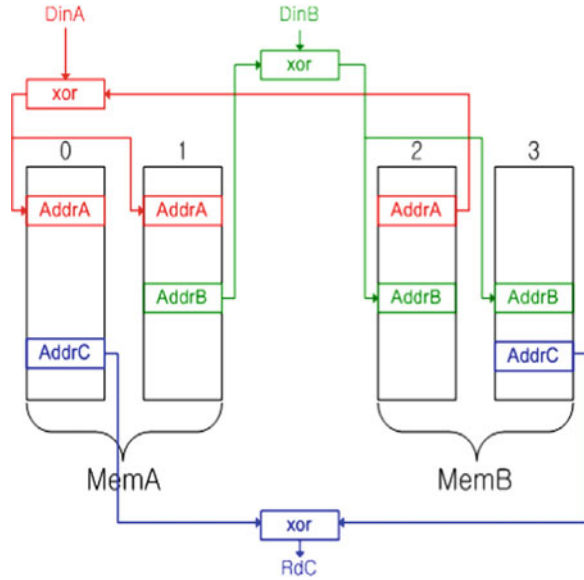
2.1 The Architecture of SIMD Programmable Shader Processor

SIMD programmable shader processor presented in this paper was basically designed with the goal to the implementation on the Virtex-4 which is one of the FPGA products of Xilinx. It is a mix of superscalar processor and vector processor architecture with the advantages of parallel processing and pipeline in order to improve performance.

2.2 The Scalable Register File Model

we propose a register file model for the SIMD programmable unified shader processor using the intellectual property (IP) provided by Xilinx FPGA. Input and output signal current of the dual-port block RAM provided by the Xilinx FPGA is present each one. Unfortunately, Input and output operation at the same time cannot be performed on a port of the dual-port block RAM. Therefore, it is difficult to use for the structure of the programmable unified shader processor required reading and writing at the same time on the same register. Thus, In order to support the one read (1R) and two write (2 W) operation simultaneously in a single cycle, we propose an improved 2W1R block RAM structure using the dual-port block RAM provided by the Xilinx FPGA as Fig. 1. The 2W1R operation is as follows.

Fig. 1 The register file with 1 read-port and 2 write-port



```

if(WriteEnableMemA)
    MemA[AddrA] := MemB[AddrA] ⊕ DinA;
endif;
if(WriteEnableMemB)
    MemB[AddrB] := MemA[AddrB] ⊕ DinB;
endif;
if(ReadEnableMemAddC)
    Rdc = MemA[AddrC] ⊕ MemB[AddrC];
    = DinA;
endif;

```

Four block RAM is used on 2W1R register file structure. Equation 1 means that the number of block RAM required in case of the write port and read port expansion.

$$\text{BlockRamNum} = 2 * (\text{WritePortNum}) + (\text{ReadPortNum} - 1) * \text{WritePortNum} \tag{1}$$

2.3 The Design of the Register File

The simulation using a benchmark program for a specific procedure frequently used in the algorithms for the graphics was performed in order to determine the structure of a register file for the SIMD programmable unified shader processor. As a result, we have verified that the Dot Product (DP), the Multiply and Add (MAD) operation is repeated in the main procedure. In addition, The Instruction Per Cycle (IPC) is increased due to the increase of the unit which is for the DP and the MAD operation. However, The IPC have not risen compared to hardware used because repeated DP and MAD operation is frequently used in the algorithms increasing more than a certain number of the Arithmetic Logic Unit (ALU). Thus, the Four ALUs was designed for programmable unified shader processor based on result of the simulation using a benchmark program.

2.4 The Time Division Mechanism of the Register File

Due to four ALUs used the pipeline architecture, the ports of register file in the SIMD programmable unified shader processor which consists of the structure to handle a large number of operands with four pixel data is accessed simultaneously. Accordingly, the time division mechanism for the proposed register file is used on the proposed register file as the Fig. 2. The clock signals for the register file and the main processor should be in order to use time division mechanism. Then, the two signals for the register file operation must be synchronized with each other. For this purpose, the buffer circuit for the register file was designed so that it is accessed from four ALUs and generates the two clock signals for the synchronization.

3 The Logic Synthesis and Verification

3.1 The Simulation for the Behavioral Level

By default, the signal validation for all registers is performed to verify the designed register file. Also, normal read and write operations are validated when simultaneously access to the register file through the various ports. The register file in the designed processor is the structure sharing the four ALUs and the ports. To this behavior, time division mechanism and internal signal selector is important. Therefore, the verification for the operation of the register file is performed with the ALUs.

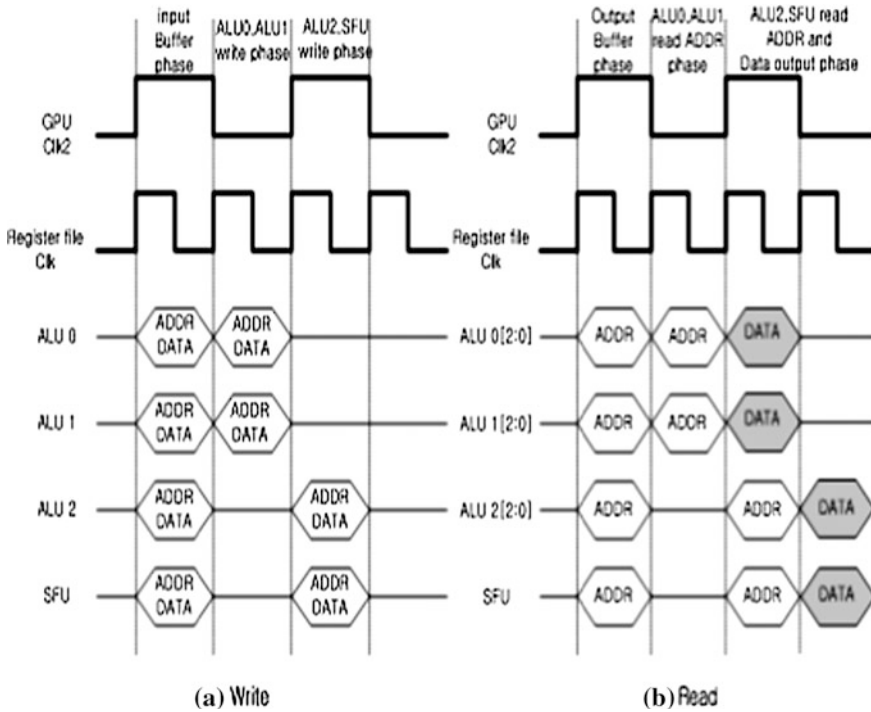


Fig. 2 Timing diagram of reading and writing the register file

Table 1 Synthesis result

Detailed module	Slices	Speed(MHz)
Top of register file	6465	206.695
Frequency divider	2	685.895
Input buffer	555	685.895
Output buffer	1927	685.895
Register file group	5461	288.663
Register file	4274	301.568
Temporary register file	1187	288.663

3.2 HDL Synthesis

The results synthesized using Xilinx ISE 9.2 shown in Table 1. Then, the binary files generated through a process of the implementation were fusing on the logic tile for xc4vlx200 based on the results of the synthesized. In addition, the binary files were targeted to the hardware by connecting the logic tile for xc4vlx200 to the versatile platform baseboard that operates as a host. As a result, we verified that the results of the simulation results with the same output.

4 Conclusions

In this paper, we proposed the multi-port register file that supports three-dimensional computer graphic on the programmable unified shader processor. Designed register file uses improved block RAM configured using dual-port block RAM quickly to process a large number of data by using a relatively small instruction in the Xilinx FPGA performing three read and one write in a single cycle. The further researches will focus on the algorithms related to the efficient management of the register file to minimize the entire system resource usage and the reduction of operating speed even increasing the size of the register file and the ports.

References

1. Hennessy JL, Patterson DA (2006) *Computer architecture, a quantitative approach*, 4th edn. Elsevier, San Francisco
2. Moya V, Gonzalez C, Roca J, Fernandez A, Espasa R (2005) Shader performance analysis on a modern GPU architecture. In: *Proceedings of the 38th annual IEEE/ACM international symposium on microarchitecture(MICRO'05)*, pp 355–364, Nov 2005
3. Atabek B, Kimar A (2008) Implementability of shading models for current game engines, *ICCES*, pp 427–432
4. Chung K, Yu C, Kim D, Kim L (2008) Tessellation-enabled shader for a bandwidth-limited 3D graphics engine, *IEEE CICC*, pp 367–370

Design Exploration Technique for Software Component Mapping of AUTOSAR Development Methodology

Kabsu Han and Jeonghun Cho

Abstract Model-based development is useful method in engineering. AUTOSAR is a kind of model-based development method for automotive E/E system which is developed by automotive industry. AUTOSAR provides several models for model-based system development, e.g. software component model, platform model and system model. Basic benefit of AUTOSAR is relocation of the software components regardless of hardware, however limitation for design exploration still exist. This paper describes constraints and requirements of software component mapping and proposes simple algorithm for design space exploration. Two kinds of constraints, software component mapping and data element mapping, are in AUTOSAR but this paper considers software component mapping only for simple algorithm. Fast design space exploration is basis of design decision and acts as input to next phase of development.

Keywords Model-based development · AUTOSAR · SWC mapping

1 Introduction

Model-based development is a useful method in engineering. The benefits of the model-based engineering are various, e.g. easy to represent, simulation, auto code generation [1–3]. AUTOSAR is open standard specification for development of

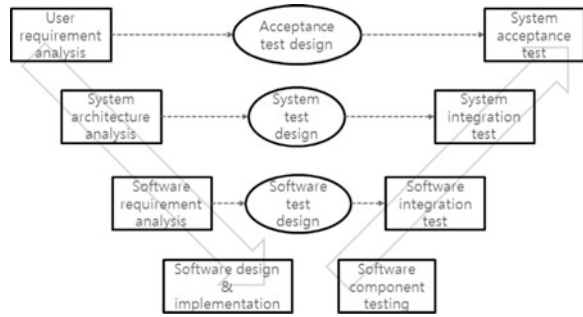
K. Han · J. Cho (✉)

School of EE, Kyungpook National University, Daehak-ro 80, Buk-gu, Daegu, South Korea
e-mail: jcho@ee.knu.ac.kr

K. Han

e-mail: kshan@ee.knu.ac.kr

Fig. 1 V-model process of development



automotive E/E system, which is developed by automotive industry [4, 5]. AUTOSAR provides software model which consists of software component and interface, platform model which contains OS, communication and basic software modules and system model which represents system topology and hardware descriptions for model based development. AUTOSAR has benefits of model-based development but some kinds of limitation still exist for design exploration of automotive E/E system. This paper considers requirements and constraints in AUTOSAR for design of automotive E/E system and proposes a simple algorithm for design space exploration.

Section 2 describes model-based development method in engineering. Section 3 describes open standard specification in automotive industry, AUTOSAR which based on model-based development. Section 4 considers requirements and constraint in AUTOSAR and propose simple algorithm for software component mapping and after that we consider some cases. Finally, Chap. 6 describes aims of research and future works.

2 Model-Based Development

For recent years, adaptation of the system architecture model and software architecture model is the one of the main trends in the area of automotive E/E system development [6]. And V-model for development of automotive E/E system is shown in Fig. 1. Various models and tools can apply to each step of the V-model, e.g. UML, Stateflow and MATLAB.

Model-based development can be divided model-based function development, which has been used in control and electronic engineering for a long time, and model-based system development, which consists of functional models, system and software models.

2.1 Model-Based Function Development

The model based function development utilizes graphical models to represent control algorithms and control flow, e.g. MATLAB, Simulink, as shown in Fig. 2.

Several reasons make the automotive industry prefers model-based approach. First, graphical functional model represents functionality of software easily. Second, verification and validation of the functionality can be performed by the executable models early in development phase. Third, the models can be used for several development phase, e.g. auto code generation, auto test-case generation and test automation which are formalized and consistent from the models to the code.

2.2 Model-Based System Development

Model based system development can cover not only functional models but also non-functional models, e.g. system architecture model for system requirement. Model-based system development consists of system models, software models and platform models normally.

2.2.1 System Models

Automotive E/E system is based on networked ECU architecture. System model represents all of the information of the system development, e.g. ECUs, network topology, I/O, functional and non-functional system requirements and system constraints. Depend on system analysis, various system models can represent

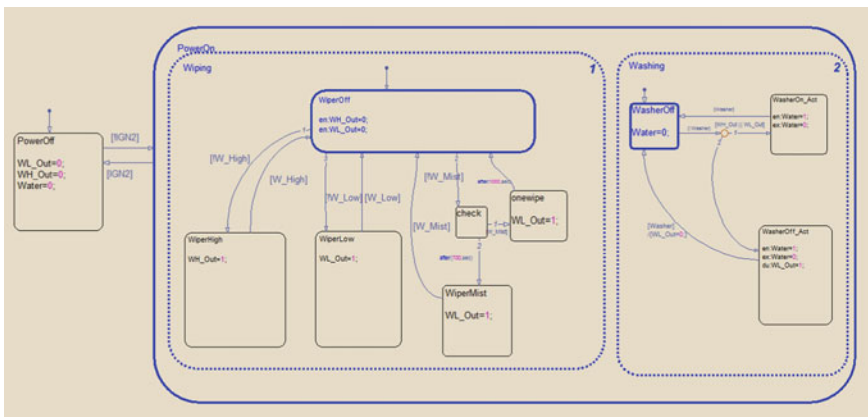


Fig. 2 Example of MATLAB/Simulink

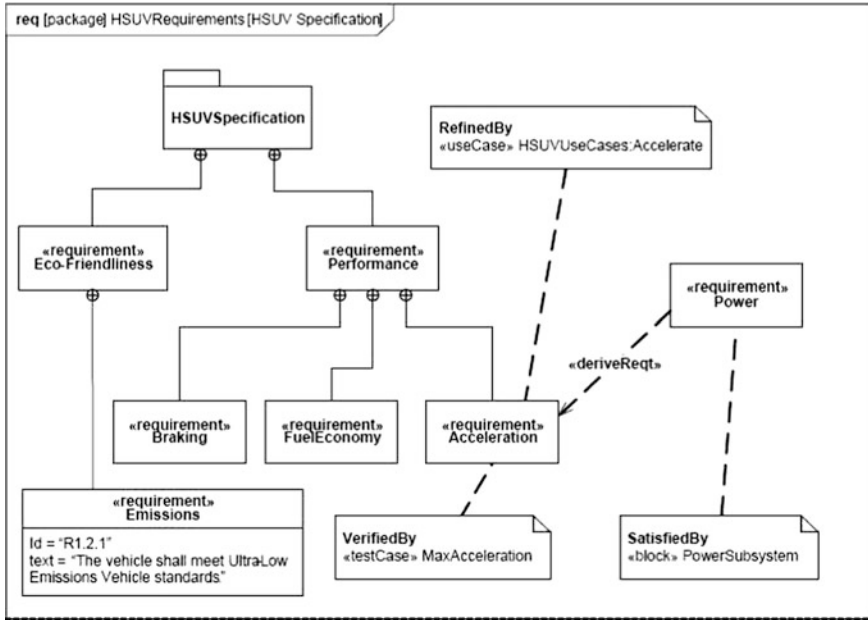


Fig. 3 Example of system modeling

system architecture efficiently. Also, several system modeling language are available, e.g. UML and SysML, as shown in Fig. 3.

2.2.2 Software Models

In automotive industry, the software models are used to model several characters of software. First, software models represent reuse of software, which based on definition of software interface for consistent usages and easy integration. Second, software models define the hierarchical interconnection of software components based on structure of software architecture. Software modeling language are available, e.g. AUTOSAR, EAST-ADL and UML.

2.2.3 Platform Models

To run software on hardware, some kind of interfaces and resource management are needed. Platform models represent basic software such as operating system, network and I/O interfaces for a system. Generally, layered models are used for platform model.

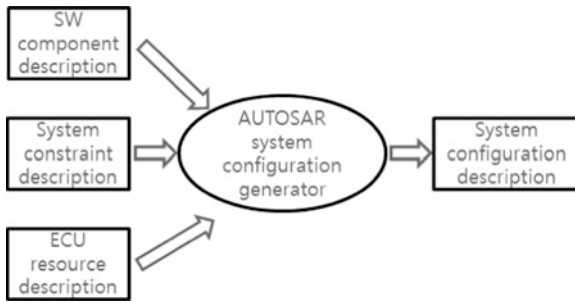


Fig. 4 AUTOSAR methodology of system level

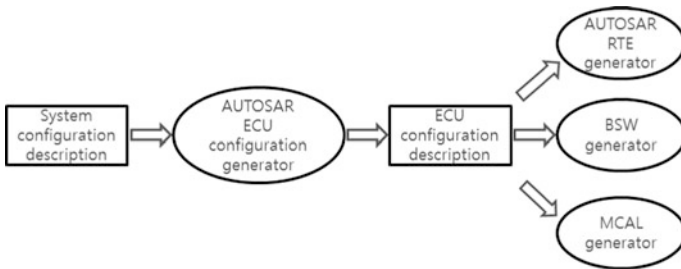


Fig. 5 AUTOSAR methodology of ECU level

3 Autosar

AUTomotive Open System ARchitecture (AUTOSAR) is an open standard specification for automotive E/E, which is developed by automotive OEM and tier1 supplier system in 2003. The main goals of AUTOSAR are management of complexity of automotive E/E system, flexibility for modification and update, scalability for variants and improvement of reliability and quality. The core technologies are software component models, standard APIs for software component and runtime environment (RTE) for communication. The process of system models of AUTOSAR methodology are shown in Fig. 4.

The models of core technologies of AUTOSAR are shown in Fig. 5. The core technologies of AUTOSAR will be done after generation of the AUTOSAR system configuration. In the generation of AUTOSAR system configuration, the main goal is the mapping of software component to ECUs [7]. Depend on the mapping of the software component, the cost and the performance of automotive E/E system varies considerably but the mapping is not pats of AUTOSAR. So this paper considers requirements for the mapping and proposes simple mapping algorithm for design exploration.

4 Mapping

The mapping of software components to ECUs is a main part of the AUTOSAR system configuration process but not a part of AUTOSAR specification. Software component description and system topology are used as input of the software component mapping. After software component mapping, the mapping of data element to bus frame will be performed. We assume that we consider the software component mapping only.

4.1 Software Component Mapping

A basic concept of AUTOSAR is the relocation of software components to ECUs. The system constraint description acts as an input to software component mapping. During iteration of system configuration, several elements are used for software component mapping.

4.1.1 Software Component to ECU Mapping

SwcToEcuMapping element can represent the mapping of software component to one ECU instance. ECU instance contains characteristics of ECU which have to be considered.

4.1.2 Software Component to Implementation Mapping

SwcToImplMapping element represents an implementation of *an AtomicSoftwareComponentType*, when various implementations are available for an *AtomicSoftwareComponentType*.

4.1.3 Software Component Mapping Constraints

MappingConstraints define invariants which have to be fulfilled by a valid mapping. Two constraints, *ComponentClustering* and *ComponentSeperation*, represent the restrictions which software components have to be mapped onto same ECU or not.

- *ComponentClustering* is used to represent some kind of software components must be mapped onto the same ECU. *ComponetClustering* means software components execute together on same ECU.
- *ComponentSeperation* is used to represent two software components shall not be mapped onto the same ECU. *ComponentSeperation* means two software components do not execute together on same ECU.

- *SwcToEcuMappingConstraint* restricts the mapping of software components to ECUs. If the *SwcToEcuMappingConstraint* is dedicated, specific software component have to be mapped onto one of numbers of dedicated ECUs. If the *SwcToEcuMappingConstraint* is exclusive, the software components cannot be mapped onto the ECUs.

4.2 Software Component Mapping Algorithm

During software component mapping, system constraints have to be considered firstly with set of ECUs, buses and software components. Among the system constrains, dedicated type is mapped onto dedicated ECUs first. After mapping of the dedicated type, search *ComponentClustering* type software components, which are clustered with already mapped software components, and map *ComponentClustering* type onto non-exclusive ECUs. The rest of software components, which are not mapped yet, can be mapped onto non-*ComponentSeperation* and non-exclusive ECUs. If software component mappings fulfill system constraints, all of software component mapping will report to system designer for manual design exploration. At least a software component cannot be mapped onto ECU, the software component mapping is fail. Search process has to start again for another design. We define set of relations between software component and ECU, software component and buses for mapping of software components to ECU. If software component and ECU are exclusive, the value of relation will be 0.

$x_{i,j}^{swc,ecu} = 1$ if software component i is mapped onto ECU j , otherwise is 0.

$y_{i,j}^{swc,bus} = 1$ if software component i uses bus j , otherwise 0.

$$\forall swc_i \in swcSet, \sum x_{i,j}^{swc,ecu} = 1 \text{ and } \sum y_{i,j}^{swc,bus} \leq 1$$

The first search for dedicated type will process all of software components, in about $O(n)$ time complexity. The second search for *ComponentClustering* and non-exclusive type will be very smaller than the number of software components, because a mapped software component is needed for *ComponentClustering* type at least. Also, *ComponentClustering* type may consist of small number of software components. In worst case, complexity will be close to $O(n^2)$ but generally will not. Rest of the software components is smaller than the numbers of all of software component definitely.

case 1	case 10	case 19
Connected SWC	Connected SWC	Connected SWC
ECU_A wiper2 / EntLamp2 / PowerWindow / OutsideMirror / TurnSignal / DoorLock	ECU_A wiper2 / EntLamp2	ECU_A wiper2 / PowerWindow / OutsideMirror / TurnSignal / DoorLock
ECU_B wiper3	ECU_B PowerWindow / OutsideMirror / TurnSignal / DoorLock / wiper3	ECU_B EntLamp2 / wiper3
ECU_C Steering	ECU_C Steering	ECU_C Steering
ECU_D wiper1 / EntLamp1	ECU_D wiper1 / EntLamp1	ECU_D wiper1 / EntLamp1
Warning hard to Scheduling ECU_A	Warning hard to Scheduling ECU_B	Warning hard to Scheduling ECU_A
==> Scheduling False ECU_A	==> Scheduling False ECU_B	==> Scheduling False ECU_A
case 3	case 14	case 23
Connected SWC	Connected SWC	Connected SWC
ECU_A wiper2 / EntLamp2 / PowerWindow / OutsideMirror / TurnSignal / DoorLock	ECU_A wiper2 / EntLamp2	ECU_A wiper2
ECU_B Steering	ECU_B TurnSignal / DoorLock / wiper3	ECU_B EntLamp2 / PowerWindow / OutsideMirror / TurnSignal / DoorLock / wiper3
ECU_C Steering	ECU_C PowerWindow / OutsideMirror	ECU_C Steering
ECU_D wiper1 / EntLamp1 / wiper3	ECU_D wiper1 / EntLamp1	ECU_D wiper1 / EntLamp1
Warning hard to Scheduling ECU_A	All ECU Scheduling available	Warning hard to Scheduling ECU_B
==> Scheduling False ECU_A	==> Scheduling Success	==> Scheduling False ECU_B
case 8	case 18	case 45
Connected SWC	Connected SWC	Connected SWC
ECU_A wiper1 / EntLamp1 / TurnSignal / DoorLock	ECU_A wiper2 / EntLamp2	ECU_A EntLamp2
ECU_B PowerWindow / OutsideMirror	ECU_B TurnSignal / DoorLock	ECU_B wiper2 / PowerWindow / OutsideMirror / TurnSignal / DoorLock
ECU_C wiper3	ECU_C PowerWindow / OutsideMirror	ECU_C wiper3
ECU_D wiper1 / EntLamp1	ECU_D wiper1 / EntLamp1 / wiper3	ECU_D wiper1 / EntLamp1
Warning hard to Scheduling ECU_A	All ECU Scheduling available	Warning hard to Scheduling ECU_B
==> Scheduling False ECU_A	==> Scheduling Success	==> Scheduling False ECU_B

Fig. 6 Test cases

	ECU_A	ECU_B	ECU_C	ECU_D		ECU_A	ECU_B	ECU_C	ECU_D		ECU_A	ECU_B	ECU_C	ECU_D
CPU	87	5	0	27	CPU	19	73	0	27	CPU	80	12	0	27
Memory	164	24	0	97	Memory	39	149	0	97	Memory	149	39	0	97
I/O Pin	85	2	0	12	I/O Pin	10	77	0	12	I/O Pin	79	8	0	12
CAN	H/L	H	N/A	H/L	CAN	H/L	H	N/A	H/L	CAN	H	H/L	N/A	H/L
	ECU_A	ECU_B	ECU_C	ECU_D		ECU_A	ECU_B	ECU_C	ECU_D		ECU_A	ECU_B	ECU_C	ECU_D
CPU	87	0	0	32	CPU	19	48	25	27	CPU	12	80	0	27
Memory	164	0	0	121	Memory	39	97	52	97	Memory	24	164	0	97
I/O Pin	85	0	0	14	I/O Pin	10	42	35	12	I/O Pin	4	83	0	12
CAN	H/L	N/A	N/A	H/L	CAN	H/L	H	H	H/L	CAN	H	H/L	N/A	H/L
	ECU_A	ECU_B	ECU_C	ECU_D		ECU_A	ECU_B	ECU_C	ECU_D		ECU_A	ECU_B	ECU_C	ECU_D
CPU	62	25	0	27	CPU	19	43	25	32	CPU	7	80	5	27
Memory	112	52	24	97	Memory	39	79	52	121	Memory	15	149	24	97
I/O Pin	50	35	2	12	I/O Pin	10	40	35	14	I/O Pin	6	79	2	12
CAN	H/L	H	H	H/L	CAN	H/L	H	H	H/L	CAN	L	H	H	H/L

Fig. 7 The results of SWC mapping

5 Case Study

SWC mapping algorithm for design space exploration is implemented on MS Windows environment on which AUTOSAR and modeling tools are based and running. Requirements and constraints for SWC mapping are provided by AUTOSAR XML document (AR-XML), which is standard input/output document type of AUTOSAR, the result of SWC mapping will be presented by AR-XML, also.

To experiment SWC mapping algorithm, we make various requirements and constraints cases with nine software components of comfort domain and 4 ECUs, shown in Fig. 6.

To make design space exploration easier, the result of SWC mapping contain several performance values which evaluated on real ECUs, e.g., CPU usage, memory usage, numbers of I/O, usage of CAN interface, shown in Fig. 7.

As the experimental requirements and constraints, 162 cases can happen. Our SWC mapping algorithm presents the correct results of every cases in few seconds. Designer can choose the SWC mapping what he want in early development time.

6 Conclusion

This paper considered requirements and constraints for the software component mapping in AUTOSAR and proposed simple mapping algorithm for design exploration. It is not a part of AUTOSAR but fundamental design of automotive E/E system. Depend on software component mapping, generation of RTE and test case varies.

We research practical software component mapping algorithm for model-based system development of automotive E/E system. It is simple but we take our first step toward practical algorithm. And it provides basis of design decision, which uses as input to next development phase, early in model-based development phase. Also, research about the mapping of data element to bus frame is needed for practical development method.

Acknowledgments This research was supported by The Ministry of Knowledge Economy (MKE), Korea, under the Convergence Information Technology Research Center (CITRC) support program (NIPA-2012-C6150-1202-0011) supervised by the National IT Industry Promotion Agency (NIPA) and the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea(NRF) through the Human Resource Training Project for Regional Innovation and Kyungpook National University Research Fund, 2012.

References

1. Oliver N, Joachim S (2008) Models for model's sake. ICSE'08, pp 561–570
2. Christian P, Andreas F, Judith H, Martin R, Sabine R, Doris W, On the integration of design and test: a model-based approach for embedded system. AST'06, pp 15–21
3. Martin R, Lars L, Ulrik E (2002) A method for model based automotive software development. 12th euromicro conference
4. Ulrich F, Vivek J, Joachim L (2009) Multi-level system integration of automotive ECUs based on AUTOSAR. SAE world congress and Exhibition'09
5. Automotive open system architecture <http://www.autosar.org>
6. Jorg S (2005) Automotive software engineering: principles, processes, methods, and tools. SAE, 2005
7. Wei P, Hong L, Min Y (2010) Deployment optimization for AUTOSAR system configuration. ICCET'10, pp 189–193

Interoperability and Control Systems for Medical Cyber Physical Systems

Min-Woo Jung and Jeonghun Cho

Abstract As quality of life is improving, a number of people pay attention to their health. Though the demand for medical services has been increased steadily, the number of doctors per capita increased only 2 % per year on average. In order to solve lack of caregiver, manufacturer of medical devices has tried to develop automatic medical systems. This system is called “Medical Cyber Physical Systems (MCPS)”. It is necessary to ensure interoperability between medical devices and closed loop system in order to automatic control. We consider IEEE 11073 to support interoperability between medical devices. We also introduce prototype platform that is based closed loop system in MCPS. This platform is our preliminary work to support interoperability between medical devices and compose a closed loop system which makes possible to control devices each other.

Keywords Personal health care · MCPS · IEEE 11073 · Closed loop system

1 Introduction

MCPS requires to ensure several important features. First, it is interoperability between medical devices. IEEE 11073 is established in order to ensure interoperability between medical devices. In spite of establishment of IEEE11073, most of

M.-W. Jung · J. Cho (✉)
School of Electronic Engineering, Kyungpook National University,
Daehak-ro 80, Buk-gu, Daegu, South Korea
e-mail: jcho@ee.knu.ac.kr

M.-W. Jung
e-mail: jungminwoo80@gmail.com

manufacturers use proprietary approach of themselves. Proprietary approach is able to make a problem for interoperability between medical devices of different manufacturers. Therefore, Integration of standard and proprietary is being researched. Second, it is high confidence software. In order to implement high confidence software, it is essential to ensure safety of software because medical device is corresponding with life of patients. In order to ensure safety of software, it is indispensable to verification and validation of software. Finally, most of important thing in MCPS is control and communication system. Because medical devices in MCPS exchange data among each other, stability of communication is important aspect.

Though the demand for medical services is been increased, the number of caregiver such as doctor, nurse cannot meet the one. In order to solve this problem, manufacturer to develop medical devices try to develop automatic medical systems. Automatic medical systems require elaborate closed the loop that is based sensing and control technologies. Automatic medical systems must be implemented elaborately due to be associated with life of patients. Implemented system must be verified for safety and effectiveness. So, automatic medical systems require verification and validation techniques [1].

We are studying integration of IEEE 11073 based standard devices and proprietary devices. In this paper, we propose conversion mechanism which converts IEEE 11073 to proprietary. We also proposed closed the loop medical system for automatic control system and designed the one using matlab in order to support verification, validation and optimized C code generation.

In this paper, [Sect. 2](#) describes backgrounds. [Section 3](#) describes Interoperability between medical devices for MCPS. [Section 4](#) considers closed loop medical system and [Sect. 5](#) deal with conversion mechanism and design of close the loop system. [Section 6](#) concludes the paper.

2 Background

2.1 IEEE 11073

IEEE 11073 work group established to develop new medical device standard specification for typical personal health care devices. Personal health care device is medical embedded system which has limited functionality and resource typically, and has network interface in order to communicate each other medical devices. New medical device standard specifications have to consider about restrictions like weight, cost, convenience, durability and duration. The work group has been developed a common base protocol for initial set of six device specializations (pulse oximeter, pulse/heart rate, blood pressure, thermometer, glucose, weighing scale) [2]. [Figure 1](#) shows categories and typical devices for personal health services. Personal health sensor devices which called agent such as pulse oximeter, blood pressure monitors, weighing scales, and pedometers, collect vital sign and information of people and send the information

to a personal health processing device which called manager such as PC, cell phone, health appliance, set top box for processing, display and backup. Manager can send the information which processed within manager to remote system for more services and backup. This is including disease management, health and fitness or aging independently applications. The communication path between agent and manager is logical point to point connection whatever physical connection is wireless or wired. Generally, an agent connected with a manager at once, called 1:1 communication. A manager can communicate with several agents simultaneously using each point to point connection, called 1:n communication. The overlay shows the focus area of the IEEE 11073 Personal health Devices Working Group. The first consideration point is the data exchange and interface between the agent and the manager. The connection between the agent and the manager, wireless or wired, may converts, exchange and send data across different interfaces. This is the scope of the IEEE 11073 in application layer but transport interface under the application layer is out of the specification [3].

2.2 Mcps

CPS is defined integrations of computation and physical processes. Embedded computers and networks monitor and control the physical processes, usually with feedback loops where physical processes affect computations and vice versa. In the physical world, the passage of time is inexorable and concurrency is intrinsic. Neither of these properties is present in today’s computing and networking abstractions [4].

MCPS is part of CPS for medical system. MCPSs are safety, interconnected, intelligent systems of medical devices. MCPS is required high-confidence systems, system of systems (SoS), closing-the-loop, guarantee of service (GoS), privacy and security, validation, certification. Figure 2 presents requirement for MCPS [5].

In order to high-confidence systems, high assurance software for safety and effectiveness is essential. Many functions traditionally implemented in hardware are increasingly being implemented in software. As medical devices get communication interfaces for interoperability, it is essential to ensure that the integrated medical devices are safe, effective, secures. Patient information exchanged during device interoperation can not only provide a better understanding of the general health of disease and generation of effective alarms in the event of emergencies.

Fig. 1 Overview IEEE 11073

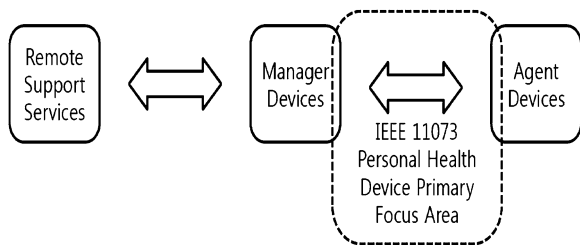
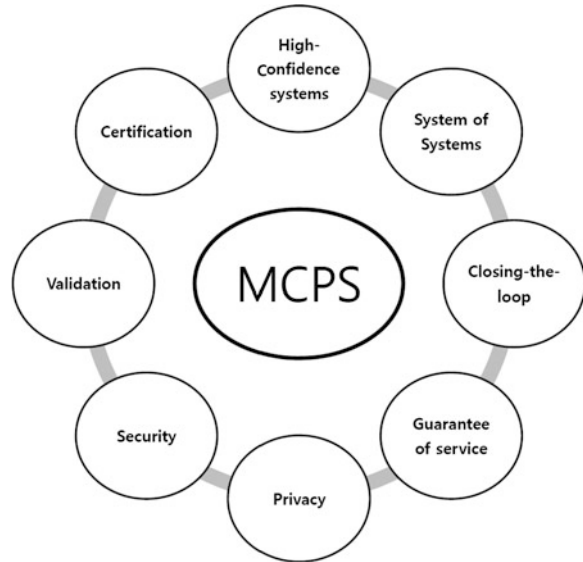


Fig. 2 Requirement of MCPS



Given the complexity of the human body and variations of physiological parameters over patient population, developing such computational intelligence is a non-trivial task. So MCPS is needed context awareness computing technology. Medical data collected and managed by MCPS is very critical [6–8]. Unauthorized access or tampering with this information can have severe consequences to the patient in the form of privacy loss and physical harm. Preserving the security of MCPS is crucial. The complex and safety critical nature of MCPS requires a cost-effective way to demonstrate medical device software dependability. Certification of medical devices provides a way of achieving this goal. The certification is an essential requirement for the eventual viability of MCPS and an important challenge to be addressed [9].

3 Interoperability Between Medical Devices for MCPS

In order to ensure interoperability between medical devices in MCPS, currently medical devices that not observe IEEE 11073 have to facilitate communication with complied-standard one. We propose the platform that convert proprietary to IEEE 11073 standard. The medical device is communicated by transport layer such as Bluetooth, Zigbee, USB. It is that exchange vital sign and device information. Proprietary compliant protocol analyzes by proprietary protocol that is based providing manual by manufacturer. Conversion Mechanism (CM) converts proprietary to IEEE 11073 standard. Converted data form suitable IEEE 11073 specializations. CM is the most of important thing in this paper (Fig. 3).

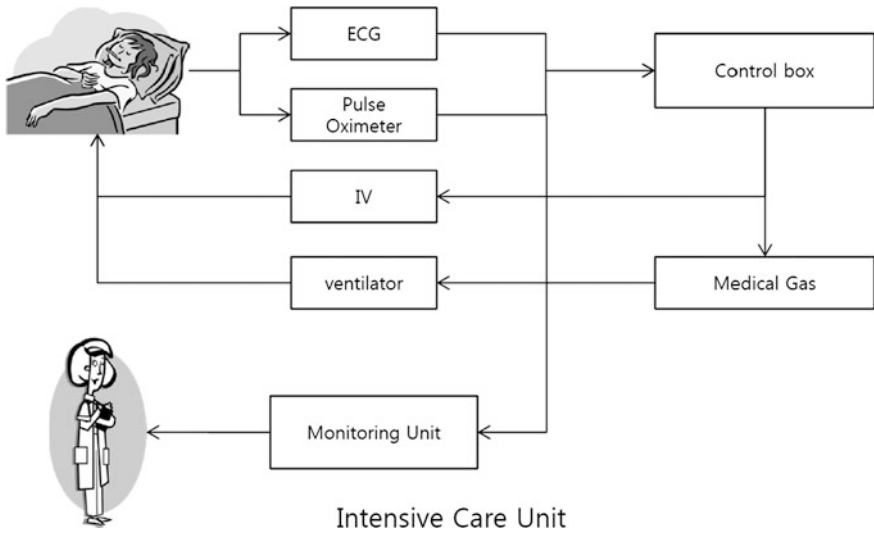
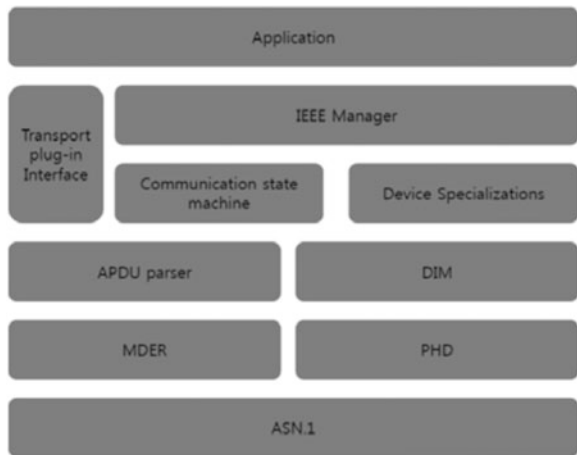


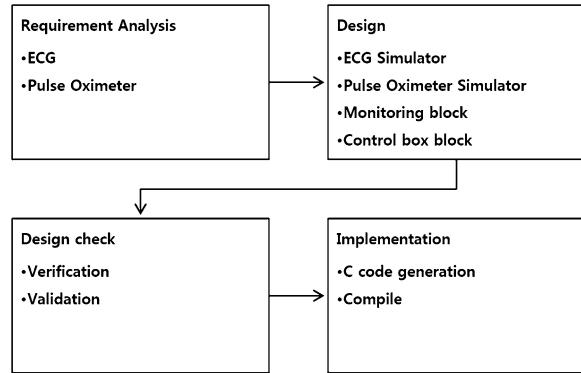
Fig. 3 Closed the loop in ICU

Fig. 4 Block of IEEE 11073 software



We deal with proprietary pulse oximeter and electrocardiogram (ECG). Pulse oximeter includes plethysmogram (PPG), SpO₂, pulse rate with vital sign. PPG presents analog data type, SpO₂ and pulse rate present digital type. Proprietary protocol is distinguishes between vital sign and information of patient. SpO₂ and pulse rate generate numeric object due to digital signal, PPG and ECG generate Real Time Sample Array (RT-SA) due to analog signal. The components have dependency to metric object. Information of medical device generates each enumeration metric object. After the procedure, proprietary data convert to IEEE 11073. Figure 4 presents overall of CM.

Fig. 5 Development process for CLS



4 Closed Loop Medical System

With the development of MCPS, complexity of medical systems has been increasing. As a result, MCPS require more compact control systems due to have relevance to patient's life.

We consider to intensive care unit (ICU) that has many device in order to measure vital sign of intensive patient. ICU is needed very compact control system that minimizes errors. The control system is most suitable in MCPS. High confidence development is very important for the safety and effectiveness of MCPSs that is based intensive software. Figure 5 presents overview of ICU in order to introduce closed the loop. Critical patient attaches ECG and Pulse Oximeter in order to monitor state of patient. The medical device such as Pulse Oximeter, ECG is observed IEEE 11073. If the medical device is not observed, it is converted through CM in our proposal. The vital sign send to control box and monitoring unit. The control box controls IV and medical gas. Though the IV is consistently supplied patient, its volume controls in state of patient. Ventilator is also controlled volume of oxygen supply.

The control box includes manager operation for communication of IEEE 11073. The monitoring unit provides not only patient's vital sign to caregiver, but also alarm functionality for emergency. The monitoring unit includes communication module such as Bluetooth, Zigbee, USB that enable communicates with medical devices.

5 Implement

5.1 Conversion Mechanism

We consider pulse oximeter in order to apply proposed mechanism. We have been implemented conversion mechanism using C language. Figure 4 presents block of software in order to observe IEEE 11073 standard. IEEE 11073 standard is written

Table 1 Specification of hardware

Sensor	Finger probe sensor
AFE	Pulse oximeter analog front-end
DSP module	TMS320C5515
Convert MCU	ATmega128
Communication	Bluetooth

using the ASN.1 language. Actual encoding of an ASN.1 message to a stream of bytes is dictated by a set of rules, like BER, XER or MDER. MDER supports a subset of ASN.1 primitive types. Medical messages are limited to those primitive types. Compound types may use any composition of the primitive types supported by MDER. Application Protocol Data Unit (APDU) is the equivalent of Bluetooth data packet. Each APDU must be decoded under MDER rules. Each DIM class contains a well-defined set of PHD type variables, and inherits from another DIM class. A specialization defines standard configurations, may also extend the collection of DIM types. The IEEE manager caches configurations in the local file-system. The communication state machine block takes care of IEEE state machine for each session, IEEE protocol timeouts, dealing with transport connections, notifying applications about data of interest. The transport plug-in interface supplies the interface that is implemented and chosen by the application. The IEEE manager is the top-level component of IEEE stack. IEEE counterparts are either agents or managers.

The hardware consists of four modules such as analog front-end (AFE), DSP Module, IEEE 11073 convert module, smart device. Pulse oximeter analog front-end module used product of texas instrument (TI). This module output analog signal such as plethysmogram and digital signal such as SpO₂, pulse rate. These signals not observe IEEE 11073 standard. The output signal transmits to TMS320C5515 DSP module through front-end connect. TMS320C5515 DSP module controls LCD in order to can be presented plethysmogram waveform and SpO₂, pulse rate.

IEEE 11073 conversion module implements using Bluetooth module and ATmega128. Conversion mechanism is porting to ATmega128. After convert standard, the data transmit to smart device through Bluetooth module. Table 1 present specification of hardware.

5.2 Closed Loop System

We consider closed loop system in intensive care unit. We design ICU system using LabView that enable c code generation and uses graphical language. We analyze system requirement such as ECG and Pulse Oximeter that contain tolerance. Based on the requirements, we design ECG simulator and Pulse Oximeter Simulator using filter library that provides by LabView. The signal that generates

in simulator operates random mode that abnormal and normal. The monitoring block is implemented systems that can monitor vital sign of patient and can inform emergency to caregiver. The control box block includes module that adjusts the amount of IV and medical gas upon patient's condition. Each block is connected with other one for verification and validation through simulations. If the system has not an error during the simulation, we obtain optimized source of the system that runs on the chosen platform. The design system has advantage that enables step-by-step checking in order to find error. Figure 3 presents the development process for closed the loop system.

6 Conclusions

In this paper, we considered interoperability between medical devices in MCPS. Due to MCPS is expensive embedded system, it requires high confidence software and interoperability between medical devices.

We propose the software platform that convert from proprietary to IEEE 11073 standard in order to ensures interoperability between medical devices and closed the loop based medical system that enable automation of medical systems. Interoperability and closed loop system is the most of important requirement in MCPS. We can obtain reduction of cost and development time through simulation of medical system. The proposed systems is prototype of MCPS, it is expected to distribute development of MCPS.

We will study model based software that enable more elaborate verification and validation of software for implement of high confidence software and develop virtual simulation that can optimize automation of medical systems.

Acknowledgments This research was financially supported by the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea(NRF) through the Human Resource Training Project for Regional Innovation.

This research was supported by The Ministry of Knowledge Economy (MKE), Korea, under the Convergence Information Technology Research Center (CITRC) support program (NIPA-2012-C6150-1202-0011) supervised by the National IT Industry Promotion Agency(NIPA).

References

1. Lim J-H, Park C, Park S-J (2010) Home healthcare settop-box for senior chronic care using ISO/IEEE 11073 PHD standard. 32nd annual international conference of the IEEE EMBS, 2010
2. Pedersen S (2004) Interoperability for information systems among the health service providers based on medical standards. *Informatik-Forschung Und Entwicklung* 18:174–188
3. Martinez I, Escayola J, Martinez-Espronceda M (2008) Standard-based middleware platform for medical sensor networks and u-health. Proceedings of 17th international conference on computer communications and networks, 2008

4. Lee EA (2008) Cyber physical systems: design challenges. International symposium on object/component/service-oriented real-time distributed computing (ISORC), 2008
5. Stylianides N, Dikaiakos MD, Gjermundrod H, Panayi G, Kyprianous T (2011) Intensive care window: real-time monitoring and analysis in the intensive care environment. *IEEE Trans Inf Technol Biomed* 15(1)
6. Fioravanti A, Fico G, Arredondo MT, Salvi D, Villalar JL (2010) Integration of heterogeneous biomedical sensors into an ISO/IEEE 11073 compliant application. 32nd annual international conference of the IEEE EMBS, 2010
7. Yao J, Warren S (2005) Applying the ISO/IEEE11073 standards to wearable home health monitoring systems. *J Clin Monit Comput* 19:427–436
8. Lim S, Chung L, Han O, Kim J-H (2011) An interactive cyber-physical system (CPS) for people with disability and frail elderly people. Proceedings of the 5th international conference on ubiquitous information management and communication, 2011
9. Lee I, Sokolsky O, Chen S (2012) Challenges and research directions in medical cyber-physical systems. *Proc IEEE* 100(1):75–91

Ubiquitous Logistics Management in the Steel Industry

Sang-Young Lee and Yoon-Seok Lee

Abstract Recently, steel industry including the ubiquitous logistics cost savings with IT skills and enhance their competitiveness in terms of traceability and logistics management through efficient management and utilization of real-time information about the importance of awareness and with increased require of needs logistics management services based on ubiquitous IT environment, focusing on the core of the many services that can be said to belong to the service area. Thus, the introduction of ubiquitous technology in the field of logistics in the steel industry, with some merits at minor degree and although some advantage is very small level. In this paper, presents the feasibility of deploying RFID for the steel industry as a tool to reduce the production costs. Steel industry that is applicable to RFID-based tracking management system is proposed. The results of this paper proved that the recognition of 100 % came from the material input and output and the location, the location indicated 99 % detection rate. Therefore, the proposed RFID-based tracking management system was found superior to the existing system in terms of productivity.

Keywords Ubiquitous · Logistics management · Steel indus

S.-Y. Lee (✉) · Y.-S. Lee

Department of Health Administration, Namseoul Universit, Cheonan, South Korea
e-mail: sylee@nsu.ac.kr

Y.-S. Lee

e-mail: yslee@nsu.ac.kr

1 Introduction

Ubiquitous technology Radio Frequency Identification (RFID) tags using radio frequency to fall from close range to tens of meters from the tag is recognized by the technology of giving and receiving information. Using these techniques, Logistics Information Logistics to handle the flow of accurate and timely information can contribute to a significant improvement [1]. RFID-tagged products at the factory and forward it to the consumer in the course of a number of existing logistics nodes to manage the flow of human resources that have been served by the use of RFID reading device, the light and computing technology to automate this process by using economic and social effects can be obtained potential power. In this context, logistics management services based on ubiquitous IT environment, focusing on the core of the many services that can be said to belong to the service area. Thus, the introduction of ubiquitous technology in the field of logistics in the steel industry, with some merits at minor degree and although some advantage is very small level.

Steel industry, particularly the weakness of RFID exist simultaneously, steel columns and metal products caused technical problems in the practical application. In addition, the costs of tags, tag attachment and damage costs are one of the challenges to be solved. Currently the most widely used in industry Ultra-High Frequency (UHF) band passive RFID tags associated with the introduction of the industry's biggest concern was found what and how the tag can be attached with the metal or liquid products. In general, metal Radio Frequency (RF) energy, and that reflect characteristics of the RF energy has the property to absorb the liquid. Thus, the received RF energy from the reader communicates using UHF band passive tag, the RF energy, adhesive molecules, depending on the actual tag reading rate of change factors will affect a lot. In particular, the spread of RFID application in various industrial environments, according to the order of attachment on the properties to overcome the special Tag antenna, material development, and packaging technologies are essentially required to be developed.

The purpose of this paper is to use RFID technology in the steel industry based on logistics management for the development tracking management system to build the efficient logistic tracking system. The present paper is to organize largely consists of four chapters.

2 Related Research

The related researches are the existing logistics information system of RFID technology and how it applies on what measures to be integrated with the proposed research, to the RFID to existing processes are identified by the practical research [2]. Firstly, RFID technology, logistics information system integration with existing research on the efficiency of distribution has been studied mainly in terms

of research. RFID into existing processes and to apply the research to the RFID technology, several empirical studies have been performed [3]. Of these empirical studies, particularly the research on logistics management of the track in the tracking system was applied to a SCM distribution center for the research and studies on tracking technology [1, 4–6].

In particular, include the tracking of tagged materials on the recognition rate and recognition rate across case studies and RFID systems for the improved reliability technological study was conducted [7]. In addition, the UHF band passive tag that provides a macro in box, pallet, truck loads in a study on the application [8], the increased systems reliabilities of antennas, readers, tags, using the research [9, 10], and GPS or mobile devices using RFID for tracking the application of researches [11, 12]. However, these studies were found hardly applicable for to small-and medium-sized steel industry.

3 Design of the RFID-Based Tracking System

3.1 RFID Systems in the Steel Industry

RFID system that includes components of the reader antenna, a radio resource transmission and with reception antenna for storing information and data exchange protocol, tags, and including server and network. RFID tags are composed with electronic circuits and electronic components, including antenna. Tag stores mounted on the circuit to define the information of user. This information is exchanged via an external antenna. In addition, the tag antenna and the transceiver which acts as intermediary between the propagation as enable or disable the tag sends a signal and plays the role of reading and writing data. These tags are separated basing on the presence or absence of power supply active tags and passive tags are separated. First, the active type that requires power to reduce the power needs of readers with advantage of being able to recognize the distance that has shut the other hand, because it requires a power supply of working hours are limited, because of expensive compared to passive with disadvantage.

Meanwhile, the passive tag reader, without the power supply is activated by the electromagnetic field. Compared with the active type passive type is lighter, cheaper, and can be used indefinitely, but the passive type has disadvantages for reader of short distance from the reader more aware of power consumption. For this reason, the passive tag when it is sent for a long time and often requires transmission, data storage is mainly used when there is no limit on. Readers of the electromagnetic information into meaningful information are the ability of interpretation. The host computer interprets the information received and stored, or made into meaningful information, enterprise application systems need to or are linked to other business systems.

The host computer interprets the information received and stored, or made into meaningful information, applied or linked with other business systems. The steel industry to apply the RFID weakness existing both of steel columns and metal, technical problems caused on the application. In addition, the price of tags and tag attachment and damage are the challenges to solve problems. Currently the most widely used in industry in the UHF band passive RFID tag, the industry's biggest concern at the introduction of metal or liquid form forwarding for the product tagging.

In general, metal has reflective characteristics of the RF energy, while the liquid absorb the RF energy. Thus, the received RF energy from the reader communicates using UHF band of passive tag, the RF energy, adhesion molecules, depending on the actual tag reading rate of change factors will affect a lot. Application of RFID in various industries, especially the spread of the environment in order to overcome the change of properties with attachments an antenna applicable for special purposes. However, the development, packaging technology necessary to develop are at level of primitive stage.

3.2 RFID System for Tracking Module

RFID tracking management system based on the basic functional modules, RFID integration module and the tracking support module are shown in Fig. 2 shows the modules of these systems. As shown in the picture of Fig. 1 first the basic functional modules of the basic features of a tracking system make possible to track the stock of required data, delete, modify, view and features.

And RFID integration module works with the default function module reads the RFID reader and tag, the tag information stored in the DB which has a function of the location tracking. Support module to track the relevant sector information on the network, the instructions and information, and factory information, directions, check real-time information, and make the factory. The module configurations for the system are shown in Fig. 2.

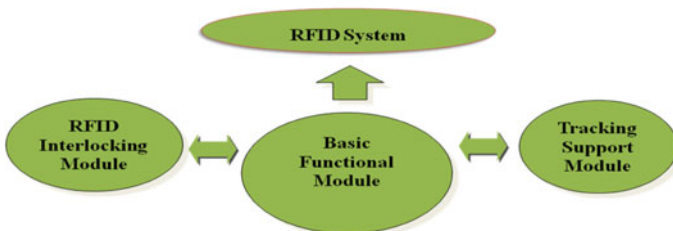


Fig. 1 RFID system module

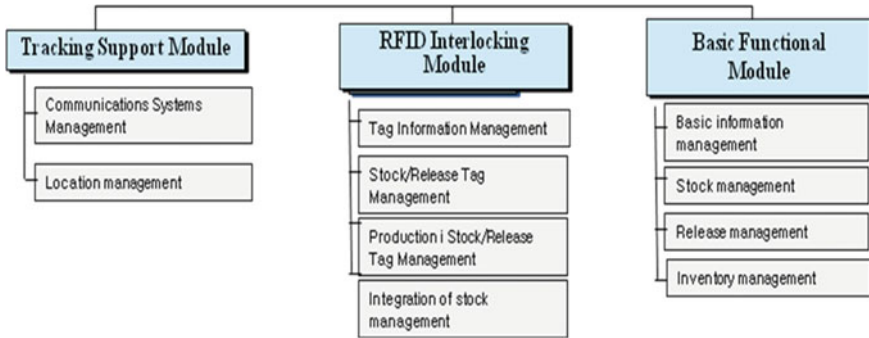


Fig. 2 RFID system structure

As shown in the picture of an RFID system the basic functional modules based are consisted of information management, stock management, release management and inventory management.

3.3 Design of RFID Systems for Tracking

RFID systems for tracking of small and medium-sized steel companies organized by the most basic features, and location tracking of plant and loading will be the basic design. System and in-plant location at the entrance to the factory management, factory floor to the four sections, separated by the storage of materials or products for easy configuration and location were identified. Firstly, the two entrances in plant material stock release-only, the command fixed here by installing the RFID antenna and reader through the materials and products can be managed. And the plant location by installing a removable antenna RFID of PDA reader designed for enable to track management.

In this paper, the steel industry in the design of RFID systems for tracking by default steelmaker, adopt reflecting the characteristics of small and medium-sized and short time to build a system to be applied for to minimizing the burden of your investment. To provide to help to secure to cut failure of in application of the new system. In addition, the failure to apply the new system a chance to maximize effectiveness while minimizing the effect was to help. Currently, the most of small and medium size steel companies barcode location management is under application for the workforce through inventory management and the reality. That Excel by hand through the management and administrative personnel are loaded due to errors and the management of the product has several problems including problems. To solve these problems by using RFID technology in the area of materials and management of the final product will be able to obtain significant effects. It is possible to pass by using RFID-plant and product loading area when

the achievement of real-time location management using RFID at each location area as to whether a product is wrapped in real-time management information and giving information of the products loaded or forklift.

4 Implementation of Trace-Based RFID System in the Steel Industry

RFID system for tracking the entire process has three detailed processes. Purchase orders of the department based on information that is directed from stock to stock and stock of materials and processes, production department, the production of information based on the release of the product from the picking instructions, release to the release process is composed. Finally, including the location process is managed is the planned location for the tag attached to the load.

In detail, we first order of the purchasing department and production department based on the information put together instructions to get off the material, as directed by Stock materials arrival material has been properly inspected and stocked wearing. RFID reader with an antenna fixed to the perception as to update the DB. And go through the production process of completion, production, production planning based on the finished product will be shipped by the factory instructions. Factory based on the instructions from any location, such as how much the factory to factory release the plan. And when release the factory are recognized as fixed RFID reader will update the DB. And, release products based on the instructions of the product placement location in one location and determine how many yards. To four locations during the assigned area of the section is in the curved form. The following figure shows the entire process of these systems (Fig. 3).

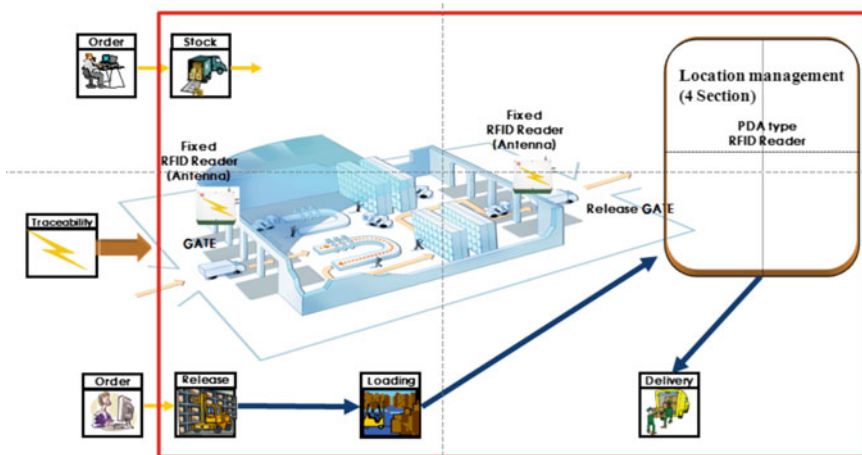


Fig. 3 RFID system process

Table 1 System testing

Classification	Percentage of recognition	
	Width (%)	Length (%)
Stocked/Released	100	100
Stocked location	99	99

Finally, for those tracking the performance of RFID systems we tested fixed reader installed in the 10 m distance is based on recognition, PDA mobile readers use expression when measured by 2 m interval. First, in order to measure the recognition rate—exit—exit the gate of an installation and teller machines were tested with the same tag testing performed on the mouth. Location on the product coming out process when loading a package of products loaded at the factory in the form of recognition from the state for the package was measured. Measurement conditions were used 100 metal tag attached, portable reader, the maximum output power (30 dBm), while the behavior was measured recognition of the tag. The table below shows the measured results from such a system. Measurement results, the pass was secured when the recognition rate of 100 %, when the product load ratio of 99 % was recognized (Table 1).

5 Conclusions

RFID systems in terms of logistics management in near future, our entire lives will be in a major impact. However, divers’ applications are different from the propagation, environment, the development of a system for each environment are needed. In particular, the steel industry in this paper to target the ubiquitous technology in the field of logistics has been introduced with some minor degree of advantage. In steel industry, especially the nature of RFID, there is weakness at the same time, steel columns and metal products caused some technical problems in the cause of the application. In this study, basically small and medium-sized steel industries, reflecting the characteristics of the building systems and by allowing for faster application time, while minimizing the capital investment.

Acknowledgments Funding for this paper was provided by Namseoul university.

References

1. Jaselskis EJ, El-Misalami T (2003) Implementing radio frequency identification in the construction process. *J Construction Eng Manage* 129(6):680–688
2. Chang Y, McFarlane D (2004) Supply chain management using auto-ID technology. In Chang

3. Makatsoris C, Richards H (2006) Evolution of supply chain management: symbiosis of adaptive value networks and ICT. Kluwer Academic Publisher, Boston
4. Kallonen T, Porrás J (2007) Embedded RFID in product identification. The proceedings of the 5th workshop on applications of wireless communications. Lappeenranta, Finland, pp 11–19
5. Garcia A, Chang Y, Valverde R (2006) Impact of new identification and tracking technologies on a distribution center. *Comput Industrial Eng* 51:542–52
6. Inoue S, Hagiwara D, Yasuura H (2006) Systematic error detection for RFID reliability. In: Conference on Availability, Reliability and Security (ARES 2006), pp 280–286
7. Furlani KM, Pfeffer LE (2000) Automated tracking of structural steel members at the construction site. Automation and robotics in construction XVII, symposium proceedings (ISARC 2000), Taipei, Taiwan, pp 1201–1206
8. Akinci B, Patton M, Ergen E (2002) Utilizing radio frequency identification on precast concrete components—supplier’s perspective. The nineteenth international symposium on automation and robotics in construction (ISARC 2002), Washington, DC, pp 381–386
9. Yagi J, Arai E, Arai T (2005) Parts and packets unification radio frequency identification application for construction, *Automat Constr* 14(4):477–490
10. Navon R, Goldschmidt E (2005) Monitoring labor inputs: automated-datacollection model and enabling technologies. *Automat Constr* 12:185–199
11. Ergen E, Akinci B, Sacks R (2007) Tracking and locating components in a precast storage yard utilizing radio frequency identification technology and GPS. *Automat Constr* 163: 354–367
12. Lindsay J, Reade W (2003) Cascading RFID tags. <http://www.jefflindsay.com/rfid3.shtml>

Performance Evaluation of Train Propulsion Control on Ethernet Network Using TrueTime

Hyeon-Chyeol Hwang, Yong-Kuk Oh and Ronny Yongho Kim

Abstract Because of its convenience and environmental benefits, public transportation like railway is getting more attention. With the growth of data traffic generated by smartphones, laptop computers and multimedia devices, in railway, there has been increasing demand for data services such as passengers' Internet access and surveillance video data transmission. In order to meet such increasing demand for data communications, International Electro-Technical Commission (IEC) has been standardizing new Ethernet-based Train Communication Network (TCN) standards. Since Ethernet is not able to guarantee fixed and sustained data rate for a certain service class due to the opportunistic nature of random access, communication schemes in the new Ethernet-based TCN standards are required to be carefully designed in order to support time-critical control applications such as propulsion and braking controls. In this paper, extensive study on the train propulsion control performance over Ethernet networks is provided. Since control system and network system need to be studied jointly in this study, co-simulation is implemented using TrueTime. Through extensive co-simulations, the train propulsion control performance on Ethernet is evaluated with the variety of interfering background data traffics.

H.-C. Hwang · Y.-K. Oh (✉)

Korea Railroad Research Institute, 176 Cheoldo Bakmulgwan-ro, Uiwang,
Gyeonggi-Do, South Korea
e-mail: hchwang@krii.re.kr

R. Y. Kim

Department of Railroad Electrical and Electronics Engineering, Korea National University
of Transportation, 157 Cheoldo Bakmulgwan-ro, Uiwang, Gyeonggi-Do, South Korea
e-mail: ronnykim@ut.ac.kr

Keywords Train communication network · Ethernet network · Train propulsion control · TrueTime simulation tool

1 Introduction

Electronic devices in railway vehicles provide various functionalities such as propulsion and braking controls, vehicle condition monitoring, on-line diagnosis and passenger information services, etc. Since electronic devices are distributed across a train, they are required to be efficiently connected in order to interoperate accurately. International Electro-Technical Commission (IEC) standardized Train Communication Network (TCN) [1]. By utilizing the TCN standards, required number of wire-lines for electronic devices' interconnection can be reduced leading to efficient information sharing among variety of devices. Therefore, electronic devices can be interconnected with minimum work load and efficient control can be provided. Currently, TCN based on Multifunction Vehicle Bus (MVB) and Wire Train Bus (WTB) is able to support data communication of only 1.5 Mbps which are mainly used for control applications such as propulsion and braking controls. In order to accommodate increasing service demands for data communications such as passengers' Internet access and Closed Circuit Television (CCTV) connection and surveillance video data transmission, IEC has been standardizing new Ethernet-based TCN standards [2–4]. Even though MVB and WTB are able to guarantee fixed data rates for time-critical control applications, since Ethernet is not able to guarantee fixed and sustained data rate for a certain service class due to the opportunistic nature of random access, communication schemes in the new Ethernet-based TCN standards are required to be carefully designed in order to support time-critical control applications such as propulsion and braking controls.

In order to provide stable train control over Ethernet, available capacity for other train applications while preserving required QoS parameters for train control related traffics should be properly and dynamically adjusted. In this paper, in order to provide guidelines for QoS guaranteed train control over Ethernet, extensive study on the train propulsion control performance over Ethernet networks is provided. Since control system and network system need to be studied jointly in this study, co-simulation is implemented using TrueTime Toolbox [5]. Through extensive co-simulations, the train propulsion control performance on Ethernet is evaluated with the variety of interfering background data traffics.

The remaining part of the paper is organized as follows. [Section 2](#) introduces Ethernet-based TCN and [Sect. 3](#) presents system model and simulation model of train propulsion control on Ethernet Network. Simulation setting and system performance analysis using TrueTime toolbox are also provided in [Sect. 3](#). Finally, [Sect. 4](#) draws some conclusions.

2 Ethernet-Based TCN

The general TCN architecture is a hierarchical structure with two network levels, Train Backbone level and Consist Network level. This hierarchical structure specifies Consist Networks based on different technologies such as MVB, CAN-open, etc., interfacing one Train Backbone. When the TCN uses Ethernet technology, Train Backbone and Consist Network can be defined as Ethernet Train Backbone (ETB) and Ethernet Consist Network (ECN). ECN of different designs and implementations may be interfaced with the same ETB which assures interoperability among Consist Networks of different types. The hierarchical architecture of TCN over Ethernet is shown in Fig. 1.

An ECN may be configured with one or more vehicles. Once an ECN is configured, its structure and topology are not typically changed. In an ECN, there are one or more Ethernet Train Backbone Nodes (ETBNs) that are used to set up communications links among ECNs of the same train. There may exist one or more ECNs within a single vehicle. The TCN is set up by two types of communication devices: Network Devices (NDs) and End Devices (EDs). NDs are primarily used for user data transmission and forwarding. There are two kinds of NDs: passive components and active components. Examples of passive NDs are cables and connectors. Examples of active components are repeaters, bridges, switches, routers and application layer gateways. EDs typically performs as sources and sinks of user data. Examples of EDs are controllers, display devices and sub-systems.

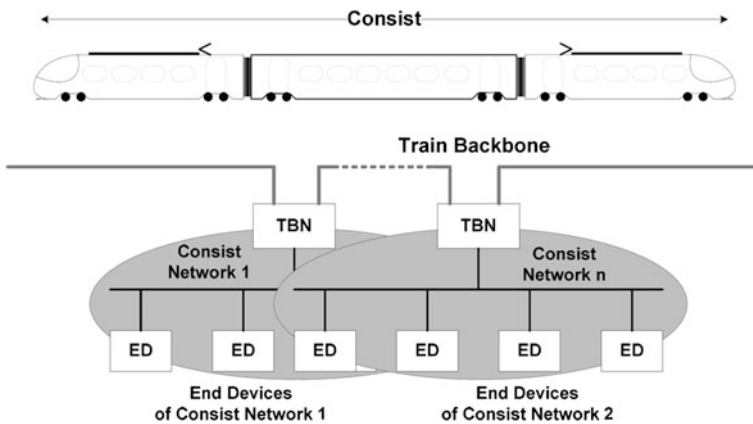


Fig. 1 Hierarchical architecture of TCN

3 Train Propulsion Control Over Ethernet Network

3.1 System Model

Following the Ethernet-based TCN standards described in previous section, the train propulsion control system model in Fig. 2 is considered for performance evaluation in this paper. Train propulsion controller receives a notch command (acceleration level) generated by driver in driver cabin, and translates the acceleration level into a torque level. The torque level is an input parameter to a train control algorithm as a reference signal. Then, a control signal is generated to minimize difference between a reference signal and a motor output torque, and transmitted to a traction motor through Ethernet network. A motor output torque measured by a sensor is also transmitted to a train propulsion controller through Ethernet network. 100 BASE-Tx fast Ethernet is standardized as ETN/ECN [3, 4], and its transmission bandwidth is shared by train control application services such as propulsion, braking, door control, etc. and train infotainment services such as Passenger Information Service (PIS), CCTV footage monitoring, passengers' Internet service, etc.

Since 100BASE-Tx fast Ethernet is able to support approximately 60 times faster transmission rate than existing MVB (maximum transmission data rate of 1.5 Mbps), it is anticipated to accommodate not only control applications but also other train applications. Whereas MVB is able to guarantee fixed and sustained data rates using dedicated channels for time-critical control applications, Ethernet is not able to guarantee fixed and sustained data rates due to its opportunistic nature of random access. In order to provide stable train control over Ethernet, available capacity for other train applications while preserving required QoS parameters for train control related traffics should be properly and dynamically adjusted. In the following section, the train propulsion control performance on

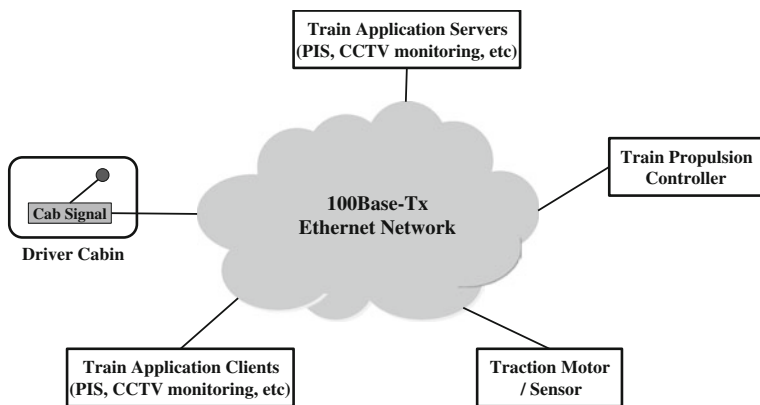


Fig. 2 System model for ethernet-based train propulsion control

Ethernet network is evaluated with variety of interfering background data traffics in order to accurately estimate required QoS parameters for time-critical propulsion control application.

3.2 Simulation Environments

The temporal non-determinism introduced by Ethernet network in the form of delays and jitter can lead to significant control performance degradation. In order to accurately evaluate the train propulsion control performance on Ethernet network, computer-based control system performance and data transmission system performance over Ethernet network should be jointly evaluated through co-simulated between control and communication systems. In order to achieve such objective, co-simulation model for Ethernet-based train propulsion control is implemented using Truetime toolbox [5]. The TrueTime is a simulation platform based on Matlab/Simulink for real-time control system and it is developed by M. Ohlin, D. Henriksson and A. Cervin, etc. of Sweden Lund University. This toolbox provides functions for co-simulation among controller task execution in real-time kernels, network transmissions and continuous plant dynamics.

Employed simulation parameters are listed in Table 1. Propagation delay and processing time are ignored because values are negligible for large network delay. Both train propulsion control signals and sensor signals are sampled at every 1 ms and those samples are transmitted with Ethernet frame of 256 bits. In the simulation, Interior Permanent Magnet Synchronous Motor (IPMSM) is considered as a traction motor. IPMSM has recently been adopted as train traction motor because it can meet the performance requirements of train operation such as high torque at low speed, high speed and static output at wide range of speed.

In the simulation, several CCTV monitoring application servers of 384 kbps and PIS servers of 256 kbps in Ethernet network generate interfering data traffics. Interfering background data traffics could be aggregated in various combinations of CCTV and PIS services resulting in 0/61/71/92/96/100 Mbps. Through extensive co-simulations, the train propulsion control performance on Ethernet is evaluated with the variety of interfering background data traffics.

Table 1 Simulation environments

Parameters	Value
Ethernet network	100 Base-Tx
Transmission rate of train propulsion control	256 Kbps
Transmission rate of motor torque sensor	256 Kbps
Aggregate transmission rate of interferer (train application servers)	0/61/71/92/96/100 Mbps
Motor type	IPMSM, 65 KW
Train propulsion control algorithm	PID control

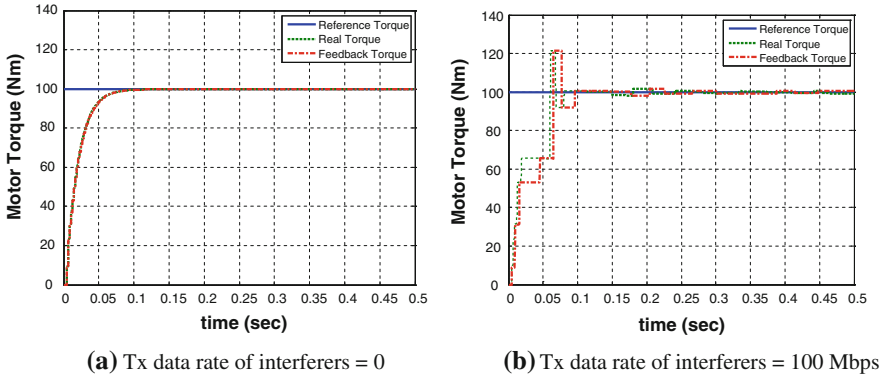


Fig. 3 Step response with respect to various transmission data rate of interferers. **a** Tx data rate of interferers = 0, **b** Tx data rate of interferers = 100 Mbps

3.3 Simulation Results

Figure 3 shows the step response with respect to various transmission rates of interferers. The interferers’ traffic causes delay in control loop of propulsion control and heavy traffic eventually causes the overshooting which usually makes system unstable.

In order to closely look at the trend of control error as the traffics of interferer increase, we define the torque error and the torque mean square error respectively as follows.

$$e(t) = y_{i0}(t) - y_i(t), \quad \text{and} \quad e_{mse} = \int_0^{\infty} |e(t)|^2 dt$$

where $y_{i0}(t)$ and $y_i(t)$ are the motor torques without interferer in the network and the motor torques with interferer in the network respectively. Interferers are defined as traffics sharing Ethernet bandwidth with the main application, propulsion control in this paper. Figure 4 shows $e(t)$ and e_{mse} at motor output with respect to various aggregated Tx rates of interferers. e_{mse} increases dramatically when interfering traffic of 92 Mbps is generated. From the simulation results shown in Fig. 4, stable control could be guaranteed under 80 Mbps traffic load (80 % of total Ethernet capacity).

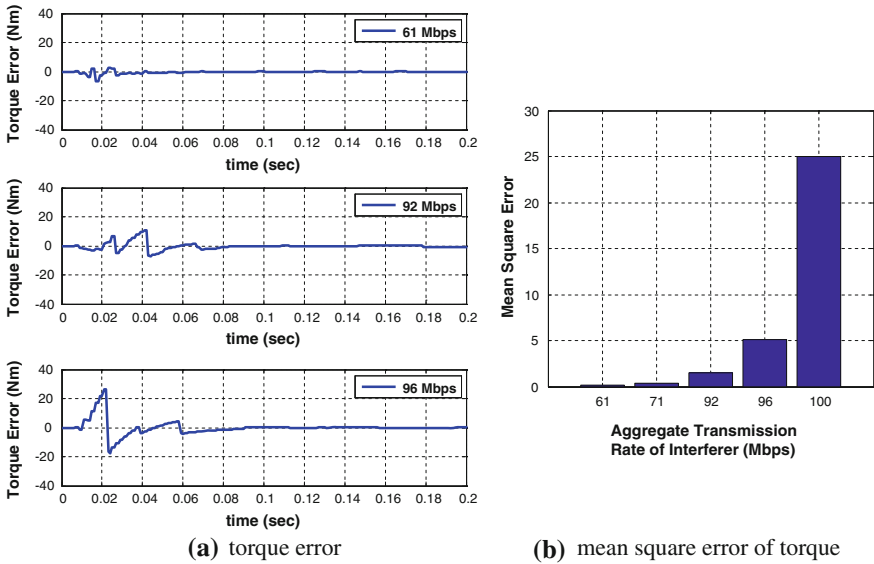


Fig. 4 Torque error and mean square error of torque at motor output. **a** Torque error, **b** mean square error of torque

4 Conclusions

Co-simulation model for train propulsion control system and network system is implemented using TrueTime Toolbox and the train propulsion control performance on Ethernet is extensively evaluated with respect to the various aggregated data rates of interfering traffic sources. From the simulation results, we can infer that train propulsion control on Ethernet is feasible as long as minimum required bandwidth is guaranteed for the train propulsion control. For future work, performance of train velocity control on Ethernet with both propulsion and braking function shall be evaluated.

References

1. IEC 61375-1 Standard (1999) Train communication network (1999): Part (1) General architecture (2) Real-time protocol (3) Multifunction vehicle bus (4) Wire train bus (5) Train network management (6) Train communication conformance testing, 1st edn.
2. IEC 61375-1, Electronic railway equipment—train communication network—part 1: TCN—train communication network general architecture, 3rd edn (under reviewed)
3. IEC 61375-2-5, Electronic railway equipment—train communication network—part 2–5: ETB—ethernet train backbone (under reviewed)
4. IEC 61375-3-4, Electronic railway equipment—train bus—part 3–4: ECN—ethernet consist network (under reviewed)
5. Cervin A, Henriksson D, Ohlin M (2010) TrueTime 2.0 beta-Reference manual. Lund University, Lund

Intelligent Control System for Railway Level Crossing Safety

Bong-Kwan Cho, Sang-Hwan Ryu, Hyeon-Chyeol Hwang,
Seoung-Chon Koh and Do-Hyeon Lee

Abstract Accident protection system at railway level crossing usually informs vehicle drivers and pedestrian that train is approaching, and eventually prevents them from passing the level crossing. This paper proposes a new level crossing safety system to solve the problem of existing system and to deal with the changing traffic condition at level crossing. The objective of the proposed system is to detect stopped vehicles at a level crossing with the state-of-the-art intelligent technology (sensor, computer, data processing, communication) and to transmit information to train allowing train drivers to stop their trains automatically and to display information of approaching train to vehicle drivers in real-time and to develop technology for accident prevention and damage reduction in connection with road traffic signal system. To evaluate the function and operation, we have demonstrated the proposed system at real railway level crossing of Young-dong line of Korail with Sea Train.

B.-K. Cho · S.-H. Ryu · H.-C. Hwang
Korea Railroad Research Institute, 360-1 Woram-dong, Uiwang-city,
Gyeonggi-do, South Korea
e-mail: bkcho@krri.re.kr

S.-H. Ryu
e-mail: shryu@krri.re.kr

H.-C. Hwang
e-mail: hchwang@krri.re.kr

S.-C. Koh
LG CNS, Prime Tower, 10-1 Hoehyeon-dong 2-ga, Jung-gu, Seoul, South Korea
e-mail: sckoh@lgcns.com

D.-H. Lee (✉)
IT Convergence Technology Research Center, Namseoul University, 21 Maeju-ri,
Seonghwan-eup, Seobuk-gu, Cheonan-city, Choongnam, South Korea
e-mail: dohyeon@gmail.com

Keywords Intelligent level crossing · Obstacle detection · Image processing · On-board monitoring

1 Introduction

The intersections between road and railroad have been increased and also level crossing danger has been increased due to the double track, electrification, speed-up of railway and increasing road traffic. According to the level crossing accident statistics, average 25.6 accidents per year from 2005 to 2007 are caused by the negligence of stop sign (43 %), barrier breakthroughs (26.2 %), vehicle faults (1.6 %), careless driving (22.3 %), and others (6.9 %). The majority of the level crossing accidents are caused by the motor vehicles (car, truck, taxi, small bus). Most of the level crossing accidents are catastrophic and the number of fatalities was 16 in 2007 and damage compensation payment per year is considerable [1].

The three elements of railway are safety, accuracy and speediness, of which the most important one is to establish safety as a public transportation system. The level crossing collision accident which comprises more than 90 % of all level crossing accidents is one of the most serious safety problems. There is a critical need for providing safe strategy and is focusing on the level crossing information rather than measures at a grade crossing [2, 3].

This paper proposes a new level crossing safety system to solve the problem of existing system and to deal with the changing traffic condition at level crossing. The objective of the proposed system is to detect stopped vehicles at a level crossing with the state-of-the-art intelligent technology (sensor, computer, data processing, communication) and to transmit information to train allowing train drivers to stop their trains automatically and to display information of approaching train to vehicle drivers in real-time and to develop technology for accident prevention and damage reduction in connection with road traffic signal system.

To evaluate the function and operation, the proposed system was set up on airport level crossing located at Yong-dong line and Sea train, and interoperation test with Sea train approaching to airport level crossing was performed 275 times for 4 months. On-board monitoring equipment, VMS equipment, and priority signal controller operated normally at this test. The result of test operation shows that the proposed system has superior effect to existing obstacle warning light drive method in view of safety and delay of train service.

2 Intelligent Railway Level Crossing System

Figure 1 shows the proposed architecture as the intelligent railway level crossing system. In Fig. 1, the intelligent railway level crossing system provides warning information to train and roadside traffic adjacent to a railway crossing, using a

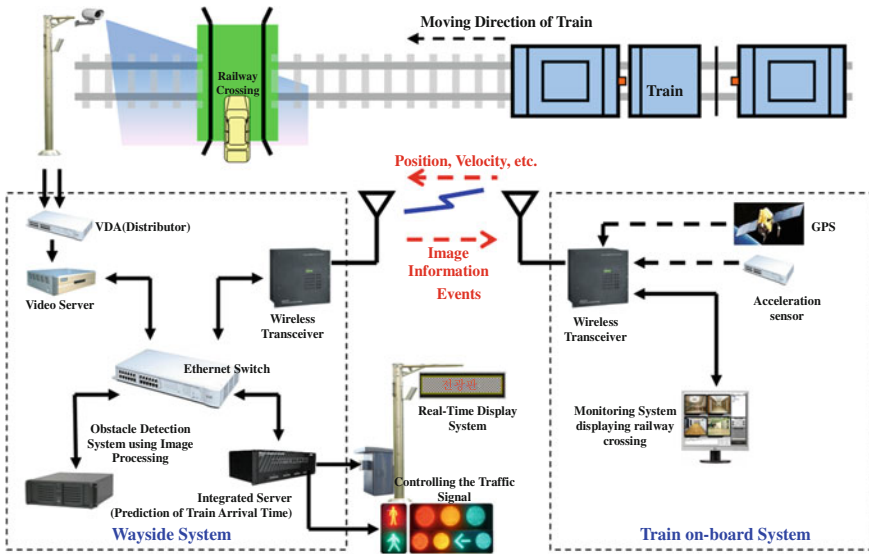


Fig. 1 Architecture of the intelligent railway level crossing system

wireless two-way communication link, for the purpose of preventing accidents and reducing damage [1, 2]. Railway crossing events (like warning messages) and video information about obstacles trapped on the crossing gate (vehicles and pedestrians, etc.) are transmitted to a train from the railway crossing, and information related to the train (direction, velocity, estimated time of arrival) is sent to the railway crossing.

The proposed system is designed to improve level crossing safety, and consist of level crossing facility, road facility and on-board facility. The level crossing facility consists of CCTV camera and obstacle detection server which collect level crossing image and detect the obstacle, level crossing integrated server which estimates train location, and wayside radio set which transfer the continuous information to the train. And, road facility consists of VMS (Variable Message Signs) which provides level crossing information to vehicle drivers on the road, and signal preemption controller which resolves road traffic congestion at level crossing. On-board facility consists of monitoring equipment which provides level crossing situation information (image, event) to train driver and processes train braking distance, GPS receiver which detects the train location, and on-board radio set.

The main components and functions of the intelligent level crossing system are as follows.

- Obstacle detection server using image processing: Obstacle detection server detects obstacle at level crossing area through image process and it determines whether obstacle exists by checking entry of object, dwell time at level crossing area.

- **Monitoring equipment:** This subsystem is installed in the cab of the train. Warning messages and real-time video of any obstacles are provided to help the train driver notice obstacles and stop the train before the level crossing. If the train driver fails to react appropriately, this system is designed to immediately deploy the emergency brake.
- **VMS (Variable Message Signs):** LED visual display device is installed between the level crossing and the adjacent road intersection to apprise vehicle drivers of accidents at the level crossing, and to provide information about approaching trains such as train moving direction, and estimated time for arrival. The VMS indicates the presence of vehicles on the level crossing and the approach of trains in real time to vehicle drivers for the purpose of reducing accidents on the level crossing.
- **Preemption signal control unit:** In urban areas, long lines of vehicles on road intersections adjacent to level crossings can occupy level crossing and cause accidents. If a level crossing is near a road intersection, the vehicles leaving the level crossing for the road intersection shall be given priority for the traffic signal. In other words, when a train is approaching, it is necessary to stop vehicles attempting to enter the level crossing and to allow vehicles already on the level crossing to quickly exit the level crossing. In this way, accidents can be prevented.
- **Integrated Server:** The integrated server estimates the time of arrival for the approaching train; information regarding time of arrival is then used by the real-time display system and the road traffic signal control unit.

3 Test Installation

The developed intelligent railway level crossing system was installed, through the cooperation with Korea Railroad Corporation, agency of Kangnung city, in driver's cabin of sea train and railway level crossing which is located near airport between Kangnung–Jeongdongjin station on Yongdong line. Henceforth, we have been verifying its performance by equipment functional and environmental test and drawing the optimization method.

Installed equipments are obstacle detection server using image processing, wireless transceiver, monitoring system on sea train, VMS (Variable Message Signs), road traffic signal control unit, and integrated server.

- **Control Cabinet and Pole Construction:** Pole, control cabinet and UPS (non-interruption electric source) were installed on the level crossing site where integration server, CCTV camera and obstacle detection system were set up. Location of pole is about 2.5 m away from rail and 2 m away from existing control cabinet. The piping excavation was executed in order to supply the power from the railway transformer and Pole/UPS base work were executed in

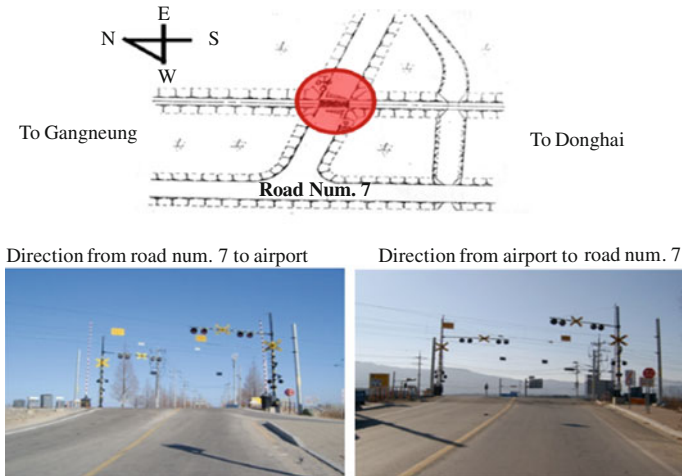


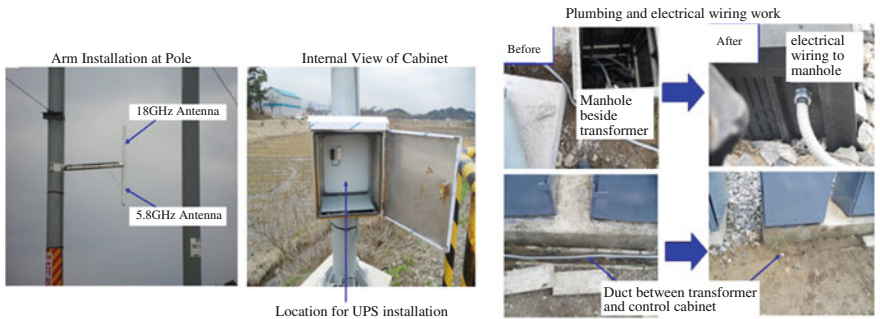
Fig. 2 Site of railway level crossing for test installation

order to install UPS. Arm structure of 1.5 m length was attached to Pole in order to install the antenna of wireless transceiver.

- Cabling Work: The UTP cable and the other cables reaching to top of pole were clearly installed inside pole in advance. Cabling work was performed after work schedule had been agreed with the person in charge of Gangwon headquarters, Korea Railroad Corporation. Control cabinet was connected with incoming line through hole under cabinet. It was locked and attached with warning sign to restrict outsider access.
- Installation of Equipments at Pole and Control Cabinet: Integration server, obstacle detection server, power supply of 18 GHz wireless transceiver, 5 GHz wireless transceiver, KVM switch, communication switch, video distribution device, power distribution device, and CCTV pan/tilt receiver were installed inside control cabinet. CCTV camera which provides images with obstacle detection server was installed on the top of pole (Figs. 2, 3).
- Installation of Bidirectional Wireless Transceiver (Ground Equipment): As shown in Fig. 4, the bidirectional wireless transceivers were installed in two places to make wireless communications possible within 2 km coverage in front of railway level crossing. The second wireless transceiver was installed on the top of pole located in railway level crossing as shown in Fig. 4. The first wireless transceiver was installed at about 1.2 km site in direction to Donghai Sea from railway level crossing (353-2, Sangsidong-Li Gangdong-myeon Gangneung City in Gangwon Province). For the first wireless transceiver, Arm mounting, control cabinet mounting, and electric wiring were also performed at existing pole.



Fig. 3 Installation of CCTV camera and control cabinet



First Wireless Transceiver (18GHz, 5.8GHz) Second Wireless Transceiver(18GHz, 5.8GHz)



Fig. 4 Installation of bidirectional wireless transceiver

- Installation of Wireless Transceiver, Monitoring Equipment, and GPS at train: High resolution GPS, 18 GHz vehicle wireless transceiver, and monitoring equipment were installed at driver’s cabin of sea train which directs toward Gangneung from Donghai, as shown in Fig. 5. Power supply of AC 220 V was provided from train electrical system.

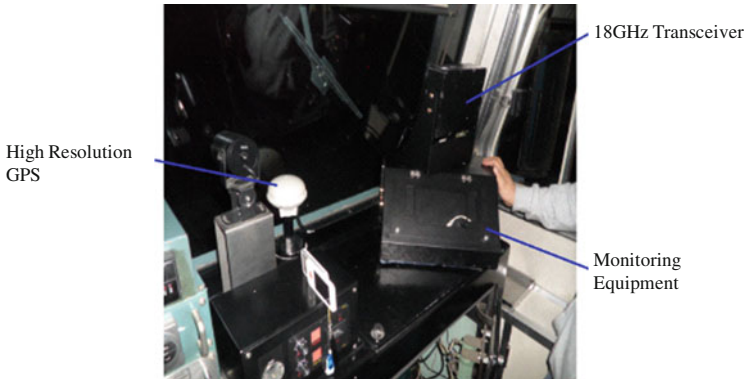


Fig. 5 Train on-board system

- Installation of Real-Time Information Display System: VMS system was installed at 200 m away from railway level crossing toward 18th combat flight airport. It helps vehicle drivers notice the situation of railway level crossing. Pole base foundation work (excavation, first class lightning grounding, 2nd class grounding of control cabinet, steel reinforcement, concrete placing, refilling, and etc.), pole installation work, and display panel installation, drawing of electric power, etc. were done.
- Approval from Public Office Concerned: The developed road traffic control unit (standard traffic signal controller (STLC-S2) 1 SET (S)) was certified by Korea Road traffic Authority. VMS system and road traffic control unit were installed with the assistance of the Transportation department, Gangneung City Hall. The use of 18 GHz wireless transceiver was permitted by Gangneung radio management office.

4 Result of Test Operation

We have tested the installed system for 35 days from 5/25/2010 to 6/30/2010. The train equipped with the intelligent system passed the railway level crossing near airport three times a day and totally about 90 times during test. Image of crossing as shown Fig. 6 was displayed to driver at train cabin. For the test period, the situation that obstacle is trapped in crossing after crossing barrier is lowered did not happen. Intentionally, we made the vehicle trapped in crossing and identified that warning signs, num. 2 and 3 of Fig. 6b were turn to red color correctly. In Fig. 6, the image of num. 1 is displayed automatically when train approaches 2 km far from crossing. Driver can cancel automatic braking by pushing num. 5 buttons when automatic braking is initiated due to obstacle detection.



Fig. 6 Monitoring System on Train. **a** Before crossing barrier was lowered, **b** after crossing barrier was lowered

5 Conclusion

In this paper, we described test installation of the developed intelligent railway level crossing prevention system which was studied by the support of the development project of damage reduction technology at railway level crossing. Interoperation between VMS system, wireless transceiver, road traffic signal control unit, obstacle detection system was checked continuously to confirm its stability during four seasons. We continuously collected the data at test installation site and upgraded the developed system. It turned out that the developed system was successfully operated and train driver was very completed.

Main features of developed system was as follows

- Obstacle detection by 1 CCTV camera at railway level crossing.
- Continuous bidirectional communication (image, control data etc.) between approaching train and ground.
- Adaptation of preemption algorithm by interface with road traffic signal control unit.
- VMS display at road side for vehicle driver.

We finally expect that the developed intelligent railway level crossing system could improve the overall level crossing control technology and reduce the accident drastically. The developed wireless transceiver can be applied to train-ground communication for control command transmission, etc. and the image processing and train tracking algorithm are also expected to be used for train detection and control.

This study could improve the existing prevention system of level crossing where many accidents occur. In case the accident occurs, the cause of accidents could be found through stored image. It would therefore be useful to clarify who is responsible for accidents. Afterward, in case of the real time monitoring system (RTMS) is planned to install, it could be utilized for supply of the real time state of railway level crossing.

References

1. Ryu SH, Cho BK, Park CH, Part JH, Gee IC (2008) Accident prevention and damage reduction technology development through intelligence of highway-railroad grade crossing. Korea Railroad Research Institute Report
2. Federal Highway Administration (2001) Intelligent transportation systems at highway-rail intersections. A cross-cutting study. Federal Highway Administration, Washington, DC
3. USDOT (2001) Vehicle proximity alert system for highway-railroad grade crossings. Final report

Moving Average Estimator Least Mean Square Using Echo Cancellation Algorithm

Sang-Yeob Oh and Chan-Shik Ahn

Abstract Echo cancellation algorithm should not only promptly adapt itself to changing environment but also minimize effects of a speech signal. However, since the color noise does not feature a consistent signal, it certainly has a significant influence on the speech signal. In this paper, the echo cancellation algorithm with a moving average LMS filter applied has been proposed. For the color noise cancellation method, an average estimator was measured by LMS adaptation filter techniques while a LMS filter step size was controlled. In addition, as it was designed to converge on a non-noise signal, the echo signal was cancelled which would, in return, lead it to the improvement of a performance. For the color noise environment, the echo cancellation Algorithm with the Average Estimator LMS filter used was applied and, a result to prove a convergence performance and stability to be improved by 10 dB comparing to the current method was gained.

Keywords Echo cancellation · Moving average estimator · Least mean square (LMS) filter · Adaptive filter · Noise cancellation

S.-Y. Oh (✉)

School of Interactive Media, Gachon University, 1342 Seong-Nam-Dae-Ro,
Su-Jeong-Gu, Seong-Nam-Si, Kyung-Gi-Do 461-702, South Korea
e-mail: coolsahn@gmail.com

C.-S. Ahn

School of Computer Engineering, Kwangwoon University, 447-1 Wolgye-dong,
Nowon-gu, Seoul, South Korea
e-mail: syoh1234@gmail.com

1 Introduction

For a smart unit which is supposed to deliver a speech through a microphone, an echo is expected along with an immediate delivery of a signal. According to a size of a space, the interference phenomenon occurs either often quickly or sometimes slowly [1]. This interference phenomenon would affect an actually-delivered signal and be a cause to lower the recognition rate. To cancel such noise phenomenon is referred as echo cancellation [2].

A noise cancellation device for the echo cancellation is used for a speech communication system to enhance noise-mixed signals. For a filter to be most frequently used for the noise cancellation is Least Mean Square (LMS) adaptive finite impulse response (FIR) filter [3].

In the process that the speech recognition system delivers a signal, which would be interfered by various noise environments that are complicated and difficult to estimate, a channel noise of the system by itself is included and this enlarges a size of an adaptive filter coefficient [4]. When it is composed of a pattern model of a speech with an incoming signal and a correlation rate of a measurable probability, a phenomenon in which a convergence time slows down by a coefficient of a pattern model with a high correlation coefficient is found increasing [5].

In order to work on such problem, this paper proposed an LMS detection Algorithm with the Average Estimator applied to LMS detection Algorithm as an adaptive filter method for the color noise cancellation. A filter with the Average Estimator applied is consisted of a linear time-invariant system and as an activated parameter threshold was controlled, an activated parameter measurement for activated parameter detection was gained. Afterwards, the acquired measurement was gone through indexing to renew the average which, in the end, not only improved the convergence speed but also confirmed the convergence performance and the stability to be increased by 10 dB.

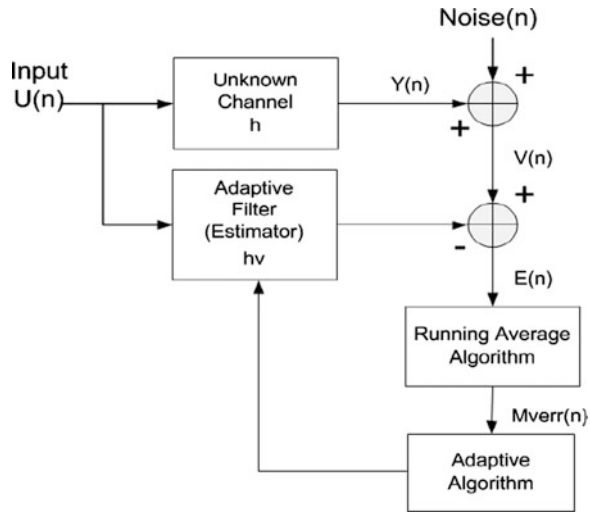
2 Echo Cancellation with a Moving Estimator Applied

In comparison with LMS, the standard detection LMS turns out to have a better convergence probability. If μ , a convergence weighting factor, gets bigger, the stability gets reduced and vice versa, if μ , the convergence weighting factor, becomes smaller, the stability becomes increased. The filter with the Average Estimator applied is consisted of the linear time-invariant system and it is presented as a following formula [6].

$$AMerr[n] = \frac{1}{L} \sum_{k=0}^{L-1} e[n-k] \quad (1)$$

The L-point Average Estimator filter calculates and prints out an average from an L-1 input of $e[n]$ per Time N.

Fig. 1 Echo canceller using moving average estimator LMS filter



In Fig. 1, an echo canceller with the proposed moving Average Estimator LMS filter used is present. The activated parameter is believed to have a greater size than the LMS adaptive echo value. The rest of the parameters of each is defined as a non-activated parameter and, for parameter detection; it is to find a location of h factor rather than 0 of m . Hence, with a cost function of Structurally Consistent Least Squares (SCLS), the detection was conducted and it is described as a following formula.

$$J_{SCLS}(N) = J_{LS}(N) + m\sigma_v^2 \log N \tag{2}$$

m indicates the number of unknown activated parameter while σ_v^2 indicates the deviation of $v(k)$.

$$J_{LS}(N) = \sum_{k=1}^N [v(k) - hvU(k)^T]^2 \tag{3}$$

For $J_{LS}(N)$ it is likely to be presented as formula (3) and to measure an Estimator value, it is used with a following formula.

$$\hat{J}_{SCLS} = \sum_{k=1}^N v^2(k) - \sum_{i=1}^m [X_{ji}[N] - \sigma_v^2 \log N] \tag{4}$$

$$X_{ji}(N) = \frac{[\sum_{k=1}^N v(k)u_{ji}(k)]^2}{\sum_{k=1}^N u_{ji}^2(k)} \tag{5}$$

Since $X_j(N)$ and $T(N)$ indicates an activated measurement and an activated threshold relatively, J_{SCLS} minimizes the $J_i = J$ index which would satisfy $X_j(N) > T(N)$. $T(N)$ as the activated threshold is described as follows.

$$T[n] = \sigma_v^2 \log N \approx \frac{\log N}{N} \sum_{k=1}^N v^2(k) \quad (6)$$

As the activated parameter threshold is controlled, an activated parameter measurement for activated parameter detection is gained and it is conducted with indexing. For the Average Estimator filter, it is likely to be gained as formula (1) is used to average $\hat{\theta}_j(k)$ a parameter vector, in k hours. It is described as follows.

$$\hat{\theta}_j(k+1) = \alpha^{1-g_j(k)} \hat{\theta}_j(k) + \mu * AMerr * g_j(k)u(k-j) \quad (7)$$

For $g_j(k)$ it is presented as a parameter of j th $g(k)$. As this is repeated, a parameter factor is analyzed. From the entered signal, noise is cancelled to converge on the signal and this would improve a convergence speed.

3 Simulation

In this paper, the echo canceller with the proposed moving Average Estimator LMS filter applied was analyzed comparing to other current system. The noise which had been used for an experimentation of the proposed method was a color noise and, as shown in Fig. 2, most of the energy appeared to be distributed throughout the frequency. For the speech DB, 445 dB produced by ETRI was used and the experimentation was carried out by 16 K, Mono and 8 K. The difference order of the proposed filter was 50 and the experimentation was conducted with a step size of 0.003. The Fig. 2 presents a wave which has been combined with color noise and a speech signal called “the middle” among 445 dB.

In Fig. 3, a signal of composite waveform was presented, in which the proposed Algorithm was applied with the incoming signal.

Figure 4 is an output signal acquired by conducting of the proposed Algorithm. It is confirmed that it has converged near on the clean speech signal. After the noise-included speech signal was filtered by the activated parameter measurement, a signal to be very much near the clean speech signal was gained.

Figure 5 provides improved results from a comparison between the color noise signal and the filtered signal. Noise of the color environment was -29 dB and, after it was filtered by the proposed Algorithm, it came up with SNR of -39 dB which was a signal improvement by 10 dB.

4 Conclusions

In this paper, the experimentation was conducted as the color echo cancellation Algorithm with the Average Estimator applied was proposed. While the Average Estimator was figured and a step size of a filter was controlled, the convergence

Fig. 2 Combined signal and energy distribution

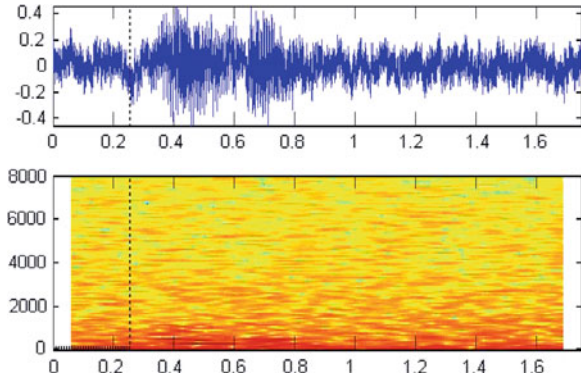


Fig. 3 Combined wave signal

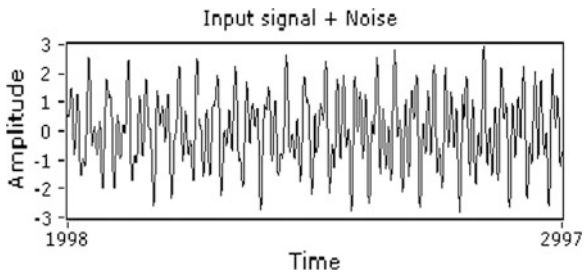


Fig. 4 Filtering signal by proposed algorithm

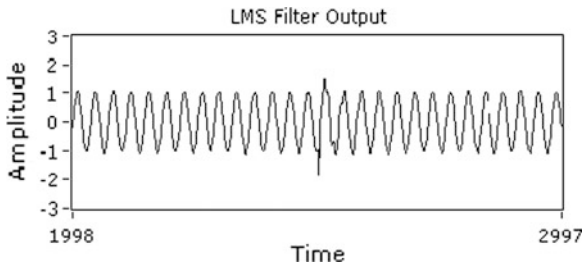
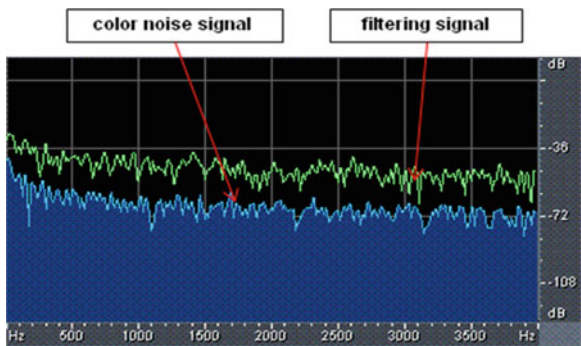


Fig. 5 Compare with color noise signal and filtering signal



performance was improved, which would cancel the color noise by adaptive filter method.

After the experimentation on the echo cancellation Algorithm with Average Estimator LMS filter applied at the color noise environment was conducted, it was discovered that the proposed detection method has a convergence performance and a stability to be improved by 10 dB than the current method does. In the end, it was concluded through the research that the proposed detection method is equipped with a more outstanding performance comparing to the current method.

Acknowledgments This work was supported by the Gachon University research fund of 2012. (GCU-2012-R168).

References

1. Homer J, Mareels I (2004) LS detection guided NLMS estimation of sparse system. In: Proceedings of the IEEE 2004 international conference on acoustic, speech, and signal processing (ICASSP), Montreal
2. Michael H (2003) Acoustic echo cancellation digital signal processing. Bachelor of Engineering thesis, the school of electrical engineering, the University of Queensland
3. Edward V (1999) Signal prewhitening schemes for detection-guided LMS estimates. Department of Electrical and Computer Engineering, University of Queensland, Brisbane
4. Gay SL (2000) Acoustic signal processing for telecommunication. Kluwer Academic Publishers, Dordrecht
5. Homer J (2000) Detection guided NLMS estimation of sparsely parameterized channels. IEEE Trans Circuits Syst II Analog Signal Process 47(12):1437–1442
6. Haykin S (2002) Adaptive filter theory. Prentice Hall, Upper Saddle River

An Abrupt Signal Detection as Accident Detection by Hamiltonian Eigenvalue on Highway CCTV Traffic Signal

In Jeong Lee

Abstract There are many limits like as shadowing occlusion and no lights in the video image detector systems. In order to make accurate detector system, we need to get rid of these problems specially accident detection system by using vehicle trace. In this paper, we introduce a method of overcoming shadow. And we propose the accurate accident detection system. We find the flow of vehicle trace is like as level spacing distribution as Wigner distribution. It is in the level statistics when we represent this vehicle trace avoidable. From this distribution we can derive a probability induced by different of position for each lane. Using this equation we can find abrupt state of vehicle flow comparing with normal flow. We shall show statistical results from some experiments for this system evaluation

Keywords Hamiltonian · Detection system · Detection abrupt signal · Calogero-moser system

1 Introduction

The goal of a traffic monitoring system is to extract traffic information, such as the vehicle volume count, traffic events, and traffic flow, which plays an important role for traffic analysis and traffic management [1]. In these extracting information, the occlusions have taken place when two or more vehicles and shadows are regarded as one vehicle because of overlapping from the viewpoint of camera.

I. J. Lee (✉)

School of Computer Engineering, Hoseo University, Sechuli 165, Baybangeup,
Asan City, South Korea
e-mail: leeij@hoseo.edu

There is explicit occlusion or implicit occlusion, the explicit case can be detected easier than implicit case because of identifying individual vehicles before occlusion. The shadows which is explicit case can cause various unwanted behavior such as object shape distortion and merging, affecting surveillance capability like target counting and identification [4].

Generally, image processing and object tracking techniques have been mostly applied to traffic video analysis to address queue detection, vehicle classification, and volume counting [5, 6]. In this case, Model-Based tracking is highly accurate for a small number of vehicles [7].

If the number of vehicle is increasing, we need new background model like as the Kalman-filter-based adaptive background model, because the background changes rapidly. In this situation, if a separate contour could be initialized for each vehicle, then each one could be tracked even in the presence of partial occlusion [2, 8]. They introduce there are three types occlusions, that is track occlusions, background object occlusions and apparent occlusions [10].

In track occlusions, in order to resolve more complex structures in the track lattice, the bounding box tracking used by appearance based modeling [9]. The appearance model is an RGB color model with probability mask similar to that used by Ismail Haritaoglu et al. [11]. In methods to solve the implicit occlusion problem in moving objects, the fusions of multiple camera inputs are used to overcome occlusion in multiple object tracking [12]. The Predictive Trajectory Merge-and-Split (PTMS) proposed to uses a multi stage approach to determining the vehicle motion trajectories and eventually the lane geometry [13, 14]. Some shadow elimination techniques have been classified in the literature into two groups, model-based and property-based technique [14]. The shadow removal approaches are based on an assumption that the shadow pixels have the same chrominance as the background but lower luminance [4, 15]. The earliest investigations in shadow removal proposed by Scanlan et al. [16], the image was split into square blocks and produced an image based on the mean intensity of each block. We know the vision-based with nighttime images was RACCOON system [17] which has been integrated into a car experiment on the CMU Navlab II, tracks car taillights. Also, another pattern classifier algorithm is Support Vector Tracking (SVT) which integrates the SVM classifier into optic-flow based on tracker [18]. The bright regions in the nighttime generated by headlights, tail lights, break lights, and reflected lights around light sources are recognized as the vehicle feature. Because of above things, in this paper, we use an advanced shadow elimination techniques for preventing occlusions, so we develop an accuracy of finding traffic information. In addition to this system, we use the vehicle trajectories for traffic accident detection system, accordingly identifying lane change patterns from the camera's field of view [2, 3]. In this case, the flow of vehicle trace is like as level spacing distribution as Wigner distribution in the level statistics when we represent this trace avoidable. From this distribution we can derive a probability induced by different of position for each lane. Using this equation we can find abrupt state of vehicle flow comparing with normal flow. For this system evaluation, we shall

show statistical results from some experiments. In Sect. 2, we introduce an advanced shadow elimination technologies. In Sect. 3, the detection method of abrupt signal is introduced on Wigner distribution.

2 Advanced Shadow Elimination Techniques

We explain the basic idea behind the tracking algorithm developed in this research. Vehicle tracking has been based on the region-based tracking approach. For individual vehicle tracking the first step, acquisition image sequences and pre-determining the detection zones at each lane. The second, we have conducted the background subtraction, deciding threshold for binary images. The background subtraction algorithm requires a relatively small computation time and shows the robust detection in good illumination conditions [19]. The third step, morphology for small particles removal as noise, sets in mathematical morphology represent the shapes of objects in an image, for example, the set of all white pixels in a binary image is a complete description of the image. The next step, it is important to remove cast shadows due to extract the vehicle area exactly, we developed the new algorithm in this paper using by edge detection and vertical projections within the vehicle particles. And the fifth step generates the vehicle ID and labeling to each vehicle, and individual vehicle's bounding rectangle data, i.e., left, top, right, bottom coordinates. These particle data are saved into reference table which can be referred to next sequence frames. In this system, the occlusion detection is easy relatively because of short length of detection zones, less than 15 m. And we have considered only limited to explicit occlusion. The explicit occlusion cases have taken place several times during the field test, that is, multiple vehicles enter a scene separately into detection zone, and merge into a moving object region in the scene. In this case, we have maintained each vehicle ID continuously as referred to previous frame.

In the nighttime, diffused reflections on the road due to vehicle headlights pose a serious concern. Thus we need to pre-processing by reflection elimination to adjust the light parameters such as luminosity or brightness, contrast, intensity. For the vehicle extraction exactly, we have to consider about background estimation, occlusion, cast shadow detection and elimination, and light condition processing at night.

Our system has covered the four lanes with single camera. As the more system has to be processed, the less performance of system has been. The reason for that if we have included the implicit occlusion process, the performance evaluation marked low grade especially the calculation of velocity. The occlusion detection of this system is easy relatively because of short length of detection zones, less than 15 m.

So many algorithms of cast shadow elimination are proposed, the various cases are occurred in the real traffic flows, for example, dark or light shadows, shadow from trees or clouds.

The proposed algorithms as mentioned before, have been applied to our experiment, the shadows cannot be extracted exactly as a result.

Thus we have developed the appropriate algorithm in our test site. The basic concept is that the shadow area has less edge because of no variance within shadow. On the other side hand, vehicle area has more edges relatively. Let B be a binary image plane and B_x be a set of number of vertical pixels which value is 1 at x . We define a function $Verti: B/x \rightarrow B_x$

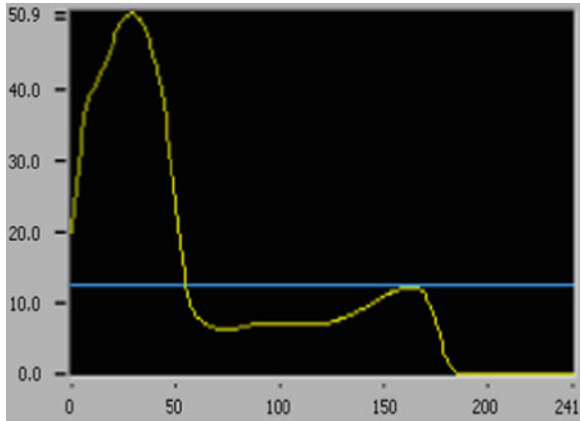
$$by\ Verti(x) = \sum_y B_1(x,y), \tag{1}$$

Where $B_1(x,y)$ is a pixel of which value is 1 at (x,y) and B/x is a projection of B into x . In Fig. 1b, the distribution of edges from moving object area can be discriminated between vehicle and cast shadow. And then discard under 25 %, that is cast shadow area, Fig. 1b.

Fig. 1 Cast shadow detection process in our test site. **a** Edge of moving object. **b** Shadow area under blue line, 25 %



(a) Edge of moving object



(b) Shadow area under blue line, 25%

3 Detection of Abrupt Signal on Wigner Level Spacing Distribution

Contents the flow of vehicle can be represented by trajectories, and a time series can be figured from this trace when the flow of vehicle volume calculated for some time interval at each lane as Fig. 2.

The distribution of position from the bottom line is Wigner distribution (2) in this time series as Fig. 3.

$$P(S) = \frac{\pi}{2} S \exp\left(-\frac{\pi}{4} S^2\right) \tag{2}$$

We can get by Hamiltonian H such as

$$P(H) = C \exp(-aTrH^2) \tag{3}$$

We know Hamiltonian

$$H(\tau) = H_0 + \tau H \tag{4}$$

has an eigenvalue.

This equation can be represented as

$$H(\tau)\Psi_n(\tau) = E_n(\tau)\Psi_n(\tau) \tag{5}$$

where τ is time series and E_n is a distance from bottom.

By diagonalization to this equation, we obtain

$$P(\{E_n\}) = \exp(- (E_1^2 + \dots + E_n^2) / (4a^2)) \prod_{m \neq n} |E_m - E_n|^v \tag{6}$$

If the Generalized Calogero-Moser system is applied to this Eq. (6) after finding the eigenvalue of Hamiltonian, we can get the Eq. (7).

$$p(x_1, x_2, \dots, x_n) = C \prod_{1 \leq i < j \leq n} |x_i - x_j|^v \tag{7}$$

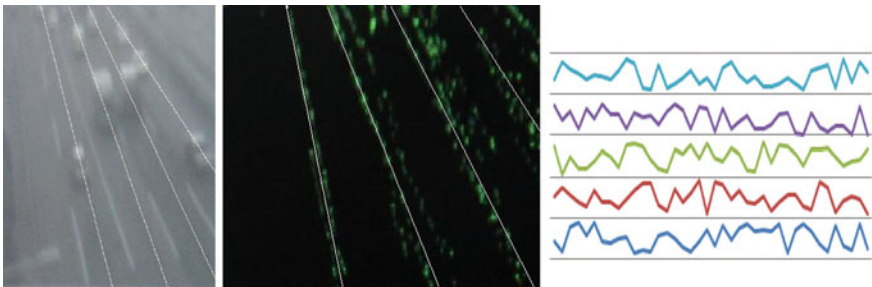


Fig. 2 Trajectories of vehicle and it's time series for each lane

Fig. 3 Wigner distribution of time series

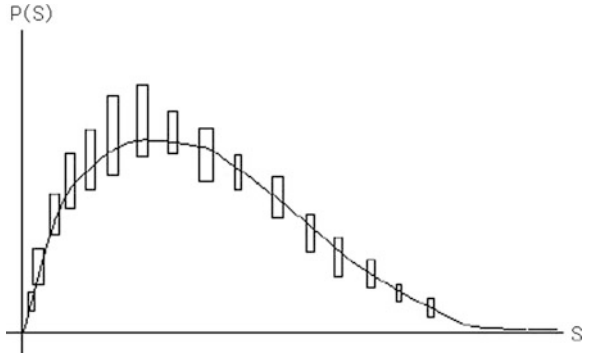
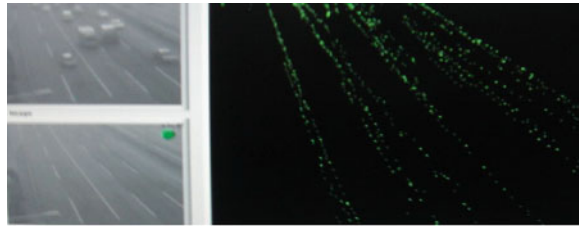
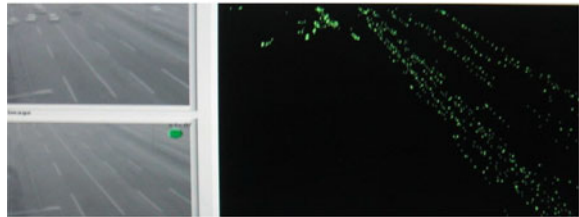


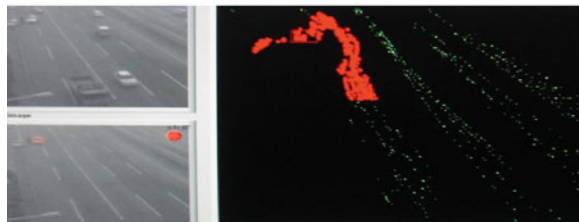
Fig. 4 Showing images from normal flow to abrupt flow and warning state. **a** Vehicle trace has no change. **b** Vehicle trace is changed. **c** Vehicle trace is changed again



(a) Vehicle trace has no change



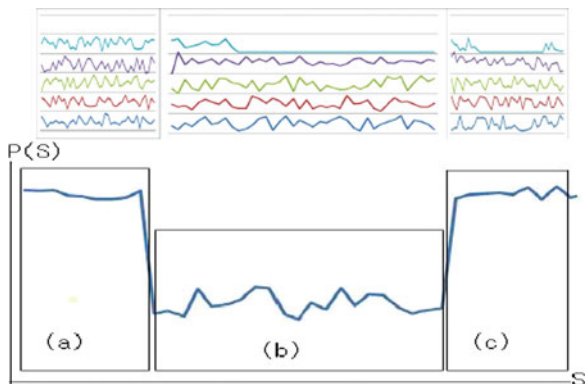
(b) Vehicle trace is changed.



(c) Vehicle trace is changed again.

Where x_i is position value of each lane. This Eq. (7) has maximum value when vehicle flow is same for every time at each lane for example each vehicle flow is same. But if one or two lane has no vehicle flow except other lane has normal flow, the value of Eq. (7) is abruptly changed. So, the detection system can be made by checking the value of (7) is abrupt value than previous value for some time. In this

Fig. 5 **a** Is an area of normal vehicle flow signal. **b** Is an area of one lane flow is stopped. **c** Is an detour area of flow



case we can choose C, ν properly for accurate precession and time interval to decision abrupt.

Since, as in Fig. 4, we can find some change of trace image (b) (c) and need to calculate abrupt signal, after that, delivered warning to manager when the calculated abrupt signal image like as Fig. 5, in this case C, ν handle the height of $P(S)$. If C increase, the noise increase and if ν increase, than normal state and abrupt state did not distinguished.

4 Experimental Result

The traffic information can be obtained by aggregating count vehicles passing through the detection zones for one minute. The more detailed measuring results, which are compared with aggregating one minute of baseline data and measuring volume counts for 30 min within each time period, are illustrated in Fig. 6 as followings.

Fig. 6 Volume counts for 30 mins within each time period

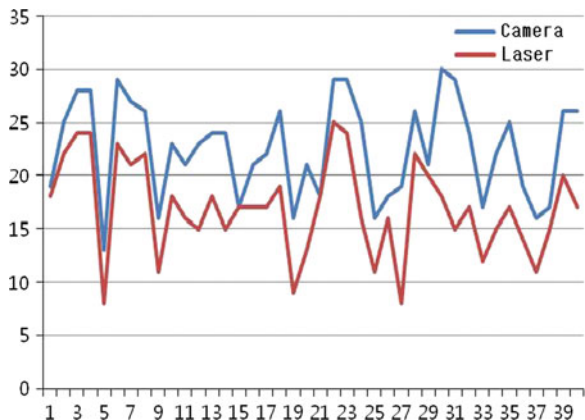
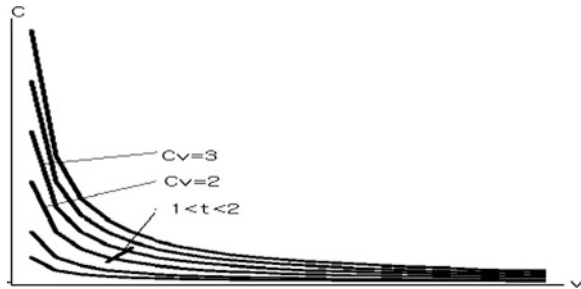


Table 1 Experimental results by changing detection time where “1” is detected case “0” is missed case

Case	1	2	3	4	5	6	7	8	9	0
0.5 min	0	0	0	0	1	1	0	1	1	1
1 min	1	1	1	1	1	1	1	1	1	1
1.5 min	1	1	1	1	1	1	1	1	1	1
2 min	1	0	1	1	x	1	0	1	0	0

Fig. 7 When $1 \leq t \leq 2$, $2 \leq Cv \leq 3$ induced from experimental data



For evaluating accident detection system, we experiments from stored moving image, because accident image is not easy searched, in this case detection time interval is changed per each experiments as shown 10 case in Table 1.

From this data we take two values as $C = 2.3$, $v = 1.2$ is properly good as in Fig. 7.

5 Conclusions

There are two issues, the one is overcoming occlusion by shadow and the other is accident detection system in this paper. The two issues are strictly related because occlusion disturbs accurate accident detection system. In order to eliminate shadow, we have developed the new algorithm using analysis of edge distribution of vehicle and shadows. If the shadow was not erased, than the volume for each lane do not calculated correctly, accordingly the accident detection is not accurate. In here, we have known the flow of vehicle trace has Wigner distribution in the level statistics when we represent each trace avoidable. From this distribution a probability was derived by representing different of position for each lane. Using this equation we could find abrupt state of vehicle flow comparing with normal flow. In this situation, the detection time interval was experimentally good for 1–1.5 min. In future works, we shall calculate the optimized detection time interval for every accident on high way.

References

1. Akio Y, Chia-Hung Y, Jay Kuo C-C (2005) Robust vehicle and traffic information extraction for highway surveillance. *EURASIP J Appl Signal Process* 14:2305–2321
2. Coifman B, Beymer D, McLauchlan P, Malik J (1998) A real-time computer vision system for vehicle tracking and traffic surveillance. *Transp Res Part C* 6:271–288
3. Oh J, Min J (2008) Development of a real time video image processing system for vehicle tracking. *J Korean Soc Road Eng* 10(3):19–31
4. Liu H, Li J, Liu Q, Qian Y (2007) Shadow elimination in traffic video segmentation. In: *MVA 2007 IAPR conference on machine vision applications*, Tokyo Japan, 16–18 May 2007
5. Chen S-C, Shyu M-L, Peeta S, Zhang C. (2003) Learning-based spatio-temporal vehicle tracking and indexing for a transportation multimedia database system. *IEEE Trans Intell Transp Syst* 4(3):154–167
6. Oh J, Min J, Kim M, Cho H (2008) Development of an automatic traffic conflict detection system based on image tracking technology. Submitted to *TRB*
7. Koller D, Daniilidis K, Nagel H (1993) Model-based object tracking in monocular image sequences of road traffic scenes. In: *J Comput Vision* 10:257–281
8. Koller D, Weber J, Huang T, Malik J, Ogasawara G, Rao B, Russell S (1994) Towards robust automatic traffic scene analysis in real time. *ICPR* 1:126–131
9. Andrew S, Arun H, Ying-Li T, Lisa B, Sharath P, Ruud B (2006) Appearance models for occlusion handling. *J Image Vis Comput* 24(11):1233–1243
10. Cucchiara R, Grana C, Tardini G, Vezzani R (2004) Probabilistic people tracking for occlusion handling. In: *Proceedings of the 17th international conference on ICPR 2004*, vol 1, pp 132–135, 23–26 Aug 2004
11. Ismail H, David H, Larry SD (2000) W⁴: real-time surveillance of people and their activities. *IEEE Trans Pattern Anal Mach Intell* 22(8):809–830
12. Dockstader SL, Tekalp AM (2001) Multiple camera fusion for multi-object tracking. In: *Proceeding IEEE workshop on multi-object tracking*, pp 95–102
13. Jose M, Andrew N, Alexandre B, Jose S-V (2004) Viewpoint independent detection of vehicle trajectories and lane geometry from uncalibrated traffic surveillance cameras. In: *International conference on image analysis and recognition*, Porto, Portugal, 29 Sep–Oct 1
14. Mei X, Chong-Zhao H, Lei Z (2007) Moving shadow detection and removal for traffic sequences. *Intern J Autom Comput* 4(1):38–46
15. Horprasert T, Harwood D, Davis L (1999) A statistical approach for real-time robust background subtraction and shadow detection. In: *Proceeding 7th IEEE ICCV Frame-rate Workshop Corfu*, pp 1–19
16. Avery RP, Zhang G, Wang Y, Nihan N (2007) An investigation into shadow removal from traffic images. In: *TRB 2007 annual meeting CD-ROM*
17. Rahul S (1993) RACCOON: a real-time autonomous car chaser operating optimally at night. In: *Proceedings of IEEE intelligent vehicles*
18. Samyong K, Se-young O, Kwangsoo K, Sang-cheol P, Kyongha P (2005) Front and rear vehicle detection and tracking in the day and night time using vision and sonar sensors. In: *Proceedings of 12th world congress of ITS*, 6–10 Nov 2005
19. Kim Z (2008) Real time object tracking based on dynamic feature grouping with background subtraction. In: *Proceeding IEEE conference on computer vision and pattern recognition*

Implementation of Improved DPD Algorithm Using the Approximation Hessian Technique and an Adaptive Filter

Jeong-Seok Jang and Gyong-Hak Lee

Abstract In this paper, Improved Digital Predistortion (DPD) Algorithm using an active filter and approximation Hessian technique is suggested. The algorithm optimized the performance of the DPD based on Quasi-Newton family method. In modeling power amplifier, the memory polynomial model which can model the memory effect of the power amplifier is used. And we compared with Least Mean-Squares (LMS) and Recursive Least squares (RLS) algorithm.

Keywords Power amplifier · DPD · Quasi-newton method

1 Introduction

The DPD technique is the linearization method using digital distortion loop. It has shown an excellence in its performance comparing with analog ones and has an advantage of miniaturization. Therefore, the power amplifier tends to be designed in small sized ones which have advanced results in their efficiency and linearity in conjunction with efficiency-oriented Class C amplifier or Switching mode amplifier.

J.-S. Jang (✉)

Navaid Div, Telemax, Uiwang-Si, Gyeonggi-Do 437-801, South Korea
e-mail: chang1022@telemax.co.kr

G.-H. Lee

School of Mechanical Engineering, Hankook University, 5-4 Sinsa-dong, Seoul, Kangnam-gu, South Korea
e-mail: Kkim@hankook.ac.kr

Generally, DPD responds quickly to the power amplifier's output variation, comes up with an adaptive filter block that makes distortion function, The most representative algorithms for DPD are Least mean-squares (LMS) and Recursive Least (RLS), etc. [1-3]. In this paper, we proposed the improved DPD algorithm by using approximation Hessian based on the Quasi-Newton family technique. Also, we verified the performance by applying the proposed algorithm on power amplifier model with the memory effect.

2 Memory Polynomial Model

The memoryless model refers to the case which the most recent output is affected by only the most recent input. This can be translated with AM-to-AM characteristic. However, in most of conventional circuit boards, the previously stored energy of the energy storing devices on the board influences on the current output with time delay. This paper considers the power amplifier's AM-to-AM and AM-to-PM characteristic using memory polynomial model.

Equation (1) refers to the Memory polynomial model of discrete signals.

$$y(n) = \sum_{k=1}^N \sum_{a=0}^Q a_{k,a} |x(n-q)|^{k-1} \bullet x(n-q) \quad (1)$$

$x(n)$: Input signal

$y(n)$: Output signal

N: The order of memory length

2.1 Adaptive Filter Algorithm

As general FIR filters or IIR filter have fixed coefficients, if we need to change the parameters of the filter by the time changes, we can not get the desired result. On the other hand, the adaptive filter can control filter coefficient in real time to obtain the closest result to the desired filter response.

Figure 1 shows the mechanism of adaptive filter block in the DPD technique. In DPD, these adaptive filter becomes signal $u(n)/G$, then ideal signal becomes input signal $x(n)$. Filter 1 makes up $w(n)$ coefficient around input signal $x(n)$ and repeats achieving new coefficient $w(n+1)$. $w(n)$ is not only an exact inverse function, but also shows similar characteristic.

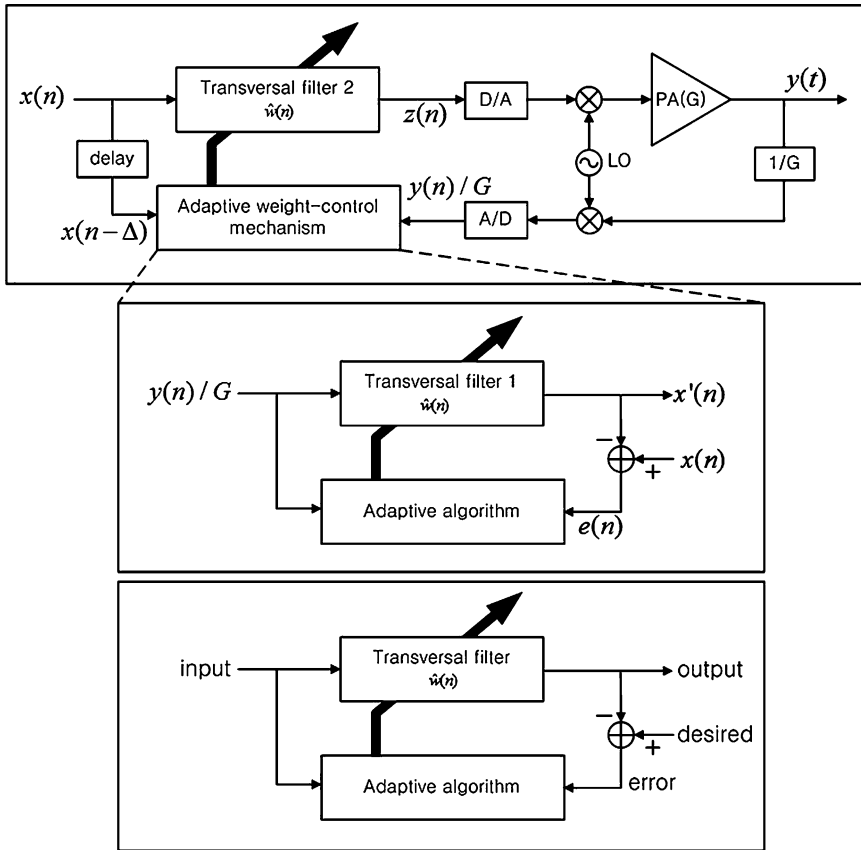


Fig. 1 Adaptive filter mechanism in indirect learning architecture

2.2 Improved Adaptive Filter Algorithm Based on the Quasi-Newton Method

This paper shows the improvement of adaptive filter algorithm in DPD based on the Quasi-Newton method tree [4]. Instead of the direct search method to directly obtain the optimal value to a specific value, we used the indirect search method to search the optimal value from the slope of phase and amplitude error. Quasi-Newton directed on Newton method consists of secondary differential information. Cost function is the optimized function of w , and it is same as Taylor series [5]. Approached diverse Hessian changes with D_n -defining new function S_n then redefined as Eq. (2).

$$w(n + 1) = w(n) + u(n)S_n \tag{2}$$

Analysing Eq. (2), $u(n)S_n = w(n + 1) - w(n)$ is defined with increase of past tap coefficient, redefined into $u(n)S_n = P_n$. Also, renewed coefficient and cost function's difference are redefined into $q_n = \nabla \zeta(w_{n+1}) - \nabla \zeta(w_n) = 2R_n p_n$, and Eq. (3) is directed.

$$D_{n+1} = D_n + \frac{p_n p_n^T}{2p_n^T q_n} - \frac{D_n q_n q_n^T D_n}{q_n^T D_n q_n} \tag{3}$$

Solving input's autocorrelation R_n into $R_n = x_n x_n^T$, becomes Eq. (4)

$$D_{n+1} = D_n + \frac{p_n p_n^T}{2|p_n^T q_n|^2} - \frac{D_n x_n x_n^T D_n}{x_n^T D_n x_n} \tag{4}$$

Deciding $u(n)$ for $\zeta(w_n - u(n)S_n)$ to renew step size $u(n)$ equals to Eq. (5)

$$u(n) = \frac{1}{2x_n^T D_n x_n} \tag{5}$$

However, Eq. (4)'s denominator renewal system is same as Eq. (6)

$$x^T H_{n+1} x = x^T H_n x + \frac{x_n^T p_n}{p_n^T q_n} - \frac{(H_n x_n^T q_n)^2}{q_n^T H_n q_n} \tag{6}$$

If this part becomes 0, this equation can't be converged and becomes radiate, so it needs positive definite [5–7]. Defining equation $a = H_n^{1/2}$, $b = H_n^{1/2} q_n$ becomes Eq. (7)

$$x^T H_{n+1} x = \frac{(a^T a)(b^T b) - (a^T a)^2}{(b^T b)} + \frac{(x_n^T p_n)^2}{p_n^T q_n} \tag{7}$$

When we look for Eq. (7), right clause gets non-negative condition by Cauchy–Schwarz inequality, and knows that the first clause doesn't disappear when a and b proportion each other. So, the equation is valid when it is $x = \beta q_n$. however this equation gets answer by a step size, $u(n)$.

$$p_n^T x = \beta p_n^T q_n = \beta u(n) \xi D_n \nabla \zeta^T \neq 0 \tag{8}$$

Therefore, $x^T H_n x > 0$ suits on any number x, instead of 0.

2.3 Simulation

In this paper, the power amplifier has been modeled with ADS ver. 2010, a RF simulation tool from Agilent Technologies Inc. The 5th degree polynomial and 3rd degree memory depth were used in this modeling method, considering memory

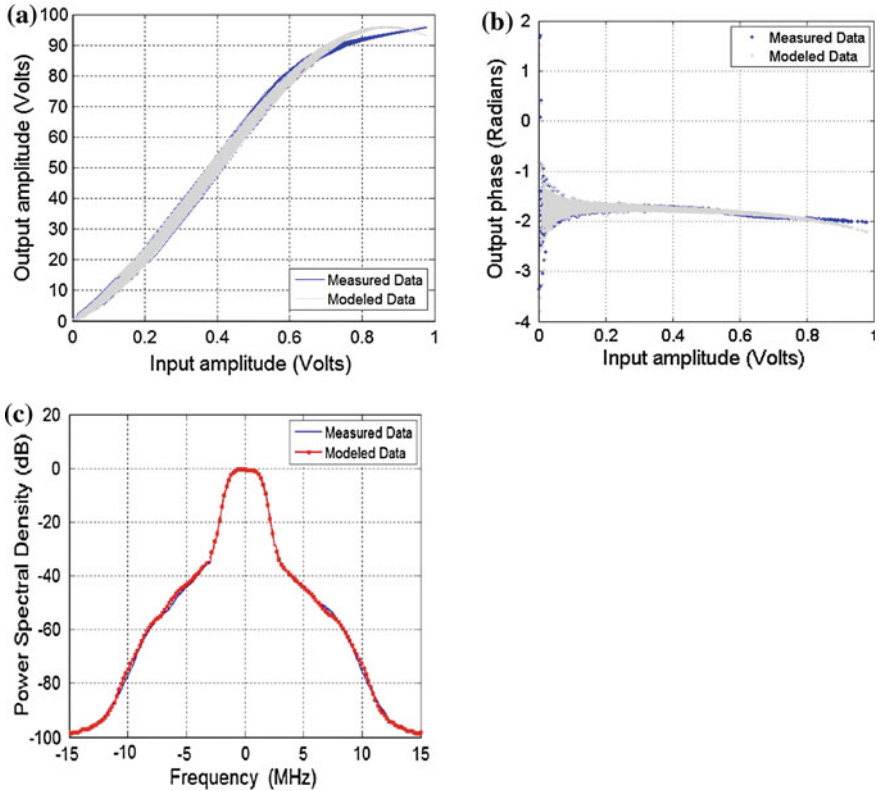


Fig. 2 The Characteristics of modeling for power amplifier. **a** Modeling power amplifier's. **b** AM-to-PM characteristic AM-to-AM characteristic. **c** modeling power amplifier's spectrum characteristic

effect. It is modeled power amplifier using WCDMA 1FA signal generator and LDMOD TR of ADS (Fig. 2).

Figure 3a is the diverse clause of AM-to-AM showing improved algorithm in this paper. Figure 3b is AM-to-PM characteristic which responds for $-1.8[\text{rad}]$ to $+1.8[\text{rad}]$ as the output undistortion.

Figure 4 shows the comparison between the application of improved algorithm using the form of Fig. 1 and LMS algorithm and RLS algorithm which have been commonly used. To compare the capacity of each algorithm, we measured Adjacent Channel Leakage Ratio (ACLR) from ± 5 MHz offset. Improved algorithm shows -72 dBc of ACLR characteristic (Table 1).

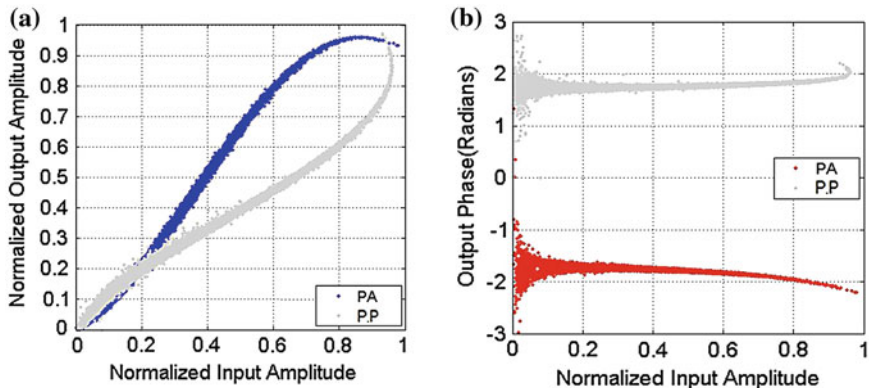


Fig. 3 Inverse function of proposed algorithm. **a** AM-to-AM characteristic. **b** AM-to-PM characteristic

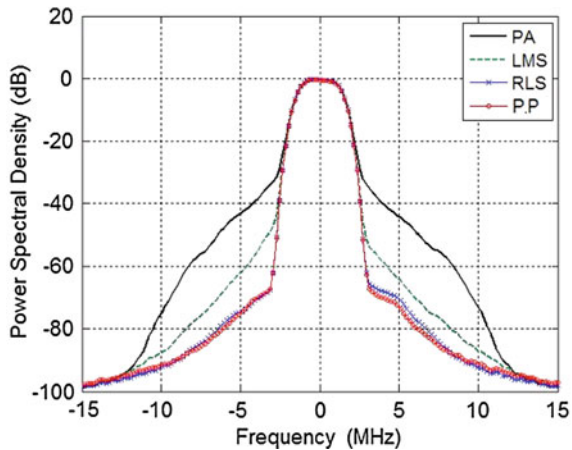


Fig. 4 Comparison of linearization results

Table 1 Comparison of linearization of each algorithm

	Lower (dBc)	Upper (dBc)	Improvement (dB)
PA only	-45	-45	-
LMS	-61	-63	About 16
RLS	-70	-70	About 25
Proposed	-72	-72	About 27

3 Conclusions

This paper proposes an improved DPD algorithm using approximate Hessian technique. This embodied algorithm guarantees amount guarantee using reverse approximate Hessian technique and is optimized to obtain better performance comparing with conventional algorithms. As a result, the linearity improvement shows the efficiency of approximately 27 dB, which is the similar performance with RLS algorithm.

References

1. Jian W, Yu C, Wang J, Yu J, Wang L (2006) OFDM adaptive digital predistortion method combines RLS and LMS algorithm. In: IEEE conference on industrial electronics and applications, pp 3900–3903
2. Yu X (2008) Stability enhancement of digital predistortion through iterative methods to solve system of equations. In: Microwave conference APMC 2008, pp 1–4
3. Eun C, Powers EJ (1997) A new Volterra predistorter based on the indirect learning architecture. *IEEE Trans Signal Process* 45(1):223–227
4. Rao SS (2008) *Engineering optimization theory and practice*, 4th edn. Wiley, New York
5. Jang JS, Choi YG, Suh KW, Hong US (2011) A design of new digital adaptive predistortion linearizer algorithm based on DFP method. *J. Korean Inst Electromagn Eng Soc* 25(3):312–319
6. Haykin S (2008) *Adaptive filter theory*, 4th edn. Prectice Hall Inc., New Jersey
7. Luenberger DG, Ye Y (2008) *Linear and nonlinear programming*. Springer, New York
8. Sun W, Yuan Y-X (2007) *Optimization theory and methods (Nonlinear Programming)*. Springer, New York

An InGaP HBT MMIC High Efficient Dual Path Power Amplifier for CDMA Handset Application

Song-Gang Kim, Hwan-Seok Yang and Seung-Jae Yoo

Abstract In this paper a high efficient dual path power amplifier (PA) for the code division multiple access (CDMA) handset applications is proposed. A dual path PA consists of two different size PAs combined parallel with single input/output matching circuits. The dual path PA is fabricated in the InGaP/GaAs hetero-junction bipolar transistor (HBT) monolithic micro-wave integrated circuit (MMIC) in the CDMA cellular, 824–849 MHz and operates at a supply voltage of 3.4 V. The dual path PA exhibits an output power of 31 dBm, a 36 % power added efficiency (PAE) at an output power, and a -46 dBc adjacent channel power ratio (ACPR) at an 885 kHz offset frequency in the high power mode and an output power of 21 dBm, a 14.2 % PAE at a 16 dBm output level, and a -49 dBc ACPR at a 885 kHz offset frequency in the low power mode. This concept is also available for the other CDMA/OFDM specifications.

Keywords Heterojunction bipolar transistor (HBT) • Parallel architecture • Adjacent channel power ratio (ACPR) • InGaP/GaAs HBT • CDMA • MMIC • Offset frequency • Power amplifiers

S.-G. Kim · S.-J. Yoo (✉)

Department of Information Communications Engineering, Joongbu University,
101, Daehakro, Chubu, Geumsan-gun, Chungnam, South Korea
e-mail: sjyoo@joongbu.ac.kr

S.-G. Kim

e-mail: kimg@joongbu.ac.kr

H.-S. Yang

Department of Information Security Engineering, Joongbu University,
101, Majeon-ri, Chubu-myeon, Geumsan-gun, Chungnam, South Korea
e-mail: yanghs@joongbu.ac.kr

1 Introduction

The high power consumption of the power amplifier is well known as a main issue in the portable handset application. The power amplifier increases the signal power level to transmit the signal from the handset to the base station [1]. A modern digital modulated signal has a high peak-to-average ratio, and the power amplifier should work at a higher power level than the average power level. For this reason, the efficiency of a PA is decreased at the high power stage. Moreover, the fluctuation of the signal power makes a spectral regrowth at the adjacent channel through the nonlinearity of the power amplifier, which could generate the interference to the adjacent channel. So the efficiency and the linearity of the power amplifier are the key design parameters. The efficiency improvement of the power amplifier could extend the operating time of the battery, and also avoid the unwanted heating problem of the handset. The commercially available handset power amplifiers have been designed to have the highest efficiency at the maximum output power level around 28 dBm. But, according to the statistical operational condition study for the CDMA signal, the probability that the PA should work at 28 dBm is only 0.01 % as shown in Fig. 1 [2]. Actually, the nominal output power required in the PA is less than 16 dBm. The power amplifier should handle the maximum power maximum output power, which causes a tremendous waste of power. This concept is also available for the other CDMA/OFDM specifications.

In order to improve the efficiency in a lower power level, the smart PA structure has been used. This technique is to control the bias level of the power amplifier according to the operating conditions. At the high power mode, it increases the bias level, and at the low power mode as shown in Fig. 2, it decreases the bias level to improve the efficiency. But in the smart PA, the efficiency improvement is not enough because the device size is still optimized to the maximum output power. The reported efficiency of the smart PA in the low power mode is about the 7–8 % PAE [3].

In this paper, a dual path power amplifier is proposed. In this approach, two different amplifiers for the high power mode and the low power mode are combined parallel through the input and output matching circuits. The high power path is activated when the output power is over 16 dBm, else the low power path is selected. The control sequence is summarized in Table 1. This method enables to optimize the device size and the bias circuit simultaneously according to the output power levels. Through this structure, the PA has been successfully developed, that shows the 14.2 % PAE of at the low power mode (16 dBm) with the same performance at high power mode.

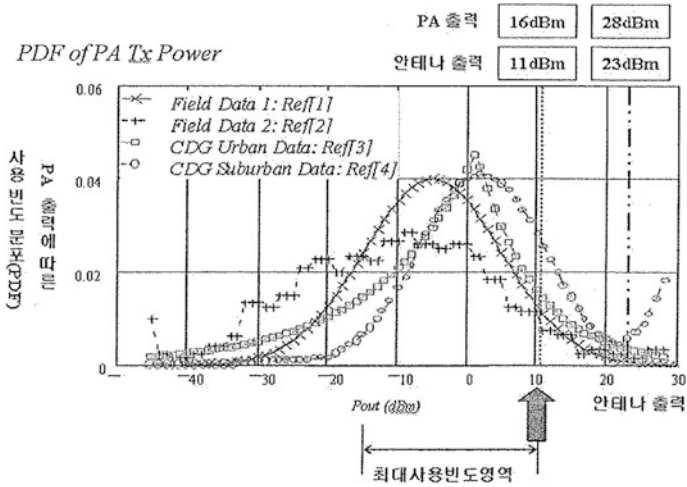


Fig. 1 Probability density function (PDF) for the output power of CDMA power amplifiers for urban and suburban profiles

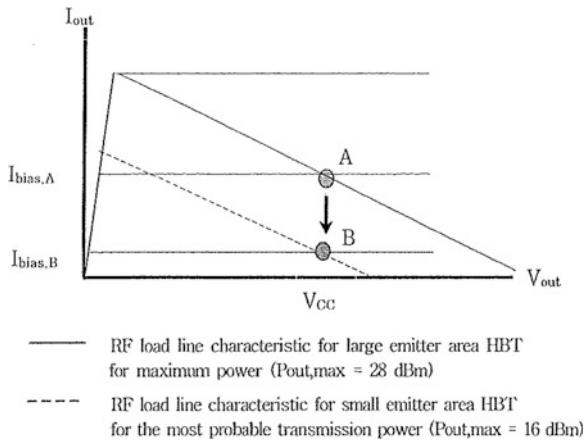


Fig. 2 Load line characteristics for the maximum output power and the most probable transmission power. Point A is the quiescent bias point for the large emitter area HBT and point B for the small emitter area HBT

Table 1 ON/OFF sequence and bias condition of the proposed PA under different operation modes

Parameter	V_{ref} (V)	V_{mode} (V)	Range (dBm)
High power mode (HPM)	2.85	Low (0)	18–28
Low power mode (LPM)	2.85	High (2085)	<18
Shut down	0	Low (0)	–

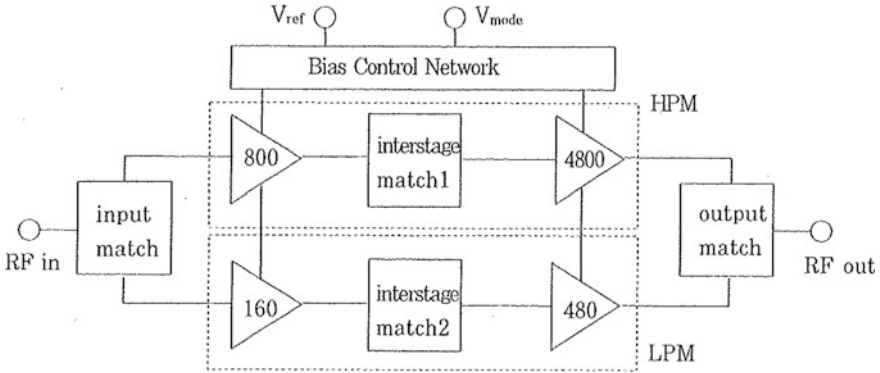


Fig. 3 Block diagram of the proposed dual path power amplifier (unit = μm^2)

2 Dual Path Power Amplifiers

A dual path PA is designed in the InGaP/GaAs HBT foundry. Fig. 3 shows the block diagram of the dual path PA. The dual path PA has two paths. The upper path is active in the high power mode and a lower path in the low power mode. In each path, the amplifier is designed in a two-stage, and the mode control voltage selects one of the high or low power modes.

The emitter area for unit cell of this foundry is $80 \mu\text{m}^2$. The emitter size of the power stage is $4800 \mu\text{m}^2$ and the drive stage is $800 \mu\text{m}^2$ in the high power mode. In the low power mode, the power stage is $480 \mu\text{m}^2$ and the drive stage is $160 \mu\text{m}^2$. The most difficult step in designing a dual path power amplifier is to combine the two paths because of the complex matching networks of the input and output. When the mode is changed low to high, the input impedance will decrease. To compensate these variations, the series capacitors are required at the base terminals

Fig. 4 Impedance of various matching points of the proposed power amplifier shown Fig. 3

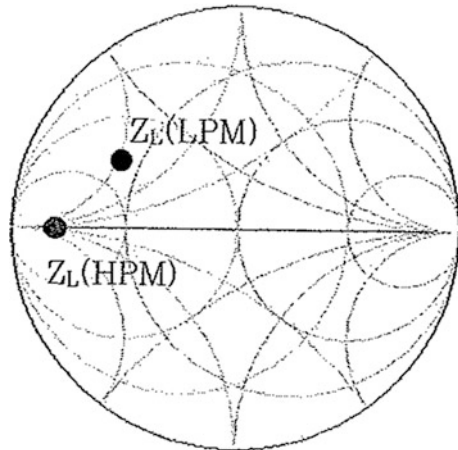
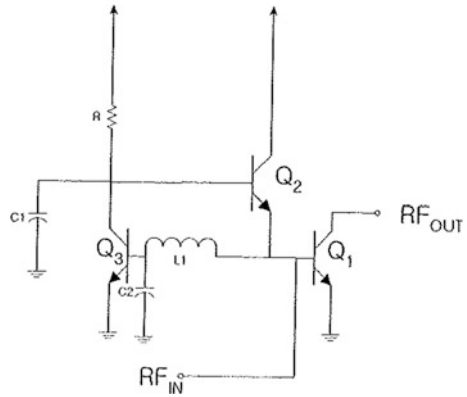


Fig. 5 Schematic of the active bias circuit with the linearizing



of each input stage, and then the additional input matching is added at the final input terminal.

Figure 4 illustrates the impedance at various output positions of the proposed PA shown in Fig. 3 the load impedance of the power stage will be determined by ZL(HPM) impedance. But in low power mode, the impedance of ZL(HPM) is transformed the optimum impedance of the low power path ZL(LPM).

For the linearity and the high efficiency, the active bias circuit using a diode has been frequently used, shown is Fig. 5, with the reference voltage of 2.85 V. The total PA chip consumes 75 mA of the quiescent current in the high power mode. But in the low power mode it only 30 Ma. The high/low power mode is controlled by the 1 bit switching transistor. The mode control signal turns the bias circuit on/off to change the modes.

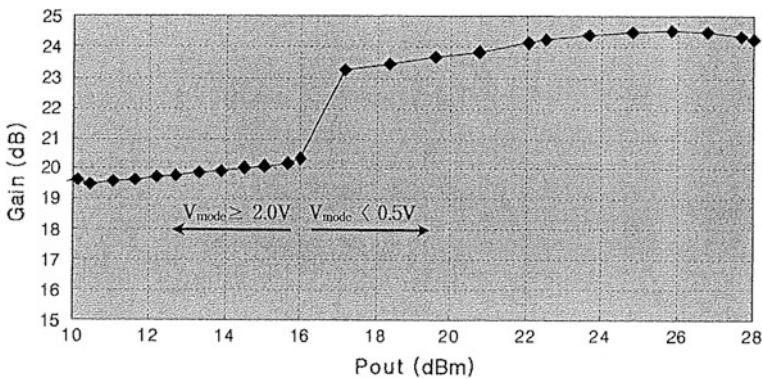


Fig. 6 The gain against output power for the dual path power amplifier in the high/low power mode

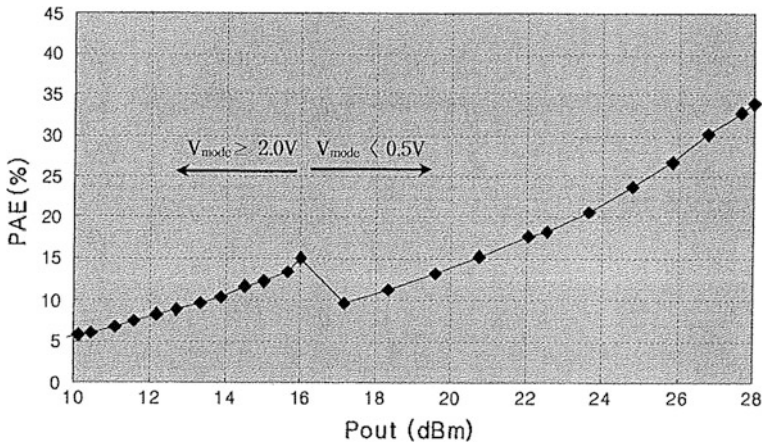


Fig. 7 The PAE against output power for the dual path power amplifier in the high/low power mode

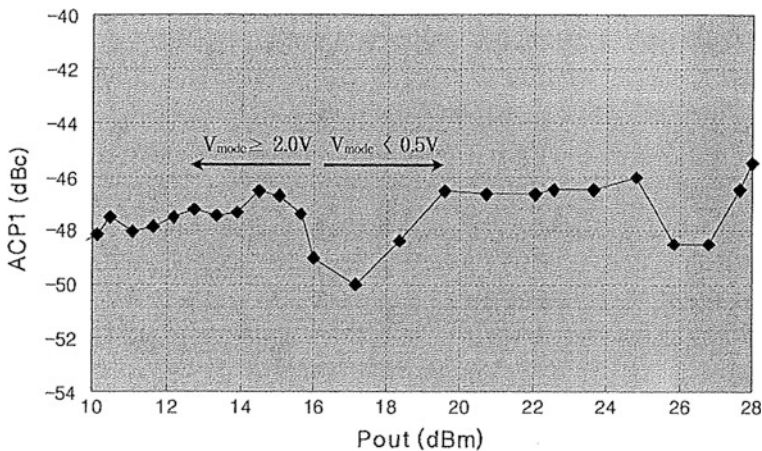


Fig. 8 The ACPR for the dual path power amplifier in the high/low power mode

3 Measurement

Within the range from 824 to -849 MHz CDMA cellular band Figs. 6, 7 shows the gain, and the PAE against output power of the dual path PA in two modes. A 1Db compression point of the amplifier is 28 dBm in the power mode and 21 dBm in the low power mode. The gain is 24.5 dB at 28 dBm output power and 14.2 % at 16 dBm output power in the high power mode and the low power mode, respectively. Considering the PAE measured at 16 dBm output power in the high power mode, it is clear that the PAE is improved much better in the low power mode.

Table 2 Summary of the measured performance of the dual path power amplifier

	Performance	
	High power mode (HPM)	Low power mode (LPM)
Operation voltage	3.4 V	3.4 V
Output power	30 dBm	21 dBm
Power gain	24.5 dB	23 dB
Power added efficiency	36 %	14.2 %
ACPR	-46 dBc	-49 dBc
Quiescent current	75 mA	30 mA

And the quiescent current is a total of 75 mA in the high power mode, and 3 mA in the low power mode. The results of ACPR test, a criterion of the linearity of the PA, show less than -46 dBc at the output power up to 28 dBm average output power in the high power mode, and -49 dBc at the output power in the low power mode, shown in Fig. 8. The measured performance is summarized in Table 2.

4 Conclusion

A high efficient dual path power amplifier for the CDMA handset applications has been studied. The dual path PA is fabricated in InGaP/GaAs HBT and consumed the 75 mA/30 mA quiescent current in the high/low power mode with the 3.4 V supply voltage. Using the single input/output matching circuit, the dual PA combined well. The measured results show that a 1 dB compression point of 28 dBm and a 36 % PAE at 28 dBm output power in the high power mode, and a 1 dB. Compression point of 20 dBm and a 14.2 % PAE of at 16 dBm output power in the low power mode and -49 dBc at 16 dBm average output power in the low power mode. The measured result meets the linearity requirement of the dual path power amplifier. This concept is also available for the other CDMA/OFDM specifications.

References

1. Hanington G, Chen PF, Asbeck PM, Larson LE (1999) High-efficiency power amplifier using dynamic power-supply voltage for CDMA applications. *IEEE Trans Microwave Theory Tech* 47:1471-1476
2. Sevic JF (1997) Statistical characterization of RF power amplifier efficiency for CDMA wireless communication system. In: *Wireless conference*, Boulder, pp 110-113
3. Noh YS, Park CS (2001) Linearised InGaP/GaAs HBT MMIC power amplifier with active bias circuit for W-CDMA application. *Electron Lett* 37(25):1523-1524

U-Health Platform for Health Management Service Based on Home Health Gateway

Jong-Hun Kim, Si-Hoon Ahn, Jae-Young Soh
and Kyung-Yong Chung

Abstract The Ubiquitous Health, or u-Health, service is an IT health care service using the ubiquitous computing environment. u-Health provides customized medical services. As it is a service that has developed from the current hospital visiting medical system, the u-Health service provides a patient with healthcare anywhere and anytime. In this paper, we propose a health management service model using home health gateway based on the u-Health platform. Using home health gateway, u-Health can provide health monitoring, diet, and exercise services using the Healthcare Decision Support Module (HDSM) in the ubiquitous environment. This approach would offer specialized services using an external content provider of DB. In addition, a doctor can provide advice to patients using the monitoring service. The proposed u-Health platform provides effective services using home health gateway in ubiquitous environments to customers, which will improve the health of chronic patients.

Keywords U-Health platform · Health gateway · Health management · Health monitoring

J.-H. Kim · S.-H. Ahn · J.-Y. Soh
U-Healthcare Department, Bit Computer, 1327-33 Bitville, Seocho-dong,
Seocho-gu, Seoul, South Korea
e-mail: kimjh@bit.kr

S.-H. Ahn
e-mail: ashinfo@bit.kr

J.-Y. Soh
e-mail: jysoh@bit.kr

K.-Y. Chung (✉)
School of Computer Information Engineering, Sangji University,
83 Sangjidae-gil,
Wonju-si, Gangwon-do, Korea
e-mail: dragonhci@hanmail.net

1 Introduction

Korea's continuing economic growth and the five-day workweek policy have led to an increased interest in medical services, leisure activities, and culture among Koreans. In particular, an increasing number of Koreans are showing a keen interest in health and happiness, which is due to new diseases and greater life expectancy among Koreans due to this economic growth. The search for an efficient way to respond to this increasing interest in health and wellness has placed the u-Health service in the spotlight. With the u-Health service, users are given medical services anytime and anywhere as their vital signs and environmental signs are collected. At present, the development of the u-Health service environment system and the middleware are actively being conducted [1]. For patients of chronic diseases, the u-Health service should include the medical services of a hospital and medical guidelines to encourage patients to improve their health via self-regulation [2]. Lately, a new u-Health service for high pressure patients and diabetic patients has been introduced, and much investment and research are being applied to developing various devices and solutions. However, most of the services are performed as a user enters only the vital signs that have been measured by a Personal Health Device (PHD) [3–6]. Current services only allow users to check their health information and accumulated medical vital sign information. The use of ubiquitous-based IT [7] is necessary for patients to be able to efficiently conduct their medical checks at home. Therefore, development and distribution of wireless PHDs as well as development of a home health gateway that gathers the measured vital signs should be promptly carried out. In this thesis, a gateway to automatically gather a user's measured vital signs is developed as well as a system that would provide the user with medical information by using HDSM. In addition, for the development of a service connected to the hospital, the thesis designs a u-Health platform in which exchange of clinical data and multi-device service are possible.

2 Relative Work

When the current u-health platform is looked into, it is found out that, based on the input of a patient's body information and disease information, the customized health workout and diet provision platform through personal body information of Welltizen [8] have provided not only workout prescriptions and diet but also a graph of consumption condition, nutritional balance and calories. Dreamcare platform of BIT Computer [9] has been liberated from the current visiting prescription and the web service to finally provide the Smart TV platform-based service. Through a patient's body composition analysis and self medical examination, the diet information is offered as well as recipes and experts' advices of the provided diet. For the current u-health platform, it was only possible for self-

visiting and for a particular space such as web or a digital TV. There has been no intellectual information provision service conducted by auto-monitoring and system in a social life of a patient. In addition, a service model based on the medical guidelines has not been presented. The u-health platform proposed in this thesis features the connection with a hospital and automatic gathering of vital signs. With this u-health platform, real-time service is provided through the smartphone, TV and web.

3 U-Health Platform for Health Management Service

3.1 Home Health Gateway

As a short distance telecommunication technique of low electricity and high reliability, Bluetooth is believed to be a skill that is proper for an application device in the medical field. In the past, independent ubiquitous healthcare service was developed in hospitals, houses and care facilities based on Serial Port Profile (SPP) of Bluetooth. Through the development, personal detection devices and servers were distributed. However, since these independent applications are conducted via each different data, they are not likely to be compatible. In order to work on the problem, IEEE 11073 standard has been announced in ISO. IEEE 11073 has introduced a definition of an agent and a manager. The agent is PHD and the manager is capable of showing the user data from the agent by having a bigger arithmetic resource. Each of the agents communicates with one manager and one manager can communicate with multiple agents. The bluetooth file which has been designated by IEEE 11073 standard is Health Device Profile (HDP). The PHD agent that supports HDP will be connected only to a manager that supports HDP. Even if the communication method is the bluetooth, it would not be connected to a communication method based on the legacy SPP profile and others which is, for example, a server that connects to TCP/IP. Likewise, the legacy private medical measuring device would not be connected to HDP profile manager and even if it is conducted for pairing by the bluetooth, it will not be compatible. Furthermore, since PHD is only capable of one pairing with one manager, if the manager changes, inconvenience would be expected. Therefore, for PHD, a gateway to accept both HDP-based PHD and SPP-based PHD is required while for the healthcare application, a gateway to support all the other communication methods including the bluetooth is necessary.

When the home health gateway is connected to PHD and when the home health gateway is operated as the built-in agent and the gateway mode that would deliver the healthcare data to the built-in manager, the service layer or a service device which would receive the healthcare data packet from PHD, the home health gateway connects to PHD through the built-in manager. In the packet, delivered via the built-in manager from PHD, the healthcare data is extracted and, this data is

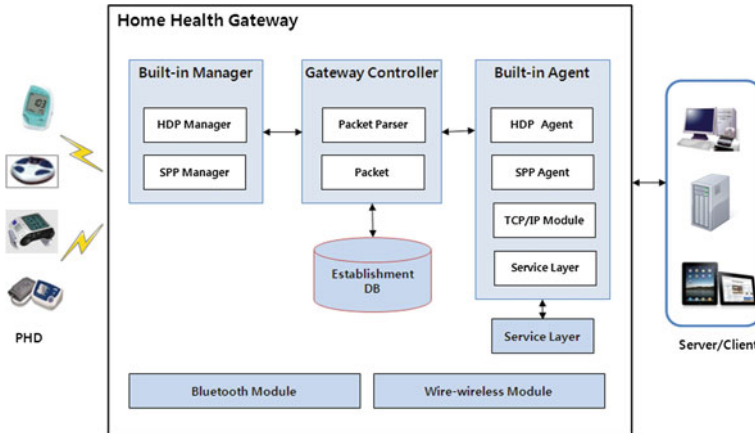


Fig. 1 Diagram for home health gateway composition

delivered to the built-in agent. All these parts described so far is a gateway control of the home health gateway and, in Fig. 1 the diagram of the home health gateway is presented.

3.2 *U-Health Platform*

At present, the healthcare service provision should be realized based on the provision of various services via many different devices in the multi service environment but not in the current single service environment. In addition, to conduct decision-making on a user's medical condition, an intellectual information provision service with the clinical information used and a service to automatically collect and monitor the user's vital signs are necessary. The u-health platform for the healthcare service proposed in this thesis is described in Fig. 2. Through PC and smart devices, the user contacts U-Health Center, requests for a service and is given a personalized service. And, through various PHDs, the user checks vital signs which are gathered via the home health gateway and sent to the platform. In the healthcare center, the information gathered through the home health gateway is analyzed via Health Decision Support Module (HDSM) and for users, they can be provided with the present medical information through smart devices anytime and anywhere. For doctors, they can perform the medical service to a patient through the u-health platform after they are given monitored information and clinical information. With HDSM, the doctors also acquire information that is necessary for the medical examination. Through the intellectual information provision service, the quality of the medical service is improved. The u-health platform supports users to enjoy contents of various special service providers in the multi-device environment. The service distributors provide lots of information on diets,

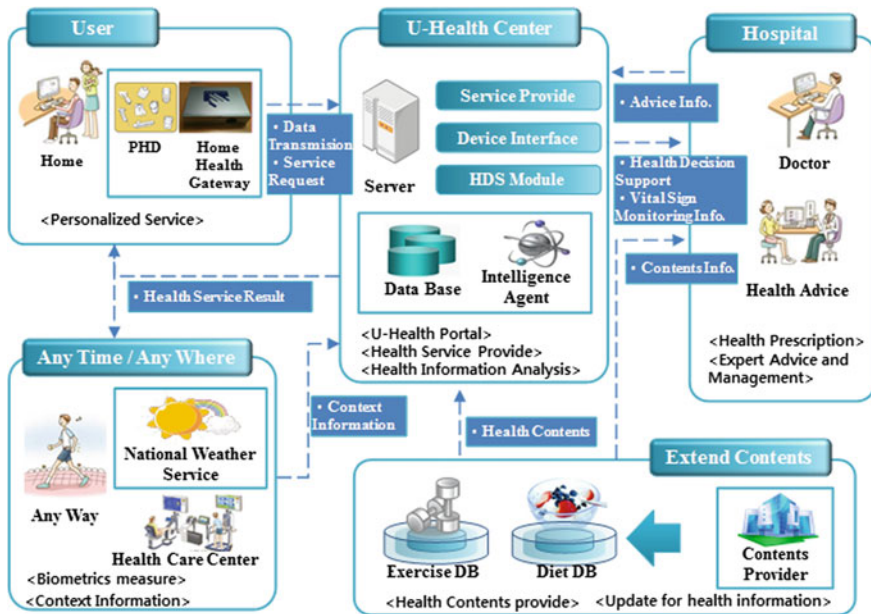


Fig. 2 U-health platform for health management service

workouts and health and the users are given the most appropriate contents through the intellectual information provision service.

4 Conclusion

In the thesis, the u-health platform for the home health gateway based healthcare service has been proposed. For the chronic patients, the proposed u-health platform provides medical information while for the doctors; it gives information that is necessary for the medical examination. By offering the home health gateway based vital sign monitoring service, the u-health platform reduces the inconvenience of the user's vital sign entry but increases the service uses. In addition, by supporting various PHDs and in/output devices, the u-health platform helps the present smart devices to provide users with efficient u-health service. In consequence, the patients are now able to get real-time customized medical service anytime and anywhere as the u-health platform is capable of providing analysis information. Since both the patient and the doctor are given necessary information, efficiency of medical examination and healthcare will be improved. In a follow-up research, how to apply the proposed u-health platform to an actual clinical experimentation should be studied. Also, through a verification of efficacy and the user's satisfaction of the u-health platform, the platform will be proved for its functions. Last

but not least, as a complementary process is conducted, it is expected that the u-health platform will be commercialized to provide the users with high-quality healthcare service.

Acknowledgments This research was supported by a grant (SS100020) from Seoul R&BD Program funded by the Seoul Development Institute of Korean government. Sincere thanks go to Mr. Jaekwon Kim who provided the development for this thesis.

References

1. Kyung Ryu JK, Kim JH, Kim JK, Lee JH, Chung KY (2011) Context-aware based U-health environment information service. *J Korean Inst Inf Technol* 11(7):21–29
2. Loring K, Holman H (2003) Self management education: history, definition, outcomes, and mechanisms. *Ann Behav Med* 26(1):1–7
3. Biospace. <http://www.biospace.co.kr/>
4. A&D. <http://www.andonline.com/>
5. Omron. <http://www.omron.com/>
6. Allmedicus. <http://www.allmedicus.co.kr/>
7. Weiser M (1991) The computer for the twenty-first century. *Sci Am* 265(3):94–104
8. Welltizen. <http://www.welltizen.com/>
9. BIT Computer. <http://www.bit.kr/>

A 2-D Visual Model for Sasang Constitution Classification Based on a Fuzzy Neural Network

Zhen-Xing Zhang, Xue-Wei Tian and Joon S. Lim

Abstract The human constitution can be classified into four possible constitutions according to an individual's temperament and nature: Tae-Yang (太陽), So-Yang (少陽), Tae-Eum (太陰), and So-Eum (少陰). This classification is known as the Sasang constitution. In this study, we classified the four types of Sasang constitutions by measuring twelve sets of meridian energy signals with a Ryodoraku device (良導絡). We then developed a Sasang constitution classification method based on a fuzzy neural network (FNN) and a two-dimensional (2-D) visual model. We obtained meridian energy signals from 35 subjects for the So-Yang, Tae-Eum, and So-Eum constitutions. A FNN was used to obtain defuzzification values for the 2-D visual model, which was then applied to the classification of these three Sasang constitutions. Finally, we achieved a Sasang constitution recognition rate of 89.4 %.

Keywords Sasang constitution · Tae-Yang (太陽) · So-Yang (少陽) · Tae-Eum (太陰) · So-Eum (少陰) · Ryodoraku (良導絡) · Fuzzy neural network

Z.-X. Zhang
School of Information & Electric Engineering, Ludong University,
Yantai, China
e-mail: billzhenxing@gmail.com

X.-W. Tian · J. S. Lim (✉)
IT College, Gachon University, San 65 Bokjeong-dong, Sujeong-gu,
Seongnam, Gyeonggi-do, South Korea
e-mail: jslim@gachon.ac.kr

X.-W. Tian
e-mail: tianxuemaog@gmail.com

1 Introduction

Sasang constitutional medicine is a traditional Korean medicine, which was founded by Jae-Ma Lee and systematically theorized in his book *Dongguisusebowon* (東醫壽世保元) in 1894 [1]. Major distinctions of the body are formed during incubation period of life, rendering person's predisposed weakness and strength. This prenatal imbalance determines a person's constitution, physiology, pathology, emotional orientation, and nature. As a result, each type shares the similar aspects of bodily structure, function, and metabolism, as well as psychological and behavioral characteristics. In Sasang medicine, four major body types are distinguished. They are called Tae-Yang(太陽), So-Yang(少陽), Tae-Eum(太陰), and So-Eum(少陰) [3–5, 8–11]. Most of people can be categorized in these four body types.

Energy medicine is a new science and technology that relies on the measurement and analysis of 12 standardized sets of meridian energy signals (each set includes one point on the left-hand side and one on the right-hand side) of the human body, based on a long history of research [2, 7], and these points are shown in Fig. 1.

Serial measurements of the meridians using Ryodoraku theory and the Ryodoraku device have been used to provide a precise, continuous, and objective check of a patient's medical progress. The meridians for the hands are designated as H1 to H6 for the Ryodoraku system, and those for the feet as F1 to F6, making 12 meridians in all. These Ryodoraku meridians are equivalent to those of energy medicine and traditional oriental medical theories, whereby most health conditions can be discovered through energy signals of the 12 sets of meridians (again, each set has one left and one right point). The meridian equivalents of the Ryodoraku are shown in Table 1.

In this paper, we analyze a data set that records 12 sets of meridian energy signals from 35 healthy people by using a Sasang constitution classification method. Sasang classification is accomplished using a fuzzy neural network, named NEWFM (neuro-fuzzy network with a weighted fuzzy membership function) [6], which classifies the Sasang constitution by calculating Takagi–Sugeno defuzzification values that we use for two-dimensional (2-D) visual modeling.



Fig. 1 Twelve measurement points for meridian energy

Table 1 Meridians description

Hand				Foot			
Meridians	Description	Left side feature' name	Right side feature' name	Meridians	Description	Left Side feature' name	Right side feature' name
H1	Lung	LH1	RH1	F1	Spleen	LF1	RF1
H2	Pericardium	LH2	RH2	F2	Liver	LF2	RF2
H3	Heart	LH3	RH3	F3	Kidney	LF3	RF3
H4	Small Intestine	LH4	RH4	F4	Urinary Bladder	LF4	RF4
H5	Triplewarmer	LH5	RH5	F5	Gall Bladder	LF5	RF5
H6	Large Intestine	LH6	RH6	F6	Stomach	LF6	RF6

2 Materials and Method

2.1 Materials

All of the data for this research, including measurements of Sasang constitutional signals, have been obtained from the Oriental Medicine Hospital (in Korea). Participants in the study were given simultaneous medical examinations with the Ryodoraku device, which is based on the Ryodoraku theory that abnormalities or disease of the viscus or the internal organs are reflected by changes in the biological electric current.

Generally, the Ryodoraku device measures the energy of the meridians via a special probe that touches the measuring points shown in Fig. 1 [1]. The 12 left-side and 12 right-side meridian energy signals are recorded, as shown in Fig. 2. This figure also shows the user interface of the Ryodoraku device, which was developed by URACLE (Uracle Technologies, Co., Korea). We obtained and collected meridian energy signals, including So-Yang, Tae-Eum, and So-Eum constitutions, from 35 healthy people. Of the 35 subjects, 6 had a So-Yang constitution, 16 had a Tae-Eum constitution, and 13 had a So-Eum constitution. We used the sum of the quantized absolute value of left-side energy and right-side energy for each meridian to represent the total energy of that meridian.

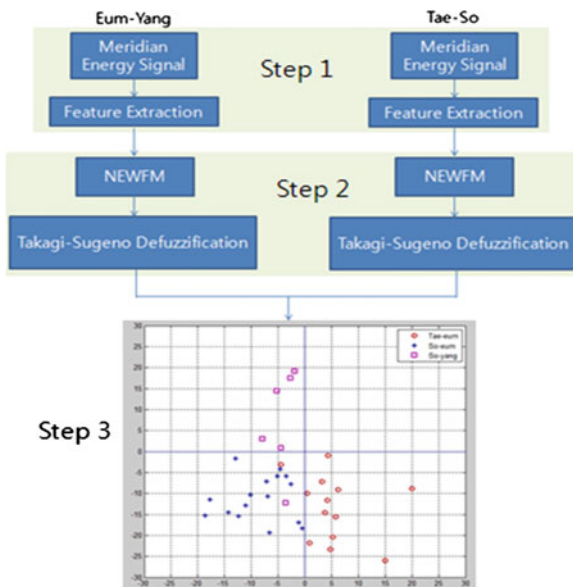
2.2 Two-Dimensional (2-D) Visual Model

As discussed in Introduction, the Sasang constitution can be characterized as the intersection of four basic factors (the terms Eum, Yang, Tae, So). To represent these factors of the Sasang constitutional classification, we built a 2-D visual model (shown in Fig. 3, Step 3).

Fig. 2 User interface of the meridian-energy recorder of Ryodoraku, which was developed by URACLE (Uracle Technologies, Co., Korea)



Fig. 3 Structure of the 2-D Sasang Constitution Classification using NEWFM



2.3 Sasang Constitution Classification Using NEWFM

We identified three categories of the Sasang constitution: Tae-Eum, So-Eum, and So-Yang. NEWFM produced a Takagi–Sugeno defuzzification value for each category from which we built a 2-D visual model for the Sasang classification, as shown in Fig. 3. The processing steps are as follows:

Step 1 (Feature Extraction): We extracted 24 basic features by Ryodoraku device, and made 18 extensional features based on these 24 basic features. Features descriptions are shown in Table 1 and Table 2

Table 2 Extensional features description

Feature' Name	Description	Feature' Name	Description
Median_LH1-LF6	The median value of all the left side detection points.	LH4/RH4	The ratio of LH4 and RH4
Median_RH1-RH6	The median value of all the right side detection points.	LH5/RH5	The ratio of LH5 and RH5
Mean_LH1-LF6	The mean value of all the left side detection points.	LH6/RH6	The ratio of LH6 and RH6
Mean_RH1-RH6	The mean value of all the right side detection points.	LF1/RF1	The ratio of LF1 and RF1
Median_LH1-LF6/ Median_RH1-RH6	The ratio of Median_LH1-LF6 and Median_RH1-RH6	LF2/RF2	The ratio of LF2 and RF2
Mean_LH1-LF6/Mean_ RH1-RH6	The ratio of Mean_LH1-LF6 and Mean_RH1-RH6	LF3/RF3	The ratio of LF3 and RF3
LH1/RH1	The ratio of LH1 and RH1	LF4/RF4	The ratio of LF4 and RF4
LH2/RH2	The ratio of LH2 and RH2	LF5/RF5	The ratio of LF5 and RF5
LH3/RH3	The ratio of LH3 and RH3	LF6/RF6	The ratio of LF6 and RF6

- Step 2 (Takagi–Sugeno Defuzzification): Get Takagi–Sugeno defuzzification values using NEWFM
- Step 3 (Distribute 2-D Sasang constitution Space): Use Takagi–Sugeno defuzzification values to distribute 2-D Sasang constitution space and classify Sasang constitution using the 2-D Sasang constitution space

3 Experimental Results

The performance results of the NEWFM-based 2-D Sasang constitution classification model are shown in Step 3 of Fig. 3 (diamonds are So-Yang, circles are Tae-Eum and snowflakes are So-Eum). Using the Takagi–Sugeno defuzzification values, we described constitutions in the 2-D Sasang constitution space. An overall classification accuracy of 89.4 % was achieved using the 2-D Sasang constitution classification system. The individual classification accuracies for So-Eum, Tae-Eum, and So-Yang were 100 %, 84.6 %, and 83.3 %, respectively.

4 Concluding Remarks

This paper presents a systematic attempt to develop a general purpose signal classification method that can be used to classify different categories of meridian

energy signals. We proposed a simple method for building a 2-D visual model that links meridian energy signals to Sasang constitution. Our preliminary experiment shows that the constitutions of Sasang constitution can be well distributed in the 2-D space. Automatic constitution recognition in Sasang constitution is currently an active research area with a wide range of applications in oriental medicine. So-Eum constitution was detected at a significantly high recognition rate of 100 %, and the overall classification accuracy was 89.4 %.

Acknowledgments This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the Convergence-ITRC (Convergence Information Technology Research Center) support program (NIPA-2012-H0401-12-1001) supervised by the NIPA (National IT Industry Promotion Agency).

References

1. (2002) System Operation Manual for Meridian Energy Analysis Device MedPex Enterprises Ltd. Taiwan
2. Charchago AY (2006) New approach to interpret the Ryodoraku map to estimate the functional state of the body's regulatory systems in scenar-therapy. In: 12th scientific and practical conference
3. Chin SJ (2002) Evaluation of acupuncture treatment of low back pain using a new Ryodoraku neurometric system. *J. Chin Med*
4. Hirohisa ODA (1989) Ryodoraku textbook-Ryodoraku autonomic nervous system therapy. Published by Naniwasha Publishing, Osaka
5. Hong WS (1985) 精校黃帝內經素問. *Orient Med Res Cent* 2:83–94
6. Lim JS (2009) Finding features for real-time premature ventricular contraction detection using a fuzzy neural network system. *IEEE Trans Neural Netw* 20(3):522–527
7. McGill ER (2005) Investigative study of long term effects of life wave[®] patches using electro meridian analysis system (EMAS). Life Wave[®] Convention Las Vegas
8. Nakatani Y (1956) Skin electric resistance and Ryodoraku. *J Auton Nerve* 6:52
9. Park YB (1996) The theory and clinical application of Ryodoraku. *J Third Med* 22(2):83–94
10. Sun JK (1998) Study on meridian and clinical application. *J Meridian Diagn* 32(3):51–57
11. <http://taoofmedicine.com/sasang-constitutional-medicine/>

Graph Coloring Algorithms and Applications to the Channel Assignment Problems

Surgwon Sohn

Abstract This paper presents graph coloring algorithms and their applications to the channel assignment problems. Two application problems of frequency assignment of low power FM broadcasting and reader collision problem of RFID system are modeled as graph coloring problems. Instead of performing an exhaustive search for the optimal result, we provide both search space reduction and variable ordering heuristics to obtain good approximate solutions. Constraint optimization modeling and variable ordering enforce the backtracking process in graph coloring algorithms, so the search space is greatly reduced and the approximate solution is found quickly. A great deal of outstanding work on graph coloring algorithms is described and applied to the simulation results.

Keywords Graph coloring algorithms · Channel assignment · Variable ordering

1 Introduction

In 1980, Hale introduced the relationship between frequency assignment problems (FAPs) and graph coloring [1]. Since then, many authors have noted that frequency assignment is a graph coloring problem if only the co-channel constraint is involved [2–4].

Graph-coloring is defined as a way of coloring the vertices of a graph such that no two adjacent vertices share the same color. The most common type of graph coloring seeks to minimize the number of colors for a given graph. Minimal

S. Sohn (✉)

Graduate School of Venture, Hoseo University, 9 Banpo-daero, Seoul, South Korea
e-mail: sohn@hoseo.edu

coloring can be achieved using a brute-force search or exhaustive search that consists of systematically enumerating all possible candidates for the solution and checking whether each candidate satisfies the problem's statement when the problem size is limited [5]. Brelaz's heuristic algorithm can also be used to produce a satisfactory, but not necessarily minimum coloring of a graph [6]. To find an optimal solution, Giortzis tried the branch-and-bound method of integer linear programming [2].

On the other hand, many frequency assignment problems can be expressed as constraint satisfaction problems (CSPs) [3, 4]. A CSP consists of a set of variables. Each variable is associated with a finite set of possible values and a set of constraints restricts the values that can be simultaneously assigned to the variables. Most of these techniques produce approximate, near-optimal solutions. In order to assess the quality of these approximate solutions, some lower bounds are needed. The lower bound of the chromatic number of graph coloring has been studied for a few decades [7].

In this paper, we describe a formal definition of the channel assignment problems in Sect. 2. Furthermore, graph coloring algorithms are described in Sect. 3, and simulation results are presented in Sect. 4, and a conclusion is provided in Sect. 5.

2 The Channel Assignment Problems

We build CSP models for two different but similar channel assignment problems. One is frequency assignment for low power FM broadcasting and the other is reader collision problem in RFID systems. The CSP model gives us a drastic reduction of search space of the problems.

2.1 *Frequency Assignment for Low Power FM Broadcasting*

FM radio signals are transmitted on VHF frequencies in the range of 88–108 MHz and this band is usually partitioned into a set of channels, all with the same bandwidth: 200 kHz of frequencies. For this reason, the channels are usually numbered from 1 to 100. Due to the vastly increasing number of radio stations and the very limited availability of frequencies, it is difficult to find vacant frequencies that can be assigned to the low power FM radio [8] operators. Consequently, a major challenge to improve the frequency utilization efficiency is needed.

Our research objective is to find a frequency assignment that satisfies all the constraints using a minimum number of frequencies while maximizing the number of radio stations served in a given area. A given area is partitioned into cells, which are served by FM radio transmitters. In the problem, each radio transmitter can be viewed as a variable. The set of possible frequencies makes the domain for each transmitter variable can take. Several types of frequency separation constraints

exist and these make the problem more complex than the simple graph coloring as shown below.

- Co-channel constraint (CCC): Any two transmitters located at different cells cannot use the same channel within interfering distance. If f_i and f_j are the frequencies assigned to transmitter i and j , $d(i, j)$ is the distance between transmitter i and j , k is a sufficiently separated distance, then these give rise to constraints of the form:

$$f_i \neq f_j, \text{ if } d(i, j) \leq k$$

- Adjacent channel constraint (ACC): Any two transmitters located at different cells within interfering distance cannot use *adjacent channels*, unless they are separated by sufficient frequencies because there is still the potential for interference normally within one to three channel-distance. The ACC can be represented as an n by n symmetric compatibility matrix corresponding to a constraint graph with n vertices. Therefore a number of constraints arise in the following form:

$$|f_i - f_j| \geq c_{ij}$$

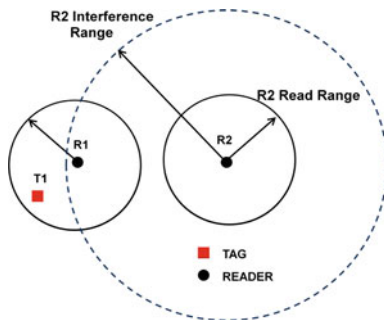
where $i \neq j$ and c_{ij} is the required number of channel separation.

- IF constraint (IFC): An intermediate frequency (IF) a beat frequency between the signal and the local oscillator in a superheterodyne radio receiver. The transmitters should not be apart between any two channels assigned to the same cell or adjacent cells for better reception of signals.
- Pre-assigned frequency constraint (PFC): Some frequencies are already occupied by local FM stations within the service area. Thus, these channels cannot be assigned.

2.2 The Reader Collision Problem

In RFID systems, interference refers to the collision of different radio signals with the same frequency, which leads to distorted signals being received. Occasionally, if tags or readers broadcast radio signals of the same frequency, collisions occur [9, 10]. There are two primary types of reader interferences experienced in RFID systems: reader-to-reader interference and reader-to-tag interference [11]. Reader-to-reader interference occurs when a reader transmits a signal that interferes with the operation of another reader, thus preventing the second reader from communicating with tags in its interrogation zone [10, 12]. Reader-to-tag interference occurs when a tag is located in the interrogation zone of two or more readers and more than one reader tries to read the same tag simultaneously [10]. We consider reader-to-reader interference only in this paper.

Fig. 1 RFID reader interference



For example, in Fig. 1, the signal strength of a reader is superior to that of a tag and therefore if a frequency channel occupied by R2 is the same as that of T1 and R1, R1 is no longer able to listen to T1's response. This problem can be solved when the channel occupied by R1 and T1 is different from the channel of R2. In the case of an identical channel being used, it should be assumed that R1 and R2 keep their distance long enough for T1's responding signal strength to become more powerful than R2's interference signal [13].

Some researchers indicate that the reader collision problem of RFID systems is equivalent to the channel assignment problem, which, in turn, is equivalent to the simple graph-coloring problem [12]. In the reader collision problem, the following types of frequency separation constraints exist.

- Co-site constraints (CSC) : Any pair of channels in the same cell must have the specified frequency separation, typically 3 channels in RFID reader collision problem. r_i means reader i , therefore $f(r_i)$ is the channel assigned to the reader i .

$$|f(r_i) - f(r_j)| \geq k$$

- Adjacent channel constraints (ACC): Any two transmitters located at different cells within interfering distance cannot use *adjacent channels*, unless they are separated by sufficient frequencies because there is still the potential for interference normally within one to two channel-distance. Therefore a number of constraints arise in the following form:

$$|f_i - f_j| \geq c_{ij}$$

where $i \neq j$ and c_{ij} is the required number of channel separation. The ACC can be represented as an n by n symmetric compatibility matrix corresponding to a constraint graph with n vertices.

This problem is very similar to the frequency assignment of LPFM broadcasting because we assign different channels to avoid interference between each

transmitter. In this case, an RFID reader can be thought of as a transmitter and a tag as a radio receiver.

3 The Graph Coloring Algorithms

A graph $G(V,E)$ is a pair of finite sets where the set V contains the vertices of the graph and the set E contains distinct unordered pairs of vertices called edges. The set of vertices V may represent the transmitters in the FM radio stations or readers in the RFIS systems, i.e. $V = \{v_1, v_2, \dots, v_n\}$ is the set of readers in a RFID system or the set of FM radio transmitters. The set of edges E represent potential reader collision between readers in a RFID system or interferences between FM radio transmitters. The set of possible frequencies makes the domain for each variable. Therefore, D_i is the domain of frequencies that a variable i can take.

In RFID system, if two readers have overlapping regions between their interrogation zones, we connect these two readers by an undirected edge. This way, the reader collision in a dense reader environment is constructed as a multigraph $G = (V,E)$ [14].

The traditional approach to minimizing the number of assigned frequencies is to perform graph coloring on the channel graph. Because each color can be identified with one channel, the minimal number of colors, which is called the *chromatic number* in graph theory, equals the minimal number of channels needed.

Channel assignment problems are modeled as constraint satisfaction problems in Sect. 2. A solution to a CSP is the assignment of every variable to a value in its domain in such a way that all the constraints are satisfied. The total number of frequencies assigned to the transmitters can be considered an objective. The problem is solved by a systematic search through the set of all possible assignments to the decision variables.

Experiments show that good variable ordering moves a CSP solution to the left of the search tree, so the solution can be found quickly using a heuristic backtracking algorithm. The backtracking method essentially performs a depth-first search of the search space and it can be enhanced using variable ordering heuristics.

One of the best-known dynamic variable-ordering heuristics is *dom*, which chooses the next variable with the smallest remaining domain. Therefore, it is also called the *smallest-domain-first* principle. This ordering heuristic was introduced by Golomb and Baumert [15] and Haralick and Elliott [16], who show analytically that *dom* minimizes the depth of search trees.

Static heuristics order the variables from the highest to the lowest degree by repeatedly selecting a variable in the constraint graph. The typical algorithm is *deg*, presented by Freuder [17]. Sohn applied the *deg* algorithm to the frequency assignment of LPFM problem to generate a satisfactory, but not minimal result [8]. We can also think of the $\langle deg, dom \rangle$ heuristic in addition to simple *deg*. In this case, *dom* acts like a tiebreaker when the degree of the variables is the same.

Bessiere and Regin [18] show that static variable ordering, which considers variables in descending order of degree, gives good results in comparison with *dom* when the constraints are sparse. However it performs very badly on complete constraint graphs. They propose a combined heuristic *dom/deg* that does not give priority to the domain size or to the degree of variables, but uses them equally in the criteria. This chooses the next variable, minimizing the ratio of domain size to degree. Boussemart et al. [19] propose *dom/ ω deg* to divide domain size by weighted degree. As another efficient combined heuristic of *dom* and *deg*, we propose *α dom-deg* to select the first variable in the smallest value of domain size and the biggest value of degree, where α is a weighting factor in the domain size. The algorithm *α dom-deg* chooses the first transmitter in the center of the service area because the center is most constrained. Once a transmitter has been selected, the algorithm generates a frequency for it in a nondeterministic manner. Once again, the model specifies a heuristic for the value ordering in which the frequencies must be tried. To reduce the number of frequencies, the model says to try the values that were used most often in previous assignments first. This is called the *most-used* value-ordering heuristic.

4 Simulation Results

4.1 Frequency Assignment for LPFM Broadcasting

For simplicity, we have chosen a grid of square cells for the topology of the region. We consider a 5,000 by 5,000 m square service area for each instance. The each cell is 500×500 m in size and we have a total of 100 cells in the service area. As is clear from the problem statement, transmitters are contained within cells, and each cell may have one radio transmitter at the most. Some cells may not have a transmitter due to a geographic reason. Some frequencies cannot be assigned to the transmitters because commercial broadcasting operators already use the frequencies.

Widely used intermediate frequencies (IF) are 10.7 MHz in the FM superheterodyne radio receiver. This makes IF constraints, so $10,700/200$ kHz = 53.5 channels (53 and 54 for real channels) should not be apart between adjacent transmitters within the service area. We assume all LPFM transmitters are at 25 m height above average terrain (HAAT) with 1 watt effective radiated power (ERP). The radio service area is within about 2 km radius for each transmitter. Table 1 clearly shows the minimum channels when each variable ordering heuristic is applied to the problem.

Table 1 Performance comparison of variable ordering heuristics for frequency assignment of LPFM broadcast radio stations

Variable ordering heuristics	Minimum # of channels	Choice points	Time (s)
Dsatur	15	71,898	18.01
<i>dom/deg</i>	17	488,259	64.30
<i>dom/odeg</i>	15	26,346	7.80
<i>αdom-deg</i>	15	8,296	2.25

Table 2 Performance comparison of variable ordering heuristics for the reader collision problem

Variable ordering heuristics	Minimum # of channels	Choice points	Time (s)
Dsatur	24.0	55,859	1.89
<i>dom/deg</i>	27.0	277,853	8.45
<i>dom/odeg</i>	23.6	54,959	7.80
<i>αdom-deg</i>	21.8	79,027	2.45

average channel for 5 instances

4.2 The Reader Collision Problem

We consider a 100 m by 100 m square service area for each instance. Each cell is 10 m by 10 m in size, and we have a total of 100 cells in the service area. As is clear from the problem statement, RFID readers are contained within cells, and each cell may have zero to two readers where we use 67 readers on average in the service area. Some cells may not have readers due to geographic reasons. The bandwidth is 200 kHz, and we use a total of 32 channels in RFID frequency ranges from 917 to 923.5 MHz. Each reader's output power is 10 mW effective radiated power (ERP).

Table 2 shows the minimum number of channels when each variable ordering heuristic is applied to the problem. We carry out our simulation experiments using an ILOG OPL 3.7 optimization tool to verify the effectiveness of the proposed ***αdom-deg*** variable ordering algorithm over our simulation environment.

5 Conclusions

Solving a frequency assignment problem is a graph coloring and many researchers have attempted to find approximate solutions using heuristic searches. We presented the CSP models for the two different problems to reduce the search space: one is the frequency assignment for low power FM broadcasting and the other is the reader collision problem of RFID systems. We also applied the proposed variable ordering heuristic, ***αdom-deg***, to traverse the search tree efficiently and reduce search time greatly. The simulation results show that the proposed algorithm, ***αdom-deg***, is far better than the conventional algorithms in terms of three comparison parameters, taken as a whole.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (KRF-2008-313- D00954)

References

1. Hale WK (1980) Frequency assignment: theory and applications. In: Proceedings of the IEEE, vol 68, pp 1497–1514
2. Giortzis AI, Turner LF (1997) Application of mathematical programming to the fixed channel assignment problem in mobile radio networks. In: IEEE Proceedings of Communication, vol 144, pp. 257–264
3. Carlsson M, Grindal M (1993) Automatic frequency assignment for cellular telephone using constraint satisfaction techniques. In: Proceedings of the tenth international conference on logic programming, pp 647–665
4. Walsher JP (1996) Feasible cellular frequency assignment using constraint programming abstractions. In: Proceedings of the workshop on constraint programming applications, Cambridge
5. Skiena S (1998) The Algorithm Design Manual, Springer, New York
6. Br'elaz D (1979) New methods to color the vertices of a graph. Commun ACM 22(4):251–256
7. Achlioptas D, Naor A (2005) The two possible values of the chromatic number of a random graph. Ann Math 162(3):1333–1349
8. Sohn S, Jo GS (2006) Solving a constraint satisfaction problem for frequency assignment in low power FM broadcasting. Lect Notes Artif Intell 4304:739–748
9. Zhou S, Luo Z, Wong E, Tan CJ, Luo J (2007) Interconnected RFID reader collision model and its application in reader anti-collision. In: Proceeding of the 2007 IEEE international conference on RFID, pp 212–219
10. Engels DW, Sarma SE (2002) The reader collision problem. In: Proceedings of the 2002 IEEE international conference on systems, man and cybernetics, pp 92–97
11. Sayeed SK, Kim YS, Yang H, Yook JG (2011) A solution to the RFID reader interference problem using adaptive beam-forming approach. IETE Tech Rev 28(1):17–28
12. Song I, Hong S, Chang (2009) An improved reader Anti-collision algorithm based on pulse protocol with slot occupied probability in dense reader mode. In: Proceeding of the IEEE 69th vehicular technology conference, pp 1–5
13. Yu J, Lee W (2008) GENTLE: reducing reader collision in mobile RFID networks. In: Proceeding of the 4th international conference on mobile Ad-hoc and sensor networks, pp 280–287
14. Tian J, Fan Y, Zhu Y, Hu K (2008) RFID reader anti-collision using chaos neural network based on annealing strategy. In: Proceeding of the world congress on intelligent control and automation, pp. 6128–6132
15. Golomb SW, Baumert LD (1965) Backtrack programming. J ACM 12(4):516–524
16. Haralick RM, Elliott GL (1980) Increasing tree search efficiency for constraint satisfaction problems. Artif Intell 14(3):263–313
17. Freuder EC (1982) A sufficient condition for backtrack-free search. J ACM 29(1):24–32
18. Bessi'ere C, R'egin J-C (1996) MAC and combined heuristics: two reasons to forsake FC (and CBJ?) on hard problems. Lect Notes Comput Sci 1118:61–75
19. Boussemart F, Hemery F, Lecoutre C Sais L (2004) Boosting systematic search by weighting constraints. In: Proceeding of the ECAI, pp 146–150

Thermal Performance in Ecological Korean House

Jaewon Kim, Kyungeun Cho and Eunyoung Ahn

Abstract Korean Traditional house made of natural resources with low level thermal properties shows high performance in temperature and humidity control comfortable for human. The present investigation is to examine the thermal behaviors of air flows inside the traditional house. The transient numerical experiments are performed along with the different external conditions. Thermal properties of building units have been obtained by the parallel measurements and utilized in the numerical works. Consequently, the details of flows and temperature of air in the houses illustrate the thermal design of the traditional Korean house satisfy the requirements of human living.

Keywords Humidify · Flow patterns · CFD · Traditional house

J. Kim

Department of Mechanical Engineering, Sunmoon University, 70, Sunmoon-ro 221, Tangjeong-myeon, Asan, Choongcheongnam-do, South Korea
e-mail: jwkim@sunmoon.ac.kr

K. Cho

Department of Multimedia Engineering, Dongguk University, 26, Pil-dong 3 Ga, Jung-gu, Seoul, South Korea
e-mail: cke@dgu.ac.kr

E. Ahn (✉)

Department of Information Communication Engineering, Hanbat National University, San 16-1, Deokmyeong-dong, Yuseong-gu, Daejeon, South Korea
e-mail: aey@hanbat.ac.kr

1 Introduction

Korean traditional house, herein after referred as Hanok and its name comes from Korean language, was made of natural materials such as soil, stone, mud, rice straw, and wood. Thermal properties of these construction substances are not larger than modern ones, however the control of indoor flows inside the house provide driving forces for insulating flows to prevent heat transfer from heat sources. Surrounding conditions are outside high temperature in summer and outer cold in winter season. Thermal properties of the primary materials used in the traditional house were analyzed and proper substitutes were proposed in the prior work [1]. Investigations for the windows array and structure of Hanok are examined by prior work to improve the insulation, liquid leakage, the quiet, easy access for light, and fine finish [2].

Main interest of the present investigation lies on the looking for optimum arrangements of contact region to outside air, such as a window and door are made of traditional materials. The model shows equivalent thermal performance to the figure of the modern house using concrete structure with cement and composite insulators. Observation and validation have been carried out by numerical predictions according to turning of the seasons. Details of flows both inside and outside of the house are prepared for the explanations of the merits of the traditional unit for Hanok. In addition, insulating flow patterns inside a room of Hanok are depicted for prefixed probe region.

In oriental traditional architecture design, the roof has distinctive characteristics from other parts like as stereobate, shaft and other ornamental parts. Though there are some suggestions for 3D design of the traditional wooden building [3–5], there are few developments of a 3D roof design until now. Because that a roof is made of hundreds number of elementary component called ‘Giwa’ in Korean word and their combination appears to be a gentle curved surface. For this reason architecture designer take much time and efforts for drawing the roof by connecting roof tile and form an esthetic and elegant surface manually [6, 7]. Once it is designed, modification of the roof surface is much more difficult. This research focuses on this troublesome and we propose an automatic roof generating method which is immediately applicable into a commercial 3D CAD program. The proposed method is implemented based on the Building Information Modeling (BIM) tool, becoming a conversation topic in architecture, from the auto-generated roof can offer information about component quantity evaluation and error check list, it is useful in the whole process of architecture design.

2 Numerical Predictions

Thermal Properties For the present numerical calculation looking for the flow details and temperature distribution inside house and surrounding fluid, the model is defined by a lot of surfaces with thermal properties (thermal conductivity,

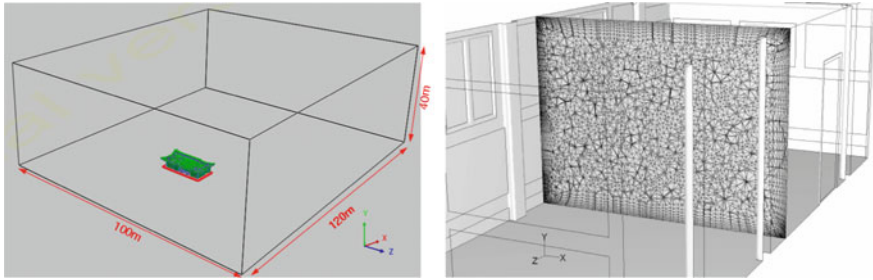


Fig. 1 Problem domain with house (left) and discrete region for numerical integration

thermal capacity and specific heat, etc.) depending on substances such as air, water, soil, stone, wood and so on. The thermal properties are measured and adopted in the calculation. Boussinesq approximation on temperature (or seasonal) dependence of the thermal properties is adopted in this research. As outdoor conditions, wind flows from the south-east direction at 5.0 m/s with temperature of 22 °C. Variation of wind in vertical direction is determined by Eq. (1).

$$V_Z = V_R \times \left(\frac{Z}{Z_R} \right)^\alpha \tag{1}$$

In Eq. (1), α denotes boundary layer thickness and is equal to 0.14 in typical Korean geophysical condition. Distribution of wind velocity is depicted in parabolic shape with aspect ratio depending on α . In Eq. (1), α denotes boundary layer thickness and is equal to 0.14 in typical Korean geophysical condition. Distribution of wind velocity is depicted as a parabolic profile according to index, α is decided by ground condition. Finite element method is used for the prediction flows and thermal behaviors and the problem domain is divided into about 1.3 million tetrahedron cells and prism layers near the solid wall with large shear stress. Figure 1 shows model for outer flows. Size of the domain is 100 m (width) by 120 m (length) by 40 m (height) and the region is divided into approximate discrete 300 million cells. The distance between centers of the adjacent cells is from 0.07 m to 1.12 m. Left of Fig. 1 displays the cell-plane in a house. Temperature condition for the present calculation is summer season with 22 °C at outdoor and 18 °C indoor, respectively. Heat flux toward indoor region is assumed by conduction, convection and radiation. The present numerical work is carried out for the steady state.

3 Result and Discussion

Velocity vectors on the mid-plane are plotted in Fig. 2. As shown in the figure, flows occur from the left to right and right upward flows appear in the plane due to the fluid viscosity. Flows over the house generate large vortex and reattachment on

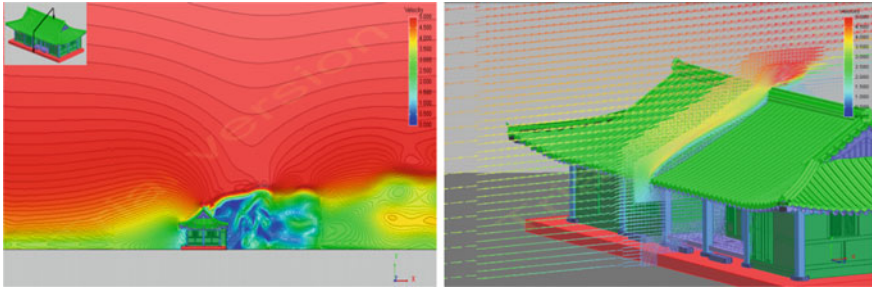


Fig. 2 Contours (left) and vector plots (right) of wind Vorticity at middle plane

Fig. 3 Contours of vortices around house at mid-length plane

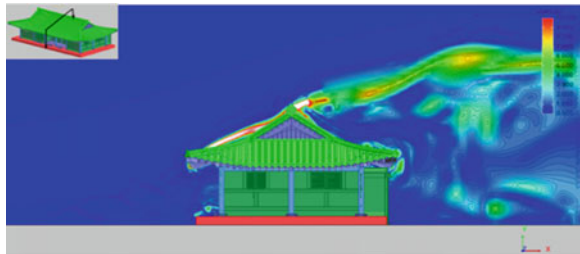


Fig. 4 Trajectory path of air-flows near windows

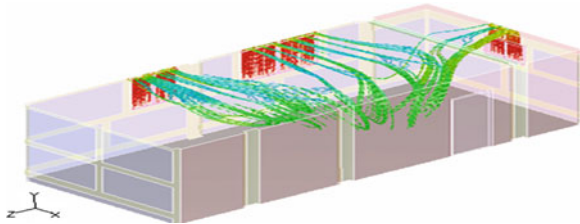
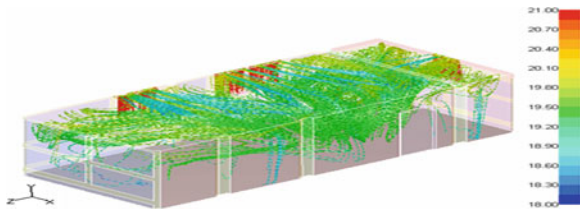


Fig. 5 Trajectory of fluid particles near windows



the vertical plane in the longitudinal direction. Flow separation takes place near the ridge of a roof. These flow patterns are identified by the velocity vector plots in right plots of Fig. 3. Vorticity distribution is shown in Fig. 3. Vortex flows generate at the protruding corners of eaves and move upward along with the down-ridge. The circulating flows separate at the ridge of a roof and reattach neat the opposite side downward surface. The vortex strength decreases as the flow is

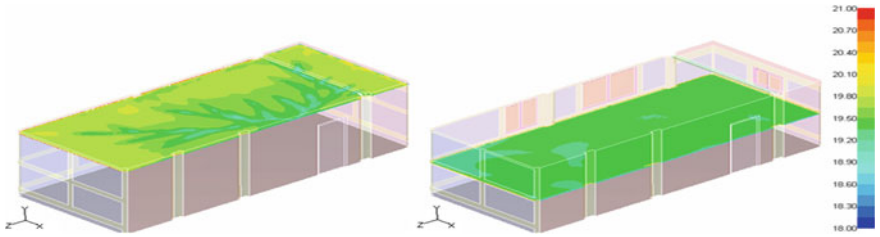


Fig. 6 Vertical variation of temperature in room

passing through the roof region. The flow patterns explain the house poses toward the south for the vital flows around the main gates.

The weak units for thermal insulation are windows which are made of translucent paper for solar access. Temperature variation near the windows cause air-flows and drive thermal exchange. Figures 4, 5 depict the trajectory of fluid-particles with color lines denote temperature degree. Upward flows from the windows toward the ceiling and consequent downward flows along the opposite wall are dominant in the plots. The indoor air flow provides fresh air to human living there.

Temperature inside human living space should be uniform for comfortable living and health life. Plots of temperature inside the living area are prepared in Fig. 6. Temperature at the junction area with windows and the outer wall is higher than average value over the whole volume and the center region of the house lower than average one. However, the difference of the maximum and minimum temperature is less than 0.5 °C and this temperature distribution creates the comfortable living space.

4 Conclusions

The present investigation numerically analyzes the thermal behaviors of air flows inside Korean traditional house made of the construction materials transmitted. The research methodology is numerical predictions of air flows depend on the temperature distribution inside the house. The thermal properties for the calculation are procured by parallel measurements. The transient numerical experiments are performed along with the different house types, weather condition, and operating time of heating. Consequently, the details of flows and temperature of air in the houses illustrate the thermal design of the traditional Korean house satisfy the requirements of human living. Energy efficiency and thermal performance of Hanok constructed by natural resources are close to those of the modern living building.

Acknowledgments This research is supported by Basic Science Research Program through the NRF (National Research Foundation of Korea) funded by the Ministry of Education, Science and Technology (2010-0021154).

References

1. National Research Institute of Cultural Heritage (2008) The characteristics evaluation of traditional architecture and probability of their application. National Research Institute of Cultural Properties, Tokyo
2. Architecture, Urban, Research Institute (AURI), Institute of Seoul Studies (ISS) (2008) Hanok building construction technology plan for the promotion of industrial research. Ministry of Land Transport and Maritime Affairs, Bogeumjari pp 129–222
3. Ahn EY, Kim JW (2010) Numerical analysis of air flows inside Koran traditional house. In: Proceedings of the Korea contents association, pp 469–471
4. Kim D, Oh H (2001) A study on changes in the space composition of Each room in Bukchon Hanoi—Focused on Open-Hanoks and publicly purchased Hanoks. *J Korean Home Manag Assoc* 26(2):115–127
5. Hanamann RJ (1981) Microelectronic device thermal resistance: a format for standardization. In: Proceedings of heat transfer in electronic equipment, pp 39–48
6. Oktay S, Hannemann R, Bar-Cohen A (1986) High heat from a small package. *Mech Eng* 108(3):36–42
7. Kraus AD, Bar-Cohen A (1983) Thermal analysis and control of equipment. McGraw Hill, New York p 302

An Automatic Roof Frame Design Method for Korean Traditional Wooden Architecture

Eunyoung Ahn and Noyoon Kwak

Abstract A designers try to design traditional wooden house, HANOK, are suffering from drawing roof frame, especially. Because that a roof frame appears in curved surface made of a number of architectural components with repetitious work. For the same reason, modification is difficult job, also. Constituent components become a part of curved roof shape and elegant line of eaves is one of the outstanding features in the Korean traditional wooden architecture. Roof shape has architectural functionalities such as passive solar system and preventing the walls from rain water etc. Moreover, the curved roof surface makes a gorgeous figure of the Korean traditional buildings in the visual viewpoint. This study suggests an efficient and systematic method to create many constituent components automatically for making curved roof surface in the Korean traditional architecture design. For verifying its validity, suggested method is implemented and applied into a commercial architectural CAD system.

Keywords Automatic design · Architectural CAD · Korean traditional building · Roof frame · Lifting-curve · Waist-curve

E. Ahn (✉)

Department of Information and Communication Engineering, Hanbat National University, San 16-1, Deokmyeong-dong, Yuseong-gu, Daejeon, South Korea

N. Kwak

Division of Information and Communication Engineering, Baekseok University, 115, Anseo-dong, Dongnam-gu, Chungcheongnam-do, Cheonan, South Korea
e-mail: nykwak@bu.ac.kr

1 Introduction

In recent years, new endeavors and trials has been made to convey and develop Korean traditional house. However, the increased cost due to the traditional customized constructing method is one of the major factors that deter the vitalization of the traditional architecture [1]. In addition, designers have no choice but to design architectural components one by one without standardized design system. In the process of traditional house design with CAD program, drawing of a roof frame is the most important and difficult job. Because that the constituent components gather and form a curved roof surface in outline. This study suggests an automatic generation system for creating major architectural components such as rafters and the corners of the eaves (angle rafter or diagonal rafter) in order to make curved roof surfaces of a traditional building.

2 Understanding the Structure of Roof Frame

2.1 Eaves

The eaves may terminate in a fascia, a board running the length of the eaves under the tiles or roof sheets to cap off and protect the exposed rafter ends and to provide grounds on which to fix gutters [2]. The eaves are formed with a number of rafters that stretch out of the columns. The optimum degree of protruding of eaves is about $28\text{--}33^\circ$ between the end point of the eaves and the column on the cornerstone [3] as shown in Fig. 1a. For the reason of the structural stability, there is a limit to extrude eaves. Generally, the distance from columns to end of the eaves is shorter than the distance from the major purlin (Jusim-Dori) to the middle purlin as shown in Fig. 1b. The slope of the roof is influenced by the size of the building and regional weather condition. The shape of roof surface is classified into three types. They are one of the flat, concave, or convex. In Korea, 5-Ryang house is most widely used. 5-Ryang is a house type that adds the middle purlin between the main purlin and highest purlin as shown in Fig. 1. As a result of differences between the slope $s1$ and the slope $s2$ it makes concave roof is formed naturally [3].

2.2 Analysis of Roof Curvature

The Korean traditional roof has three leading types: gambrel roof, hipped roof, and hipped-and-gable roof. While the gambrel roof is appeared in ordinary private houses, the rest two types are used in the large scale building such as temples and palaces. This study analyzes the roof curve of the Korean traditional wooden architecture focused on the roof timber structure of hipped-and-gable roof.

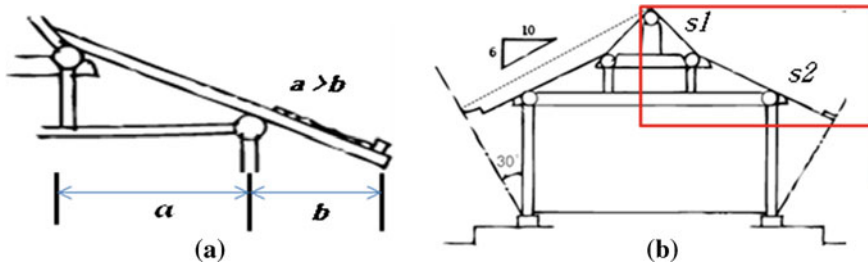


Fig. 1 Relationship between the sticking-out and slope of roof: **a** 5-Ryang House and roof slope, **b** amount of eave’s extrusion

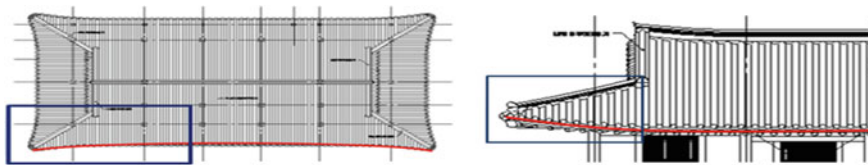


Fig. 2 Waist-curve (left) and lifting-curve (right)

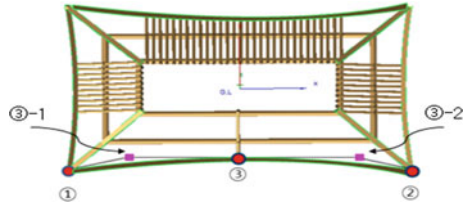
The roof surface of Korean traditional architecture is defined with two curves namely “lifting-curve” and the “waist-curve”. The lifting degree for the eaves of the corner in the front view is called “lifting-curve”. And the curve gently bent to the end of Choonyeo from the middle of eaves in the floor plan image is named “waist-curve”. The amount of lifting-curve and waist-curve depends on the building size and visual effects. Though there is no pattern to decide them, it is usually 1/4 longer than the extruding length in the middle [4] (Fig. 2).

3 Automatic Generation of Roof Frame

3.1 Control Points for Defining Roof Curves

The entire shape of a roof is defined by lifting-curve and waist-curve. Their end points are exactly identical to the end points of Choonyeo that is structurally important component. The roof curves, namely a raised lifting-curve and waist-curve, are three-dimensional curves that connects three control points: two end points of the corner eaves above the column (① and ② in Fig. 3) and an edge point of the rafter in the middle (③ in Fig. 3). Bezier spline is used and its control points are decisively affected by three points: The ① and ② are the end points of the Choonyeo and the point ③ is the end point of the rafter in the middle. Actual inner

Fig. 3 Control points for a roof curve



control points ③-1 and ③-2 of the curves are generated from point ③. For bilateral symmetry, the line from ③-1 to ③-2 is parallel to the purlins.

3.2 The Processing for Components Generation

To go into details, constituent parts involved in making the roof shape are Choonyeo, PyeongGodae, rafter, eaves, GalmoSanbang, and corner rafter. The auto-generating system for a roof frame has a roll to design a framework for making the roof surface by starting from the generating four pivotal component Choonyeos in each four corners. Then the end points of Choonyeos are used for calculating the roof curves. The system decides the shapes of rest components and generating them on the exact positions along the roof curves. The system has users altering the decision variables such as length of sticking out, degree of roof concaveness, angle of lifting-curve, and waist-curve. The auto-generating process of a roof framework is shown in Fig. 4. This figure explains how user defined variables do affect the shape and position of the components for roof frame.

4 Implementation

The proposed method is implemented on the Windows 7 using Geometric Description Language (GDL), a script language for the architectural CAD system, ArchiCAD V12 [5]. The automatic design of the components for roof frame can only be carried out by completion the design of lower structure like as purlin, crossbeam, and column. The lower structure of the building predefines the entire roof shape. User can control the roof shape in detail by decision factor like as degree of eaves extrusion, curvature for the curves and control points of the curves. The suggested automatic generating system creates the constituent components for roof fame in order of Choonyeo, PyeonGodae, eaves, etc. Figure 5 displays a result for roof curves and surfaces that is formed with many constituent components automatically generated by the suggested method.

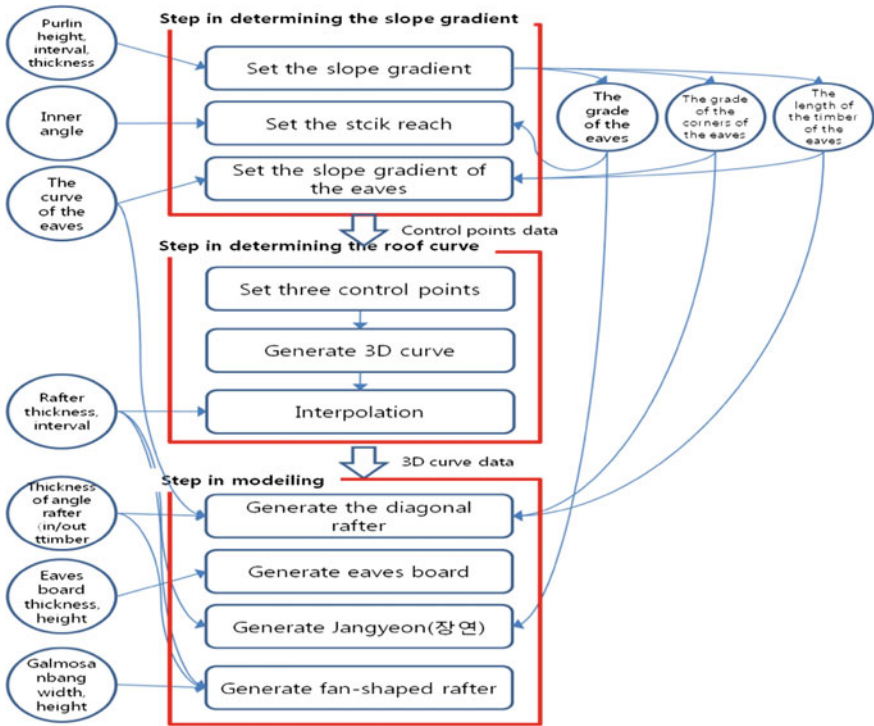


Fig. 4 Process for draw roof frame

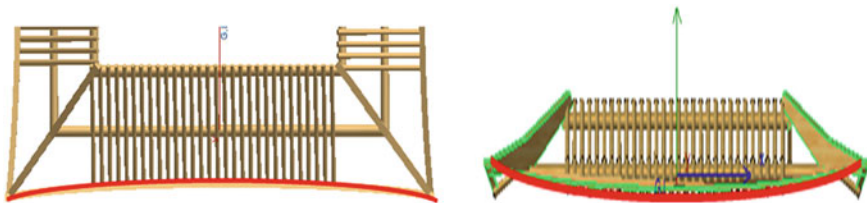


Fig. 5 Auto-generated roof surface that are composed of rafter and Choonyeo

5 Conclusions

This research defined the architectural characteristics and design factors for the roof frame through determining the slope of the roof and the slope gradient of the eaves. And we tried to find out the relations between the slope gradient of roof and the roof curvature in architectural viewpoint. Considering these elementary requisites, we implemented the automatic roof generating system by creating the related constituent components.

Acknowledgments This research is supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0021154).

References

1. Cho JB (2008) Legal and institutional countermeasures to vitalize Korean architecture, architecture and urban research institute, pp 83–96
2. <http://en.wikipedia.org/wiki/Eaves>, 2011
3. Chang GI (2003) Korean wooden architecture V. Boseong-gak, pp 247–324
4. Kim WJ (2000) Pictorial terminology of Korean architecture, Balurn
5. David NC (2003) The GDL Cookbook 3, Marmalade
6. Lee HM, Yang YD, Oh SK, Ahn EY (2012) Automatic generation method of roof shape for hipped-and-gable roof in traditional wooden architecture. In: Conference of Korea Multimedia Society, Spring, pp 90–91

Adaptive Load Partitioning Algorithm for Massively Multiplayer Online Games

Tae-Hyung Kim

Abstract Distributed Virtual Environment Systems are widely used in massively multiplayer online games. With an efficient distributed architecture and load balancing algorithm, they can support tens of thousands of players interacting with each other. An existing prevalent mechanism is to divide the virtual world into several regions and microcells and use the graph-based partitioning algorithm. But they generally limit the assignment beginning with border nodes and can't adapt to the situation when players are crowded around the center of the entire map. Furthermore, many algorithms simply focus on reducing the intra-server communication cost in a connected graph. In this paper, we propose an adaptive load balancing algorithm to solve the problem of the center-crowdedness in the DVE systems, while reducing the intra-server communication and client migration cost simultaneously.

Keywords Distributed virtual environment • MMOGs • Load balancing • Graph partitioning algorithm

1 Introduction

With the high-speed growth of players in massively multiplayer online games (MMOGs), the importance of an efficient load balancing management cannot be overemphasized in the distributed virtual environment (DVE) systems [1, 2].

T.-H. Kim (✉)

Department of Computer Science and Engineering, Hanyang University,
1271 Sa 3-dong, Ansan, Kyunggi-Do, South Korea
e-mail: taehyung@hanyang.ac.kr

In order to support the game worlds with huge numbers of avatars, we need an efficient load balancing and partitioning algorithm which tries to mitigate a heavy requirement on server resources and communicating overloads. The DVE is a parallel and distributed system that allows different users connected through high-speed communication networks to interact with each other in a virtual world. In the paper [1], they formulate the problem of load balancing as a possibly suboptimal assignment of jobs to processors by relocating a set of jobs so as to decrease a *makespan*. Therefore, the kernel problem of load balancing management in DVE is how to shed and distribute overloads in a server to relatively lightly loaded ones with minimal cost. If players are more concentrated than in other regions of the virtual environment, a hotspot is said to be formed.

The formation of hotspot is inevitable and sometimes encouraged because it is so natural for players to tend to gather in a region of strong interests in a virtual game world. Recently, many works focus on how to reduce the intra-server communication cost caused by adjacent cells which are distributed to different servers in partitioning step as an efficient way to minimize load balancing cost. They generally use connected graph-based algorithm to prevent the high correlation cells are assigned to distinct servers [2–5]. Due to recent advances in online game server architecture, servers in the same clusters are always connected by internal network and intra-server communication cost is not significant as before. Therefore, we claim there is a further aggressive step forward over the past correlated cell management.

In this paper, we propose an adaptive load balancing approach that takes into account not only the intra-server cost, but the overhead of client migration for load balancing. We also use a self-adaptive partitioning algorithm to distinguish the situation without border nodes which many graph-based partitioning algorithms cannot be reached. The rest of the paper is organized as follows. [Section 2](#) describes the heart of our mechanism is presented. Brief simulation results on the performance of our approach are presented in [Sect. 3](#) and we conclude that the results look very promising.

2 Our Approach

In this section, we present an adaptive load balancing approach for MMOG systems. Some important definitions are presented to explain the algorithm. An adaptive partitioning is to handle various situations such as no boarder node.

2.1 Area of Interest

In order for users to interact with each other, each client must maintain the local copies of other avatars' states which need to be updated periodically by region servers. If a region server has to broadcast the state information to one and each of

all users, the communication bandwidth will be wasteful. That's why the concept of the area of interest is employed: clients are able to receive and send the needed information update only from the adjacent clients with which are being actively interacted. In fact, the area of interest is a circular area whose center is decided by the coordinates of the player's location in the game world and the radius of circle usually is considered as a vision distance because it is the limit to which the player can see. In this paper, the size of a square microcell is determined by player's vision distance. As a result, a player's area of interest is defined as an 8-direction square group. Figure 1 (left) shows player P's area of interest according to our definition. Players 1–5 are inside of the player P's 8-direction square vision, so they are player P's interacting client, which means all of them are supposed to frequently exchange update data with each other.

2.2 Border Buffer Region

With the area of interest shown above, when an avatar is moving in the border area between two servers, it should receive and/or send state data to the other avatars who are in the adjacent sever border area. It will also cause the intra-server cost. In order to avoid this overhead and provide a seamless gaming experience, we use these border area microcells to build up a *buffer region*. As shown in Fig. 1 (right), when an avatar enter a buffer region, the area of interest will cover microcell which is in the adjacent server buffer region. Therefore not only the server where the avatar resides holds the state update data, but also the adjacent server will copy the avatar's data as a shadow object including establishing a connection with avatar's client.

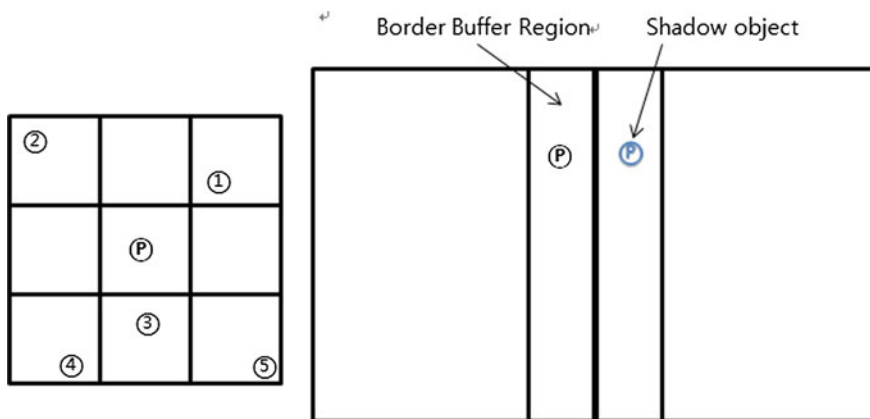


Fig. 1 Area of interest for the player P (left), border buffer region and shadow object (right)

2.3 Definitions

It is essential to give rigorous definitions on various terminologies that are used in our adaptive scheme before presenting the heart of the partitioning algorithm. Here we go:

- **Avatar's weight:** each avatar's weight depends on the frequency of state update and the number of interacting clients. Let $\{A_1, A_2, \dots, A_t\}$ be the set of the avatar P_1 's all interacting clients, then:

$$W_a(P_1) = \sum_{i=1}^t R(P_1, A_i)$$

- **Microcell's weight:** the total weight of a microcell is equal to the sum of avatars' weight in it. Consider that avatars $\{A_1, A_2, \dots, A_N\}$ is in the microcell Mc , the weight of microcell Mc can be defined as:

$$W_{mc}(Mc) = \sum_{i=1}^N W_a(A_i)$$

- **Region server's weight:** the total weight of a region server is equal to the sum of microcells' weight in it. Assume that microcells $\{Mc_1, Mc_2, \dots, Mc_N\}$ is in the Region server S_1 , the weight of Region server S_1 is:

$$W_s(S_1) = \sum_{i=1}^N W_{mc}(Mc_i)$$

- **Edge's weight:** the edge's weight represents the interaction degree between two microcells A and B . Benefited from the 8-direction square group definition of Area of interest in this paper, the value of weight can be simply calculated as the number of avatars interacting with each other which means the sum of A and B 's microcells weight as:

$$Int_{mc}(A, B) = W_{mc}(A) + W_{mc}(B)$$

- **Load thresholds and overload:** a region server's load threshold is determined during the server setting phase by developers. It represents the highest capability which a region server safely maintains game world and can be usually measured as a certain weight value like T_1 . Consider that if region server S_1 :

$$W_s(S_1) > T_1$$

We say that the region server S_1 is overloaded and need to be redistributed.

- **Load balancing cost:** Unlike other graph-based load partitioning approach, the client migration cost is also considered in addition to the overhead of the intra-server cost. Since a client is relocated to another server during the load balancing procedure, a particular client and server need to be disconnected at first, and then reconnected later. This overhead should not be overlooked. One way to reduce this overhead is to build up the border buffer region technique aforementioned and assign border cells to another server during the distribution.

Assume that if we shed loads by border cells whose total weight is $W_{mc}(B)$, the total edge's weight of being assigned microcells with their adjacent microcells is $Int_{mc}(\text{total})$ and the load weight which is needed to be shed is $W_{\text{total}}(S_1)$, the total load balancing cost of server S_1 can be expressed as:

C_s^m = Migration cost for move weight of $W_{\text{total}}(S_1) - W_{mc}(B)$ from S_1 to another adjacent server.

C_s^i = Intra-server cost for edge weight of $Int_{mc}(\text{total})$

$$C_{\text{total}} = W_1 C_s^m + W_2 C_s^i$$

In the above definitions, W_1 and W_2 represents the related importance of the migration cost and the intra-server communication cost. For instance, the server cluster is entirely connected with internal network, and then it is preferable to give a higher number to W_1 . The accurate cost calculation is depended on server's performance which can be derived through a simulation.

2.4 Partitioning Algorithm

Taking into account that servers in the same clusters are generally connected by internal network and intra-server communication cost is not important, we also suggest our partitioning approach begin from the border cells if overloaded server has like most graph-based partitioning algorithm. But to go one step further, we also need to consider that there is no border cell or there is no free adjacent server which leads to moving border nodes become no sense. Therefore, the proposed Adaptive Load Balancing algorithm can be divided into three steps and pseudo code is as Fig. 2.

- (1) There are two sets in our algorithm. Cells in S_1 will remain in the original overload server. Cells in S_d will be allocated to relevant server depend on the cell's server mark. There are also two variables in our algorithm. One records the weight which needs to be shed. One records the weight which will remain original server.
- (2) Lines 5–10 search for the condition of border node matches or not. If the conditions are met, we first let the border cells become distributed node in Lines 7–8
- (3) Lines 11–14 searches to allocate the highest weight cell to the server which is overloaded originally.
- (4) After prior assigning the border cell and the highest weight cell to the relevant set, we propose a partitioning algorithm to balance overloaded region server and select cells being assigned. This algorithm looks like the mix of ProGReGA [4] which is a greedy region growing algorithm and Layering Partitioning (LP) [5]. However, our approach aggressively reduces intra-server cost by considering the client migration cost as well. When a cell in the highest weight of an edge with free neighbor cell that is from the remaining set S_1 , it will check the assignment whether the neighbor cell to S_1 will lead to overload or not

Algorithm: Adaptive Partitioning Algorithm

```

1. if  $W_s(S_1) > T_1$  then
2.   free all cells from  $S_1$  to free_list , initialize set  $S_d$ 
3.    $w\_to\_distribute = W_s(S_1) - T_1$ 
4.    $w\_to\_keep = 0$ 
5.   loop_flag = true
6.   for each server  $S_i$  in  $S_1\_adjacent\_serverlist$  do
7.     if there is border cell in  $S_1$  and  $W_{mc}(Mc) < T_1 - W_s(S_i)$ 
8.       then  $w\_to\_distribute = w\_to\_distribute - W_{mc}(B)$ 
9.          $S_d \cup \{ \text{border cells} \}$  with a server mark
10.      end if
11.    end for
12.    for each cell  $C_i$  in free_list do
13.      select cell  $C_j$  with highest weight which is not the
neighbor of border cell in  $S_d$ 
14.       $S_1 \cup \{ C_j \text{ cells} \}$  and  $w\_to\_keep += W_{mc}(C_j)$ 
15.    end for
16.    while loop_flag == true do
17.      for each cell  $C_i$  in  $S_1$  and  $S_d$  do
18.        select neighbor free cell  $C_j$  with highest  $Int_{mc}(C_i, C_j)$ 
19.        if  $C_i$  from  $S_1$  then
20.           $w\_to\_keep = w\_to\_keep + W_{mc}(C_j)$ 
21.          if  $w\_to\_keep < T_1$  then
22.             $S_1 \cup \{ C_j \text{ cells} \}$ 
23.          else
24.             $S_d \cup \{ \text{remaining cells in free\_list} \}$ 
25.            loop_flag = false
26.          end if
27.        else
28.           $w\_to\_distribute -= W_{mc}(C_j)$  and  $S_d \cup C_j$ 
29.          if  $w\_to\_distribute \leq 0$  then
30.             $S_1 \cup \{ \text{remaining cells in free list} \}$ 
31.            loop_flag = false
32.          end if
33.        end if
34.      end while
35.      Assign cells in  $S_d$  and free cell list to relevant server
36.    end if

```

Fig. 2 Adaptive partitioning algorithm

(lines 17–23). If it is expected to cause an overload, it gives up the assignment and redistributes the remaining cells in free_list to S_d , and exit from the loop. Otherwise, it will assign that neighbor cell to S_1 . Similarly, when the cell which has the highest weight of an edge with free neighbor cell is from S_d , it will check assigning that neighbor cell to S_d will satisfy shedding requirement or not (lines 24–29). If the next movement is supposed to meet the shedding requirement, it will allocate the remaining cells in free list to S_1 and end the loop. The difference is that it will unconditionally assign selected cell to S_d .

- (5) Finally, we distribute all of selected microcells to relevant servers and finish the load balancing task.

3 Experimental Results and Conclusion

This section presents the sketch of our simulation and the results in a simulated virtual world. These virtual worlds are implemented with Microsoft Visual Studio 2005 in multi-server environment. Our multi-server environment consists of a simple 2-machine cluster of Intel core i5 3.20 GHz CPU Windows 7 connected by a 100 Mbps Ethernet Network card. One server is used as a main server to maintain the virtual world. When the virtual environment is overloaded, the simulator is also responsible for shedding the weight to another server, which is used as a secondary server to maintain the virtual world with light weight. When the main server is overloaded, it is responsible for receiving the shed weight from main server.

In the previous section, we already define a standard cost model to measure the performance of a load balancing algorithm. In order to prove this abstract cost formula is sound, we use three metrics related to the cost factors throughout the simulation—the completion time, the network bandwidth and the CPU utilization. Since most online games are based on client/server architecture, the server and network performance are of our prime interest. More discussions on the metrics issues are proposed in [6].

In this simulation, we use a small virtual world with a dimension of 6*6 cells. The load threshold of main server is set randomly. The overloaded number of clients in this virtual world is floating between 20 and 40. We also compare our algorithm to the ProGReGA algorithm by Carlos and Claudio [7] and the LP by Lui and Chan [4]. Our simulation step is divided into three phases as follows:

- (1) Initialize main server and minor server. Generate a 6*6 dimension virtual world with a certain number of players distributed into some cells randomly in the main server. The main server begins to maintain the virtual world and interacts with minor server.
- (2) Add a random number of new avatars into main server distributed in some cells randomly yet and overload the server. Main server starts the load balancing schedule because main server's load exceeds the safe threshold. It will shed part of the load to minor server using a predetermined algorithm.
- (3) When the load balancing schedule is completed successfully, the main server obtains three metrics value mentioned above and display the results.

Table 1 shows the CPU utilization. In our simulation, we vary the total number of avatars between 50 and 300. Obviously shown in the table, the CPU utilization is not a bottleneck in case of 50–300 avatars in a virtual world. The CPU utilization of three different algorithms maintains as much as around 50 %.

Table 2 presents the system cost that follows the definitions in the previous section. The main server handles 150–250 avatars randomly. We compare three algorithms with four experiment iterations. Except the first iteration, we can see that our adaptive partitioning algorithm shows the lowest system cost among the three. This is due to using two sets to find the next most suitable microcell to being

Table 1 CPU utilization

Algorithm	AP	ProGReGA	LP
Average CPU utilization (%)	50	49	48

Table 2 System Cost

	Iteration 1			Iteration 2			Iteration 3			Iteration 4		
	AP	PG	LP	AP	PG	LP	AP	PG	LP	AP	PG	LP
$W_1 = 0.5W_2 = 0.5$	215	215	130	113	113	127	104	112	113	75	95	102
$W_1 = 0.8W_2 = 0.2$	97	97	75	54	62	66	49	54	59	34	47	55
$W_1 = 0.2W_2 = 0.8$	332	332	185	171	164	187	158	171	167	116	143	149

Table 3 Completion Time and Bandwidth

Simulations count	1			2			3			4		
Algorithm	AP	PG	LP	AP	PG	LP	AP	PG	LP	AP	PG	LP
Completion time (ms)	429	446.5	431.8	476.5	519.25	512	518.2	526.72	524.5	711.85	721.4	741.6
Network bandwidth (KB)	101	161	103	188	323	178	346	450	329	996	588	870

assigned, unlike the LP and the ProGReG Algorithm that use only one set to find the next cell instead.

In Table 3, we assume that completion of the load balancing task means that all clients selected are assigned into relatively under-loaded adjacent server when a main server is overloaded. We use a MFC timer to record the total completion time. We assume that the required packet size for one avatar is 1 KB. Clients will send and/or receive the 1 KB packet to other clients within its area of interest. We observe the total size of the packets transferred between a main server and its secondary server in order to evaluate the intra-server communication cost. After four iterations for testing one and each of the three algorithms, we obtain similar results as shown in Table 2 that our adaptive partitioning algorithm achieves the best performance. Based on our experimental results, we claim that the load balancing algorithm evaluation model is reasonable enough and our adaptive partitioning algorithm is more efficient among the representative connected graph-based algorithms.

References

1. Aggrawl G, Motwani R, Zhu A (2003) The load balancing problem. In: Proceedings of ACM symposium on parallel algorithms and architectures, pp 258–265
2. Lety E, Turletti T, Baccelli F (1999) Cell-based multicast grouping in large-scale virtual environments, Technical Report No. 3729, INRIA, France
3. Vleeschauwer B, Bossche V, Verdickt T, Turck F, Dhoedt B, Demeester P (2005) Dynamic microcell assignment for massively multiplayer online gaming. In: Proceedings of the 4th ACM SIGCOMM workshop on network and system support for games, pp 1–7
4. Lui J, Chan M (2002) An efficient partitioning algorithm for distributed virtual environment systems. *IEEE Trans Parallel Distrib Syst* 13(3):193–211
5. Chen J, Wu B, Delap M, Knutsson B, Ku H, Amza C (2005) Locality aware dynamic load management for massively multiplayer games. In: Proceedings of the 10th ACM SIGPLAN symposium on principles and practice of parallel programming, pp 289–300
6. Ye M, Cheng L (2006) System-performance modeling for massively multiplayer online role-playing games. *IBM Syst J* 45(1):45–58
7. Bezerra C, Geyer C (2009) A load balancing scheme for massively multiplayer online games. *Multimedia Tools Appl* 45:263–289
8. Ta D, Zhou S, Shen H (2006) Greedy algorithms for client assignment in large-scale distributed virtual environments. In: Proceedings of the 20th workshop on principles of advanced and distributed simulation, pp 103–110
9. Huang J, Du Y, Wang C (2003) Design of the server cluster to support avatar migration. In: Processing IEEE virtual reality, pp 7–14

Game-Based Learning System Using Graduated-Interval Recall Method

Ming Jin and Yoon Sang Kim

Abstract Currently, research is actively being conducted on applying games to education. However, existing game-based learning systems mainly focus on how to increase learning motivation and enhance engagement, while achieving less success on how to make further improvements on the effectiveness of learning. To improve on these points, this paper proposes a game-based learning system using graduated-interval recall method to organize game contents and learning courses with appropriate review intervals according to learners' status. The proposed system has been verified through comparative learning effectiveness tests between two learning groups, in which one group engaged in learning with the proposed system and the other with an existing system.

Keywords Game-based learning · Graduated-interval recall · Spacing effect

1 Introduction

As games have been shown to be highly effective in education, studies are being actively conducted on applying games to education (game-based learning). Many educators and instructional designers have been developing and using digital games in schools, institutions of higher education, and commercial applications to

M. Jin · Y. S. Kim (✉)
Human Interaction Lab, Department of Computer Science and Engineering,
Korea Tech, Cheonan, South Korea
e-mail: yoonsang@koreatech.ac.kr

M. Jin
e-mail: kmyeng@kut.ac.kr

promote learning achievement [1]. In particular, innate characteristics of games, including competitiveness and enjoyment, can highly increase the motivation to learn by stimulating and enhancing learners' desire to embark on challenge and engagement.

Existing game-based learning systems have been used to enhance learning effectiveness in various ways—such as increasing learners' motivation by reforming study contents into game scenarios and prompting competition between participants [2, 3], or improving system architecture for better integration between game context and the learning environment [4, 5]. However, existing game-based learning systems lacks the effective learning method that should be adopted into the game system in order to organize learning contents in accordance with learning progress for enhancing learning effectiveness.

A game-based Mandarin learning system developed by Annisa Dwiana and Dalbir Singh [6] has shown a good game-based learning system design for attracting children's attention and increasing their engagement by integrating graphic, text, and animation effects. However, the system did not adopt a proper learning method that could help children memorize the words they learned.

In [7], it was discovered through result analysis from the test of proposed system that there was little improvement in learning performance in spite of increasing play times with the proposed game-based learning system, which had been proven to enhance learning motivation using many game features, such as roll playing and competitive activities. The authors claimed that further research on improving learning effectiveness would be necessary.

From above analyses, we found that existing game-based learning systems are mainly focused on how to improve learning motivation and engagement by making use of game playing, while achieving less success on how to enhance learning effectiveness with proper learning methods. To improve on these points, this paper proposes a game-based learning system using graduated-interval recall method to organize game contents and learning courses with appropriate review intervals in accordance with learning status. The proposed system has been verified through comparative tests on learning effectiveness between two learning groups, in which one group engaged in learning with the proposed system, and the other with an existing system.

2 Proposed Game-Based Learning System

In a study published in 1995, German psychologist Hermann Ebbinghaus identified the phenomenon referred to as spacing effect, whereby humans and animals remember or learn things more easily by studying them a few times over a long period of time rather than repeatedly in a short period [8]. Utilizing the spacing effect, the graduated-interval recall method published by Pimsleur progressively expands the review time interval. (For example, the review interval pattern used in his paper was 5 s, 25 s, 2 min, 10 min ...) [9]. The biggest advantage when

learning with the graduated-interval recall method is not only stronger consolidation of long-term memorization of acquired knowledge achieved by expanding the review interval, but also enhanced learning effectiveness by reducing unnecessary reviews and guiding learners to focus on less acquired knowledge.

The original progressively increasing time interval adopted by Pimsleur's graduated-interval recall method has the following the pattern.

$$T_1, T_2 = 5 \times T_1, T_3 = 5 \times T_2, \dots T_n = 5 \times T_{(n-1)} \quad (1)$$

In order to apply Eq. (1) in the testing environment of the proposed system, a modification was made to the increasing pattern of review time interval while maintaining its basis principle. The modified increasing pattern of review time interval used in the proposed system is as follows:

$$T_1 = 1 \times R, T_2 = 2 \times R, T_3 = 3 \times R, \dots T_{(n-1)} = (n - 1) \times R, T_n = n \times R \quad (2)$$

The R in Eq. (2) represents the minimum time unit used for measuring review interval in the proposed system. Then each knowledge point that needs to be acquired from the game is given a master degree (MD) property and a hidden ratio (HR) property. The each knowledge point's MD indicates the mastering level of corresponding knowledge point; a higher value means better memorization of that particular knowledge point. HR changes in accordance with MD; if MD is set to n, HR will also be set as n, which means this knowledge point will no longer be shown in the game during the n-th review time interval duration in the pre-defined review time intervals (T1, T2, T3, ... Tn). The MD of a knowledge point will be increased or decreased dynamically according to user's interaction judgment result for this knowledge point. If the learner's interaction judgment result from this knowledge point shows a good memorization status, its MD and HR will both be increased by a step (Fig. 1).

Thus, the proposed system expands the review time interval of a well-learned knowledge point, reduces the number of unnecessary exposure to the learner. In the opposite case, it makes a knowledge point not fully acquired by the learner is shown more frequently to enhance learning effectiveness.

3 Experiment and Discussion

3.1 Experiment Environment and Protocol

To validate the proposed game-based learning system, we developed a game on the topic of seven major continental plates for testing (Fig. 2).

The system was developed with HTML5, and the experiment environments are as follows:

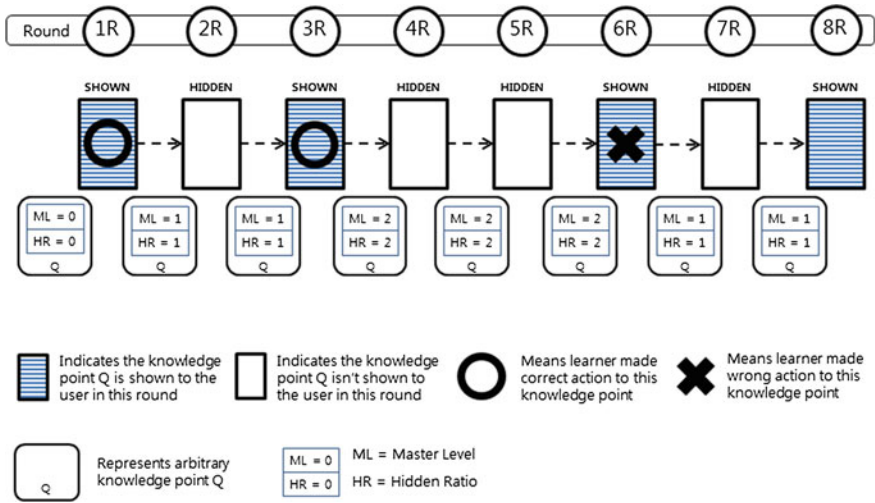


Fig. 1 The mechanism of the proposed game-based learning system using graduated-interval recall method



Fig. 2 Screen shot of the game spin the continental plates, which was developed to test and validate the proposed game-based learning system

- There are 7 plates, each with 4 descriptions, which means there are $4 \times 7 = 28$ descriptions of the plates altogether.
- The game involves the learner trying to match the descriptions and the corresponding continental plates. Thus, the 28 descriptions on 7 plates are the knowledge points, and each description has MD and HR values.
- During game play, an arbitrary description will fall down from the top of the screen. At the bottom of the game screen is a rotatable wheel made up of 7 plates, and the learner tries to match the plate that corresponds with the description falling down by rotating the wheel using the keyboard.
- When the description falls down to the wheel, the interaction result judgment of the (knowledge point is fired. If the correct plate is chosen, it is determined that the learner is acquainted with the knowledge point, and the MD and HR values are increased by a step. (For example, if the description says “The smallest plate among the 7 major plates,” the learner should match it with the “South American Plate”.)
- When the learner provides a wrong answer, the correct plate is shown on the screen to remind the learner.

In the proposed game, knowledge points (N) fall down in each round at certain speeds. Then the interaction time (t) of each description will be equals to others, and the same spent time (T) will results as follows:

$$T = N \times t \quad (3)$$

We set the total number of descriptions falling down (N) during each round of the game at 28, which is equal to the total number of knowledge points. In turn, we can make use of the number of game rounds for representing the review time interval from the duration (T) spent on each game round, which is a constant. Thus, the review time interval in the proposed game is progressively increasing, as shown by Eq. (4).

$$R_1 = R \times 1, R_2 = R \times 2, R_3 = R \times 3 \dots, R_7 = R \times 7 \quad (4)$$

The review time interval is measured in rounds in this game, and the knowledge points are shown to the learner. For example, if the review time interval of an arbitrary knowledge point from the current moment to the next is R3, the knowledge point will be shown again after 3 game rounds—in the 4th round. The knowledge point shown to the learner in each round is determined in real time according to the MD and HR values of the knowledge point, as explained in [Chap. 2](#). If the learner provides the correct answer for arbitrary knowledge point Q1 during a round, its MD and HR values will increase from M0, H0 to M1, H1, respectively, and the knowledge point will not be shown again during time interval R1. Then, a knowledge point Q2, which has never been answered correctly by the learner, can be shown instead of Q1, providing increased exposure to the learner.

For comparative analysis, we implemented a game-based learning system without incorporating the graduated-interval recall method. It has the same scenario (matching descriptions to continental plates) as the game explained above; however, every description is shown once in each round with the order of their reveal sequence varying randomly.

The experiment was conducted by dividing test subjects into two groups: Group A engaged in learning with the conventional game-based learning system, and Group B with the proposed graduated-interval recall method game-based learning system. Both groups took a pre-test on the topic, which was made up of 28 knowledge points and had a form identical to the objective test (Fig. 3). The test quizzes were presented in random orders to reduce unwanted influence. The subjects scored between 4 and 7 points in the pre-test, indicating that they had little knowledge on continental plates.

- (1) After the pre-test, both groups learned about the topic with corresponding game-based learning systems.
- (2) The learning duration was 25 min for both groups.
- (3) To reduce influence, the two groups learned about the topic in the same place and were not aware that they were using two different learning systems.
- (4) After completing the learning session, the subjects took a post-test, which contained questions identical to those in the pre-test but in a randomly varied order.
- (5) After the post-test, results were collected and compiled.

3.2 Experimental Results

The test results are shown in Table 1. The performances of the subjects that engaged in learning with the proposed graduated-interval recall method game-based learning system (Group A) were better than those who used the conventional game-based learning system. (Group B) (The average score of Group A was up to 50 % than Group B.)

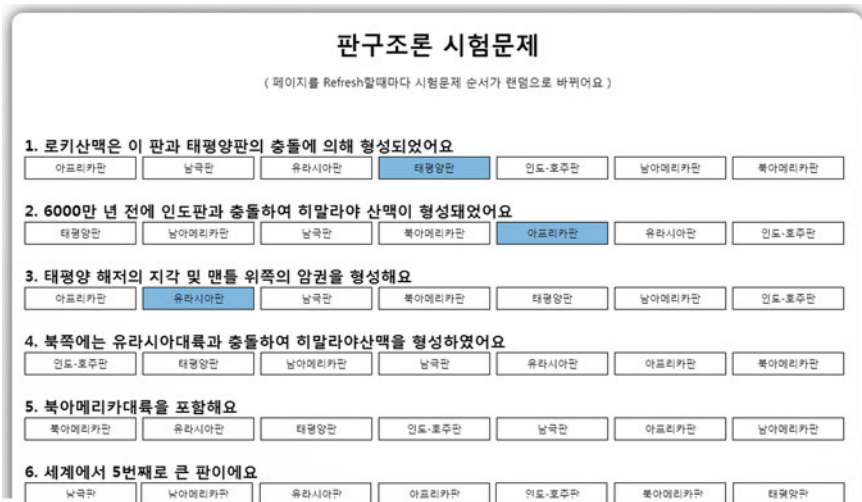


Fig. 3 Quizzes used for pre-test and post-test (in Korean)

Table 1 Comparison between score increments of the proposed system and the conventional system

		Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Max	Min	Avg. Increment
Group A	Pre-test	4	5	4	6	5	13	7	10.6
	Post-test	11	17	17	15	17			
	Increment	7	12	13	9	12			
Group B	Pre-test	5	7	6	4	5	18	13	15.8
	Post-test	22	25	21	17	21			
	Increment	17	18	15	13	16			

4 Conclusion

This paper proposed a game-based learning system using the graduated-interval recall method. To validate the effectiveness of the proposed system, we conducted an experiment that compared the learning performances between the proposed system and a conventional system had been conducted. From the results, we were able to confirm that the subjects engaged in learning with the proposed system scored approximately 50 % higher than those who used the conventional system did.

In addition to science education, the proposed system is supposed to be valid for all academic areas, including learning foreign languages. While our study involved subjects acquiring 28 knowledge points in 25 min, it is expected that the performance gap between the two systems would increase with greater number of knowledge points and longer learning time.

References

1. Wang T-L, Tseng YF (2011) Learning effect for students with game-based learning on meta-analysis, ICCSE, 2011
2. Lin K-C (2011) Online interactive game-based learning in high school history education: impact on educational effectiveness and student motivation, U-Media 2011
3. Lin K-C (2010) Using competitive digital game-based learning to improve motivation, frontier computing. Theory, technologies and applications, 2010
4. Takaoka R, Shimokawa M, Okamoto T (2011) A framework of educational control in game based learning environment, ICALT D, 2011
5. Liu G, Jiao Z, Liu S (2009) Tutoring strategy study on game-based experiential learning in vocational school, ETCS, 2009
6. Dwiana A, Singh D (2011) Computer game based learning approach for mandarin language, electrical engineering and informatics, 2011
7. Wu S-C, Tsuei M (2011) The effects of digital game-based learning on elementary students' science learning, ICECE, 2011
8. Wikipedia: spacing effect
9. Pimsleur P (1967) A memory schedule. Mod Lang J 51:73–75

An Efficient DRAM Converter for Non-Volatile Based Main Memory

Sung-In Jang, Cheong-Ghil Kim and Shin-Dug Kim

Abstract The new memory technologies having the characteristic of non-volatile such as Phase-change RAM (PRAM), Ferroelectric RAM (FRAM), Magnetic RAM (MRAM) and Resistive RAM (RRAM) that can be replaced the DRAM as main memory have been emerged. Among these memories, PRAM is especially the most promising alternative for DRAM as main memory because of its high density and low power consumption. On the other hand, the slower latency by comparison with DRAM and endurance are caused to reduce performance. In order to exploit these degradations of performance, we propose a hybrid memory system consisting of PRAM and DRAM as a converter. The DRAM converter is comprised of an aggressive streaming buffer to assure better use of spatial locality and an adaptive filtering buffer for better use of temporal locality. Our architecture is designed to enhance the long latency as well as low endurance of PRAM. The proposed structure is implemented by a trace-driven simulator and experimented by using SPEC 2006 traces. Our experimental results indicate that it is able to achieve reducing access count by about 65 %, compared with only PRAM-based main memory system. According to this result, our proposed memory architecture can be used to substitute for the current DRAM main memory system.

S.-I. Jang (✉) · S.-D. Kim (✉)

Department of Computer Science, Yonsei University, 134, Shinchon-dong,
Seodaemun-gu, Seoul 120-749, South Korea
e-mail: mckoala@yonsei.ac.kr

S.-D. Kim

e-mail: sdkim@yonsei.ac.kr

C.-G. Kim (✉)

Department of Computer Science, Namseoul University, 21, Mae Ju-ri, Seonghwan-eup,
Seobuk-gu, Cheonan-si, Chungnam 331-707, South Korea
e-mail: cgkim@nsu.ac.kr

Keywords Non-volatile memory · Phase-change RAM (PRAM) · DRAM · Main memory · Converter · Spatial locality · Temporal locality

1 Introduction

The commercial products like smart phones or tablet PCs are adopting Flash memory as in the role of Hard Disk Drive. Although Flash memory is non-volatile memory, it supports only the page unit of read/write [1]. Furthermore, the erase operation performs in term of block that is much larger than a page. In other words, Flash memory is not expected to show similar performance DRAM-based main memory.

Recently, the new memory technologies such as Phase-change RAM (PRAM), Ferroelectric RAM (FRAM), Magnetic RAM (MRAM) and Resistive RAM (RRAM) have been developed in order to replace the previous memories. In general, no-volatile memories have common advantages, including non-volatility, high density, low power and large capacity. Because of these merits, the non-volatile memories can be widely used in the future to replace the conventional memories (DRAM, SRAM, Flash, etc.). Among the new memory technologies, PRAM is one of the most promising candidates for future main memory which have similar performance with DRAM because it shows the best advantages in terms of density, power consumption, and non-volatility. Nevertheless, much slower latency than DRAM (especially, write latency) and endurance must be overcome to be used as main memory [2, 3].

In this paper, our described DRAM converter (Adaptive DRAM Converter) is located between the last level cache and PRAM-based main memory in hierarchy. The Adaptive DRAM Converter which is composed of an aggressive fetching superbloc buffer (AFSB) and a selective filtering buffer (SFB) is designed for showing similar performance of conventional DRAM-based main memory and reducing the limitations of PRAM. The AFSB exploits spatial locality aggressively by fetching superbloc from PRAM, and the other exploits temporal locality adaptively [4]. The proposed structure is simulated by using a trace-driven simulator with SPEC 2006 traces [5–7].

According to our simulation results, the access count can be reduced by around 65 % compared to only PRAM-based main memory system. Therefore, the proposed Adaptive DRAM converter can effectively hide the access latency of non-volatile memory using small buffer.

The rest of this paper is organized as follows: In Sect. 2, we will review some of previous studies regarding non-volatile memory PRAM. Section 3 presents the overall architecture and operation flow. In Sect. 4, performance evaluation is shown. Finally, we conclude in Sect. 5.

2 Related Work

There are many subsequent studies regarding non-volatile memory PRAM. We will briefly review along with the works to overcome the limits of PRAM.

Qureshi et al. [8] proposed a hybrid main memory system by utilizing non-volatile memory PRAM and DRAM. It explores the trade-offs for main memory system consisting of PRAM storage coupled with a small DRAM buffer. In this architecture, the DRAM is structured like a cache to hide read latency. The write queue is for hiding the write latency. In addition to, the fine-grained wear-leveling algorithm can supplement the endurance limits of PRAM. Their proposed scheme shows results that reduce write access latency by three times and improve the endurance from 3 to 9 years. However, it tends to be expensive because of combining 32 GB of PRAM and 1 GB DRAM.

Seok et al. [9] proposes a new page caching algorithm for hybrid main memory. Page monitoring and migration scheme was suggested to overcome the long latency and low endurance of PRAM. This scheme is to keep read-bound access pages to PRAM. The results show that it can maximally reduce the total write access count by 48.4 %. Therefore, they can enhance the average page cache performance and reduce the endurance problem in the hybrid main memory system.

3 Proposed Scheme

This section will present the overall architecture and operation flow. As explained before, our goal is to substitute volatile memory DRAM with non-volatile memory PRAM for main memory. To design, DRAM buffer is needed between the last level cache layer and main memory layer. Figure 1a shows our overall architecture.

3.1 Adaptive DRAM Converter

Figure 1b shows the structure of the proposed adaptive DRAM converter. As shown in the figure, AFSB is fetching superblock from PRAM. The superblock size is 32 KB and is consist of 256 sub-blocks which is a basic unit managed by this converter. A superblock is defined as a unit of data formed from a set of pages transferred between the converter and main memory. The sub-block unit is the same size as the last level cache block (setting 128 Byte).

When the requested data is not in AFSB and SFB, superblock is fetched from PRAM-based main memory. The eviction algorithm applied to the AFSB is least recently used (LRU). Based on LRU, an evicted superblock among the existing superblocks in the AFSB is divided hot or cold status blocks by threshold value that is the average access count of each sub-block. In case of invalid sub-block are just discarded because a clean copy exists in main memory. In case of valid

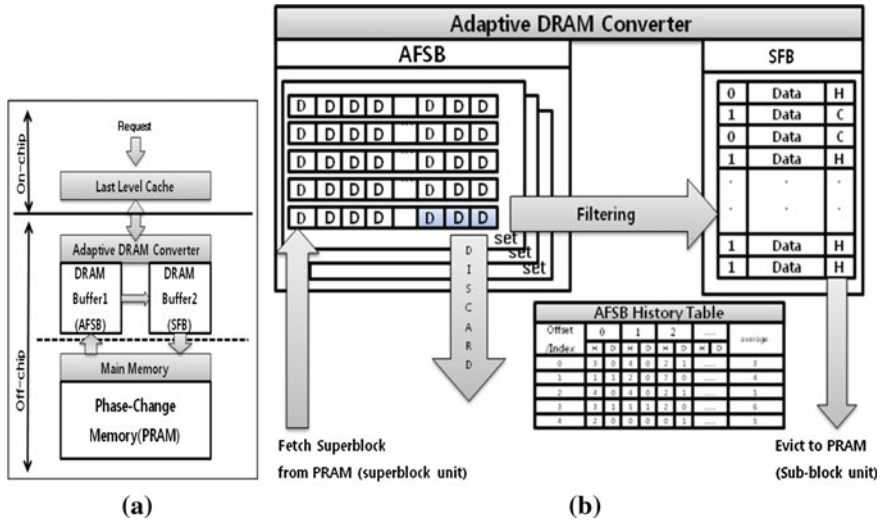


Fig. 1 Proposed model **a** overall architecture, **b** Adaptive DRAM converter

sub-blocks (hot or cold blocks) moves to SFB. When SFB is full, victim block is selected by First-In-First-out (FIFO) algorithm. If sub-blocks are changed while staying in the SFB, then the SFB plays a role similar to a write buffer.

In this process, threshold value is the criteria of dividing unnecessary and reusable data. For calculating average access count which is threshold, AFSB History Table records hit counts of each sub-blocks. The table also manages a hit bit and a dirty bit associated with each sub-block in the AFSB. If hit dirty bit is set, these sub-blocks were accessed from higher memory layer at least once.

In other words, our structure has two buffers which are AFSB, SFB. Because AFSB fetches a group of pages at once, these pages may contain reusable data and/or unnecessary data. Reusable data tend to be accessed again from last level cache because of temporal and spatial locality.

3.2 Operation Flow

In this section, we introduce a fundamental operational flow, which is specially designed for the proposed Adaptive DRAM Converter. The operation flow is as follow:

- First step: When a data is requested from last level cache, Check whether there is an data in AFSB. If the data is in AFSB, go to Fourth step. Otherwise, go to second step.
- Second step: Check whether there is an data in SFB. If the data is in SFB, go to Fourth step. Otherwise, go to Third step.

- Third step: miss managing.
- Fourth step: The adaptive DRAM converter transfers the requested data to the last level cache.

Detailed miss managing flow is as follow:

- First step: The AFSB chooses victim superblock by LRU replacement policy.
- Second step: The converter classifies sub-blocks in evicted superblock as reusable and unnecessary by using AFSB History Table. The valid sub-blocks moves to SFB and the invalid sub-blocks are just discarded. Also, the sub-blocks moving to SFB maintain hot or cold state according to threshold value.
- Third step: The victim block is substituted for a new superblock from main memory.

4 Experimental Result

In this section, we will describe the configuration of simulation and the results in terms of miss rate and access latency. Our results were implemented and simulated by trace-driven simulator.

4.1 Configuration

As shown Table 1, we extracted SPEC 2006 traces by using GEM5 is full simulator [6, 7]. The extracted traces are run on in-house trace-driven simulator [5]. The L1 instruction and data cache are 32 KB with 64 Byte block size and 4-way set associative. The L2 unified cache is 1 MB with 128 Byte block size and 8-way set associative.

4.2 Miss Rate Evaluation

The first evaluation metric is the miss rate. The adaptive DRAM converter is 20 MB consisting of 4 MB AFSB and 16 MB SFB. The 20 MB DRAM buffer is sufficient to show our efficiency of suggested architecture. Figure 2 shows that it is clearly demonstrated. The figures are the result of miss rates in contrast with the last level cache miss rate. The miss rates reduced about 65 %.

Table 1 Baseline configuration

Simulator	GEM5(Quad-core chips), Trace-driven simulator
Benchmark	SPEC 2006 benchmark suites
L1 cache	32 KB, 4-way, 64 Byte block size
L2 unified cache	1 MB, 8-way, 128 Byte block size

Fig. 2 Adaptive DRAM converter miss rates compared to LLC miss rates

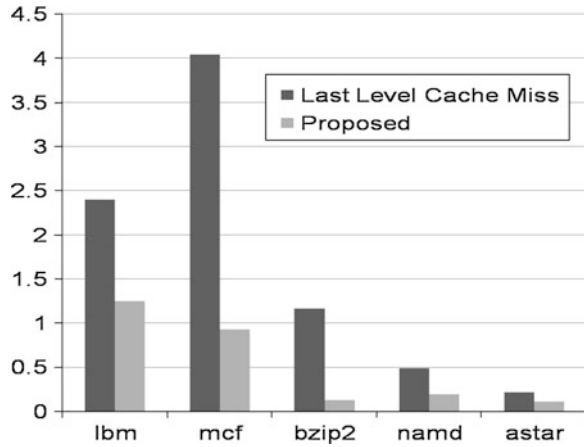
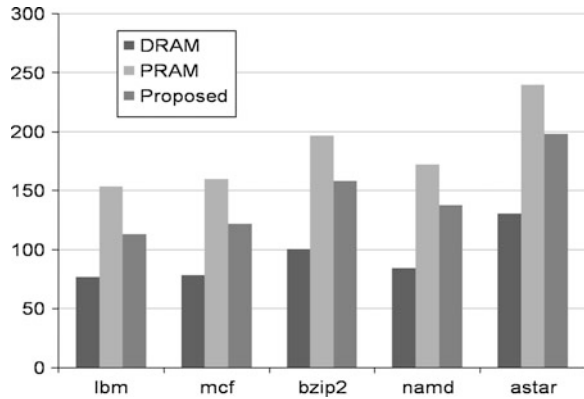


Fig. 3 Access latency of only DRAM, only PRAM and proposed main memory system



4.3 Access Latency Evaluation

Figure 3 presents the access latency in case of DRAM-based main memory, PRAM-based main memory and PRAM having DRAM converter. The memory access parameters are chosen as follows [2]. As mentioned, one of the limitations of PRAM is the very slow latency. Our results show that access latency tends to decrease about 21 %. Even though the suggested architecture is slower than DRAM-based main memory, we obtained much better figures compared to PRAM-based main memory. The predictions for future PRAM cells suggest significant latency improvement because of ongoing development of technology.

5 Conclusion

As the new memory technologies emerged, many studies are progressing for replacing the conventional memory such as DRAM, SRAM and Flash. In this paper, we suggested the PRAM-based main memory system. To realize our architecture, the limitation of PRAM must be solved. To overcome, our model presents a small Adaptive DRAM Converter. It is composed of two buffers that are AFSB and SFB. The AFSB exploits spatial locality aggressively by fetching superblock from PRAM, and the other exploits temporal locality adaptively. According to the results, our scheme can reduce the access count by 65 % and access latency by 21 %.

References

1. Park YW, Lim SH, Lee C, Park KH (2008) PFFS: a scalable flash memory file system for the hybrid architecture of phase-change RAM and NAND flash. In: SAC '08: Proceedings of the 2008 ACM symposium on applied computing, pp 1498–1503
2. Freitas RF, Wilcke WW (2008) Storage class memory: the next storage system technology. *IBM J Res Dev* 52(4.5):439–447
3. NcPRAM http://www.samsung.com/global/business/semiconductor/products/fusionmemory/Products_NcPRAM.html
4. Jung KS, Park JW, Weems CC, Kim SD (2011) A superblock based memory adapter using decoupled dual buffers for hiding the access latency of nonvolatile memory. In: Proceedings of the world congress on engineering and computer science
5. Colmenar JM, Risco Martin JL, Lanchares J (2011) An overview of computer architecture and system simulation. *SCS M&S Mag* 2(02):74–81
6. Gem5 simulator system <http://www.gem5.org/>
7. SPEC CPU (2006) <http://www.spec.org/cpu2006/>
8. Qureshi MK, Srinivassan V, Rivers JA (2009) Scalable high performance main memory system using phase-change memory technology. In: Proceedings of the 36th annual international symposium on computer architecture
9. Seok HC, Park YW, Park KH (2011) Migration based page caching algorithm for a hybrid main memory of DRAM and PRAM. In: Proceedings of SAC, pp 595–599

An M2M-Based Interface Management Framework for Vehicles with Multiple Network Interfaces

Hong-Jong Jeong, Sungwon Lee, Dongkyun Kim
and Yong-Geun Hong

Abstract The Intelligent Transportation System (ITS) can be considered the most representative example of Machine-to-Machine (M2M) applications in a standardization and commercial market. To support the applications, utilizing multiple network interfaces and providing appropriate network connectivity to them are the key issues in order to meet different network requirements of M2M applications on ITS devices. Moreover, different from applications which are controlled by a human operator, the ITS device with M2M application is required to decide the appropriate air interface and network for each application by itself according to the predefined policy and network status. In this paper, we therefore propose a novel multiple network interface management framework for ITS devices with M2M applications having their different requirements and constraints based on the CALM architecture. This can provide the M2M applications with extensible and flexible system environments based on the CALM architecture.

H.-J. Jeong

Department of Computer Engineering, Kyungpook National University,
80, Daehak-ro, Buk-gu, Daegu, South Korea
e-mail: hjjeong@monet.knu.ac.kr

S. Lee

School of Electrical Engineering and Computer Science, Kyungpook
National University, 80, Daehak-ro, Buk-gu, Daegu, South Korea
e-mail: swlee@monet.knu.ac.kr

D. Kim (✉)

School of Computer Science and Engineering, Kyungpook National
University, 80, Daehak-ro, Buk-gu, Daegu, South Korea
e-mail: dongkyun@knu.ac.kr

Y.-G. Hong

Electronics and Telecommunications Research Institute, 218 Gajeong-ro,
Yuseong-gu, Daejeon, South Korea
e-mail: yghong@etri.re.kr

Keywords M2M · Machine-to-machine · ITS · Communication · Multiple interfaces · CALM

1 Introduction

Machine-to-Machine (M2M) communication is characterized by low power, low cost, and low human intervention, supporting a wide range of applications in different domains such as u-healthcare, intelligent transportation system, smart energy, and etc. [1]. Among them, Intelligent Transportation System (ITS) can be the most representative example of M2M applications in a standardization and commercial market. ITS can be used to prevent vehicular accidents, increase the efficiency of transportation system and reduce environmental pollution, while improving passengers' convenience [2]. ETSI, European Telecommunications Standards Institute, considers several use cases of automotive applications in M2M capable networks such as electricity vehicle charging, fleet management, theft tracking, and information exchanges [3]. They have different sizes of data transmitted by each application (from several bytes to hundreds of megabytes). Moreover, the urgency of data is also subjective to the requirements of the applications.

In particular, the architecture of Communication Access for Land Mobile (CALM) is defined by ISO TC204 WG16 standard body, where ITS devices such as Road Side Unit (RSU) and On-Board Unit (OBU) can be equipped with multiple network interfaces using various access technologies to provide network connectivity [4]. CALM-compliant systems provide the ability to use the heterogeneous access technologies for data delivery with different characteristics with each other in terms of network coverage, cost, bandwidth, jitter, and etc. [5]. Therefore, an ITS device should be able to manage and use multiple network interfaces efficiently. Since each ITS M2M application has different network requirements and constraints, managing of multiple interfaces and providing appropriate network connectivity to the application based on its requirement are the key issues to providing ITS M2M applications. Moreover, as the network condition constantly changes due to continuous movements of vehicles, an ITS device with M2M applications (hereafter, called an ITS M2M device) should keep monitoring the network condition and use the best suitable network for an M2M application according to the network selection policy without any human operation.

Therefore, the ITS M2M device should be able to periodically update the network information and the policies required for selecting a suitable network regardless of manufacturers and network providers. In this paper, we propose a new multiple interface management framework for ITS M2M device based on the CALM architecture. To the best of our knowledge, this is the first trial involving multiple interface management for ITS M2M devices.

2 Related Work

The CALM architecture is an initiative to define a set of communication protocols and interfaces which are used for ITS communications [4]. Particularly, unlike the other ITS architectures such as WAVE standard [6–9], the CALM architecture can support multiple air interfaces for a variety of communication scenarios. Hence, in each of communication scenarios which have different requirements and constraints, the CALM architecture can provide the applications with appropriate communication service through interface selection.

In the CALM architecture, all communications between communication entities, called ITS stations (ITS-S) which include vehicles, personal devices, RSUs, and central stations are performed in a peer-to-peer manner. In order to support the peer-to-peer communications among them, network protocol stacks on ITS-S and their relationships should follow the ITS-S reference architecture which is defined in the CALM architecture standard.

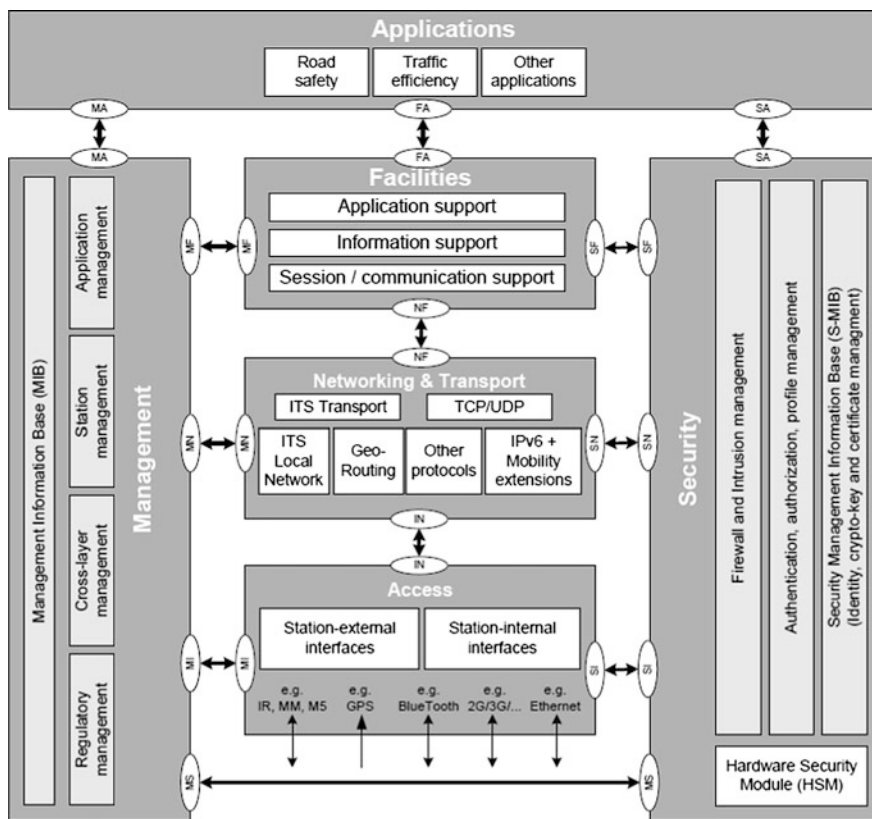


Fig. 1 Reference architecture of ITS station

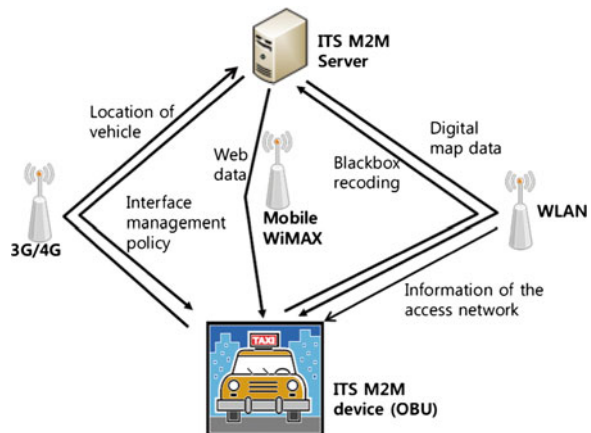
Figure 1 shows the reference architecture of ITS-S. As shown in Fig. 1, the first layer of the ITS-S architecture, called Access layer, represents the physical and link layer. Link control protocols and various physical interfaces such as Bluetooth, 3G/4G and WLAN are comprised in this layer. The second layer of the architecture, called Networking & Transport layer, corresponds to the network and transport layer in the OSI layer architecture. This layer is responsible for ITS Transport, TCP/UDP, IPv6 mobility extension, and other network protocols. The third layer of the architecture, called Facilities layer, provides upper layers with application/information/session support.

Additionally, in order to provide security and interact with the layers of the ITS-S reference architecture, the security entity and management entity are resided outside the ITS-S protocol stack and connected with each layer of the ITS-S reference architecture, respectively. Finally, the application layer is operated over all these layers.

3 Proposed MI ITS M2M Model

In this section, we introduce our proposed Multiple Interface (MI) management framework for ITS M2M devices based on the CALM architecture. As shown in Fig. 2, it consists of three parties: (a) An ITS M2M device (OBU) which is equipped with multiple interfaces and contains several M2M applications (b) An ITS access network (RSU) which provides network connectivity to the ITS M2M device and its parameters of the access network, and (c) An ITS M2M server which provides the MI management policy to the ITS M2M device and may make a decision to select an appropriate network for ITS M2M device.

Fig. 2 Conceptual model of our MI ITS M2M system



Different from a human operating system, which is controlled by a human operator considering overall information of networks and requirements of applications, the ITS M2M device is required to decide the appropriate network for the each application by itself according to the predefined policy and network status. For example, an ITS M2M application to report vehicular accidents should transmit its message to the ITS M2M server despite the high cost of network usage. On the other hand, the transmission of non-time critical data of huge size can be postponed until the ITS M2M device succeeds in establishing a network connection with low cost or free of charge.

4 M2M MI Management Framework

Figure 3 shows our proposed the structure of the MI ITS M2M framework based on the CALM architecture. In MI ITS M2M systems, several ITS M2M applications which have different communication requirements are assumed to run on an ITS M2M device.

In order to manage multiple interfaces efficiently for several M2M applications on an ITS device, we define four agents into Management Information Base (MIB) of the CALM architecture as follows. As shown in Fig. 3, the agents are interconnected via several Service Access Points (SAPs) such as MA-SAP, MF-SAP, MN-SAP and MI-SAP. According to the reference model of CALM Architecture, we define SAP primitives which are responsible for sending and receiving commands as well as reading and setting parameters.

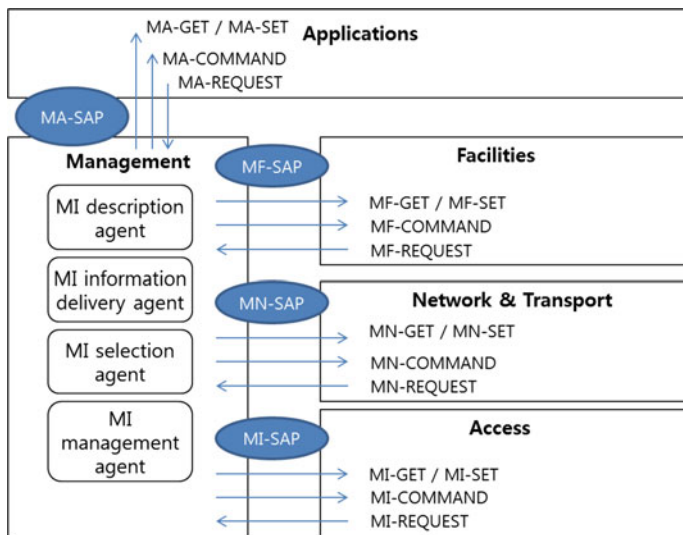


Fig. 3 MI ITS M2M architecture

- MI Description Agent:
 - Exchanges the network interface information Profile (data rate, cost, bandwidth, error rate, jitter, and etc.) and the network selection policy (Preference of networks, cost metrics, and etc.) with ITS M2M server and RSU.
- MI Information Delivery Agent:
 - Delivers the network interface information profile and network selection policy through a transportation scheme such as http and CoAP.
- MI Selection Agent:
 - Selects an interface for an M2M application according to the policy from the M2M server.
- MI Management Agent:
 - Monitors the network status and manages the network profile of each network.
 - Triggers re-configuration for the interface selection when the previous selected interface violates the policy.

Figure 4 gives an information flows between the interface selection agent and the horizontal layers. Multiple Interface Selection agent maintains three information: MI selection policy, Interface information table and Network information table. MI selection policy contains the policy information required for network and interface selection such as preference of networks and cost metrics which is delivered from ITS M2M server and RSU. Interface information table and Network information table maintain the parameters of air interfaces and networks such as data rate, cost, bandwidth, error/loss rate and jitter. These information is obtained from MA-SAP, MF-SAP, MN-SAP and MI-SAP through Mx-REQUEST command. The decision of interface selection is sent to the Network & Transport layer via MN-SPA using MN-Command.

4.1 Data Exchanges and Usage

In order to provide interoperability for exchanging the network interface information profile and network selection policy among M2M devices and servers from different manufactures, the system should use a description language in a form of XML schema, and can describe policies and functions of selecting a network interface. Figure 5 shows an example of the network interface information profile.

In order to select a network interface for a specific M2M application by itself, each of ITS M2M devices is required to exchange data as follows.

- ITS M2M Server → ITS M2M device: policies for interface selection, authentication and billing information of each network, and etc.
- ITS M2M device → ITS M2M Server: status of network connectivity, available networks, and etc.

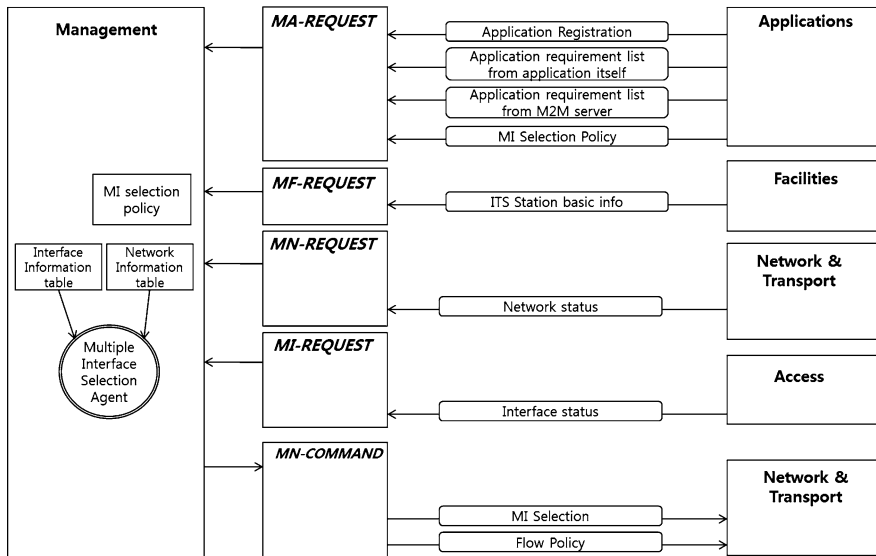


Fig. 4 Information flow between interface selection agent and the horizontal layers

```

<?xml version="1.0" encoding="UTF-8"?>
...
<miitsm2m:profile>
    <e2edelay unit = "ms"> 100 </e2edelay>
    <capacity unit = "mbps" req=true> NULL </capacity>
</miitsm2m:profile>
    
```

Fig. 5 Example of the network interface information profile

- ITS M2M RSU → ITS M2M device: network channel, bandwidth, jitter, QoS, and etc.
- ITS M2M device → ITS M2M RSU: authentication for the connection establishment, and etc.

5 Conclusion

In this paper, we proposed a multiple network interface management framework for ITS M2M devices in order to meet different network requirements and constraints for each of their ITS M2M applications. Our proposed framework consisting of four agents keeps monitoring the status of network interfaces and selects the best suitable interface for each M2M application according to the network selection policy without any human operation.

The framework is designed based on the CALM architecture and uses a description language in a form of XML schema for exchanging the network interface information and network selection policies. Through this framework, the interoperability and flexibility for ITS M2M systems can be provided, regardless of manufacturers and network providers.

Acknowledgments This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the CITRC (Convergence Information Technology Research Center) support program (NIPA-2012-H0401-12-1006) supervised by the NIPA (National IT Industry Promotion Agency). This research was also supported by the ICT Standardization program of KCC (Korea Communications Commission). This research was also supported by Kyungpook National University Research Fund, 2012.

References

1. Wu G, Talwar S, Johnsson K, Himayat N, Johnson KD (2011) M2M: from mobile to embedded internet. *IEEE Commun Mag* 49(4):36–43
2. Chen B, Cheng HH (2010) A review of the applications of agent technology in traffic and transportation systems. *IEEE Trans Intell Transp Syst* 11(2):485–497
3. ETSI (2010) Machine to machine communications (M2M): use cases of automotive applications in M2M capable networks, ETSI TR 102 898 V0.4.0, Sept 2010
4. ISO/TC TC204/SC WG16 (2012) Intelligent transport systems—communications access for land mobiles (CALM)—architecture, ISO/WD 21217, March 2012
5. Williams B (2006) CALM handbook v3, CALM forum, March 2006
6. IEEE P1609.1 SWG et al (2009) IEEE P1609.1 trial-use standard for wireless access in vehicular environments (WAVE) resource manager, IEEE P1609.1 D0.6, June 2009
7. IEEE P1609.2 SWG et al (2009) IEEE P1609.2 trial-use standard for wireless access in vehicular environments—security services for applications and management messages, IEEE P1609.2 D07, June 2009
8. IEEE P1609.3 SWG et al (2009) IEEE P1609.3 wireless access in vehicular environments (WAVE) networking services, IEEE P1609.3 D1.2, June 2009
9. IEEE P1609.4 SWG et al (2009) IEEE P1609.4 trial-use standard for wireless access in vehicular environments (WAVE)—multi-channel operation, IEEE P1609.4 D12, June 2009

A Framework of the Wireless Sensor Based Railway Signal System

Tarun Kumar, Ajay Chaudhary, Ganesh Singh
and Richa Sharma

Abstract In present era of railways transportation the presence of the railway signal system in railway networks makes the train control system much easier & safer. Today the all railway signal systems are based on electrical devices. In this paper we propose a new signal system for railway network which will be based on wireless sensor devices. Proposed frame work helps the train control mechanism to control the train even in low visibility weather condition like dust, haze, snow, heavy rain, fogginess conditions. In this paper we proposed the idea and the framework for the new wireless sensor device based signal system which can be implemented in modification in existing current signal systems.

Keywords Algorithms · Management · Performance · Design · Reliability · Experimentation · Security · Human factors · Zero-visibility · Wireless sensor device · Anti-collision device · Vigilance control device

T. Kumar (✉) · R. Sharma
Department of Computer Science, Government Engineering College, Bikaner,
Karni Industrial Area, Pugal Raod, Bikaner, Rajasthan, India
e-mail: ertarunkumar@yahoo.co.in

R. Sharma
e-mail: sharma.richa676@gmail.com

A. Chaudhary (✉) · G. Singh
Department of Information Technology, Government Engineering College,
Bikaner, Karni Industrial Area, Pugal Raod, Bikaner, Rajasthan, India
e-mail: ajaychaudhary_in@rediffmail.com

G. Singh
e-mail: ganesh.badhu@gmail.com

1 Introduction

The train control system is brought and developed to ensure traffic safety, improve transport efficiency and working conditions. Railway signal is the most important factor in the train control system. It carries instructions and information to control the train line, time interval and speed, displays current train line and equipment status, so as to ensure the safety of high-speed train traffic [1].

In European and other developed countries, the research and applications of train control system starts early. Especially in Europe, there are seven big railway signal companies, like Alcatel Company and Alstom Company in France, Adtranz Company of Sweden, Ansaldo Company in Italy, Siemens Company in Germany and WestingHouse Company in the UK [2]. Different kinds of Train Control Systems are designed by different company in different country. For example, LZB Train Control System is Used in Germany, while TVM Train Control System is Used in France.

East Japan Railway Company has been developing a new signal control system based on an IP-network to solve those problems [3]. This system composes a network by using optical cables and controls signal devices by an exchange of digital information. This can reduce cable construction work, and increase efficiency of system test and operation, and also save construction cost [4].

2 Current Signal System

2.1 Problem in Currant Signal System

The signal control systems are dramatically developed and advances, mainly because of the current computer technologies. Nevertheless, they are some problems about signal control and route control.

2.2 Problem Due to Low Visibility

In the current signal system the signal is identified by the train driver by looking at the signal pole and driver analyze that signal according to the color of the signal (RED/YELLOW/GREEN) and take appropriate action according to the signal but in case of high raining and fog there is problem of low visibility of the signal to the driver and the problem of identification of signal arise which cause the train driving difficult. In some cases even it is really hard to indentify signals correctly hence leads to sudden accidents. As a result approximately 177 accidents that occurred in 2008–2010, even more than half of accidents occurs due to climate conditions like fog and heavy rain etc. [4].

3 Scope of Proposed System

3.1 Solution to the Problem of Low Visibility

As we have seen earlier that the train driving in low visibility is difficult due to electrical signal system but our proposed signal system framework uses the wireless sensor (WSN) devices. WSN devices used to send the signal on the railway track to the engine and the problem of low visibility or heavy rain of the signal has no effect on this system.

3.2 Scope of Automatic Driving Mode in Train

The wireless sensor device for signal system in railway may allow to send the signal to the wireless sensing device present in the engine and it will analyze the signal according to the message it received by the device and by using the microcontroller in the train engine, we can control the speed of the train according to the demand of the signal so this concept can add the option of auto driving mode in train engine.

4 Concept and Framework of the System

4.1 Definitions of WSN Device

For implementation of the new railway signal system based on wireless sensor devices we use mainly the three types wireless sensor devices these device are called nodes.

1. Wireless sensor device at the pole of signal system which will transmit the current signal status, in this paper we called this device signal sender.
2. Second wireless sensor device we will use in the engine of the train which will receive the signal when it will in the range of any signal and also it will analyze the signal status. In this paper we called this device signal receiver.
3. Third sensor device we will put on the track where the two track are meets are locking of tracks which will used to update the current track number in the device used in the engine and in this paper we called this device track changer.

4.2 Working of Proposed Framework System

Now our system will work on the modification of the current railway signal system. It will implemented by placing the signal sender device on the signal pole of the current system. This device will sense the light color of the signal and

transmits the message with the trackID, time and the signal status of the pole. In this system we will use the trackID system due to possibility of the multiple parallel tracks and the system will define the unique trackID for each track. It will set to the signal sender device of the every track. The signal sender device now transmits the signal in required range of approximate 2 KM.

Now the signal receiver device in the engine comes in the range of signal sender device it will receive the message which contains the trackID and the status of the signal after that it match the trackID with its own trackID to ensure that the signal is for this train only after the matching the trackID it reads the signal status and activate the corresponding action for the signal.

The third device which is track changer is placed in the interlocking of the tracks. This device is used as our system maintain the track position of the engine, because the signal are based on the trackIDs. Whenever the track of the train is changed its current track position should be updated that's why the track changer device is placed at the interlocking of the tracks. Whenever train change its track changer device update the current track position of the engine so that the engine can receive the signal which is only for its current track.

The implementation of this system is shown in Figs. 1 and 2.

4.3 Working of the Framework for Proposed System

In this section we will describe the proposed logic at the all three devices signal sender, signal receiver and the track changer with the message structure.

The signal sender will send the message which contains the following information.

TrackID-for the track id for which signal is working.

Signal_Status-is the current status of the signal

Time- is the time at which signal is detected by device.

the class structure for message at the signal sender device will be.

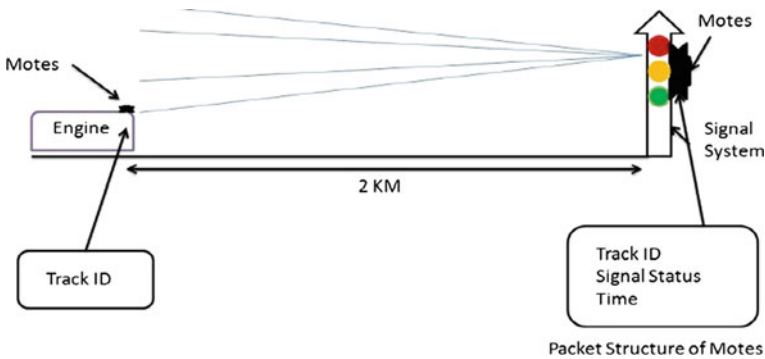


Fig. 1 Implementation of motes (signal sender and signal receiver) in current signal system

```
Class signal_sensor  
{  
  Int TrackID;  
  Int Signal_Status;  
  Time T;  
}  
void set_signal_status(int st)  
{  
  T=getTime();  
  Signal_Status=st;  
}  
}  
  
Int main()  
{  
  Int st=RED;  
  Signal_sensor S1;  
  While(1)  
  {  
    St=read(Signal from Motes);  
    S1.set_signal_status(st);  
    Send(S1);  
  }  
  Return 0;  
}
```

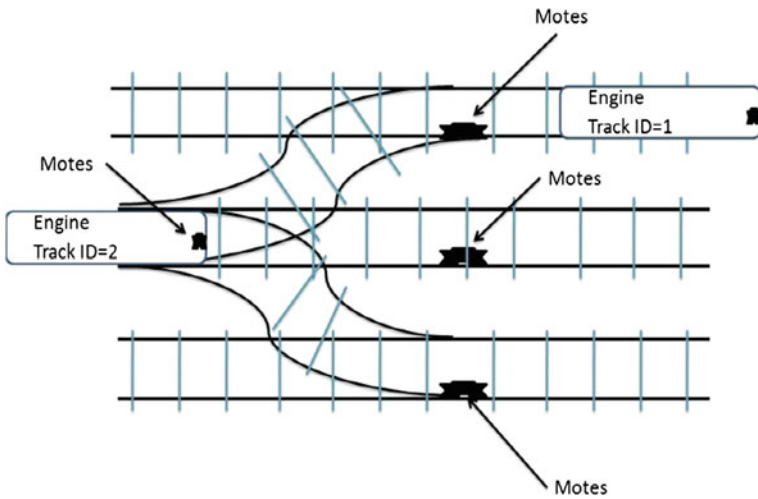


Fig. 2 Implementation of motes (track changer) at the interlocking of current signal system

The signal receiver contains the following data structure at the engine.

Current_TrackID-it will store the track id on which train is running.

Time-it will contains the time at which signal is received.

the class structure will be as follows.

Class signal_receiver

```

{
Int Current_TrackID;
Time T;
Int set_Current_Track(int tid)
{
Current_TrackID= tid;
T= getTime();
return 0;
}
}
Int main()
{
signal_receiver E;
signal_sensor S;
while(1)
{
S= read(message from motes);
If(S.TrackID==E.Current_TrackID)
{
If(S.Signal_Status==RED)
{
//activate RED LED in the engine with Alarm
}
else if(S.Signal_Status==YELLOW)
{
//activate YELLOW LED in the engine
}
else if(S.Signal_Status==GREEN)
{
//activate GREEN LED in the engine
}
}
}
}
}

```

The third device track changer will called the set_Current_Track() method of the engine as the engine comes to the interlocking and the motes on which the

engine running will activate from the load of the engine and it will change the current trackID of the engine.

5 Conclusion

After modification of the current railway signal system with the our proposed railway signal system we can overcome from the problem of the signal tracking in low visibility condition as well as this new system can introduce the concept of the auto driving mode in the engine so that for the train driver can relax also it is completely depend on the wireless sensing device so this this system will free from the human readable error in the signal tracking also it can be implemented with less effort and at very low coast by the modification of the current railway signal system.

6 Future Work

In this paper we proposed the only framework and the concept of the new railway signal system which will based on wireless sensor devices but in future we will implement this feature and will design the model for our this signal system and we will implement the concept of automatic train operation .

References

1. Kong Y, Do P-X, Tan Z-H (2009) The simulation system of railway signal system transmission and on-board cab signal receiving. In: Proceedings of the IEEE 70th vehicular technology conference fall (VTC 2009-Fall)
2. Xiao DN, Liu HY (2008) Foreign train control system in high-speed railway, railway standard design
3. Nishiyama J, Sughara H, Okada T (2007) A signal control system by optical LAN and design simplification. In: Proceedings of the ISIC. IEEE international conference on systems, man and cybernetics
4. Mathur V, Singh K, Chawhan MD (2012) Zero-visibility navigation for the indian railways. In: Proceedings of the MPGI national multi conference 2012 (MPGINMC-2012)

Evolutionary Bio-Interaction Knowledge Accumulation for Smart Healthcare

Sung-Kwan Kang, Jong-Hun Kim, Kyung-Yong Chung,
Joong-Kyung Ryu, Kee-Wook Rim and Jung-Hyun Lee

Abstract The range of ubiquitous computing technology available for use in healthcare continues to evolve, allowing for an increasing variety of wireless sensors, devices, and actuators to be deployed in changing environments. This paper presents a robust distributed architecture for adaptive and intelligent bio-interaction systems, called Evolutionary Bio-inspired Knowledge Accumulation. This system is designed to its capability to increase knowledge enhancement even

S.-K. Kang

HCI Lab Department of Computer Science and Engineering, Inha University,
Yong-Hyun Dong, Incheon, South Korea
e-mail: kskk1111@empas.com

J.-H. Kim (✉)

U-Healthcare Department, Bit Computer, 1327-33 Bitville, Seocho-dong,
Seocho-gu, Seoul, South Korea
e-mail: kimjh@bit.kr

K.-Y. Chung

School of Computer Information Engineering, Sangji University,
83 Sangjidae-gil, Wonju-si, Gangwon-do, Korea
e-mail: dragonhci@hanmail.net

J.-K. Ryu

Department of Computer Science, Daelim University College,
Anyang, Korea
e-mail: jkryu@daelim.ac.kr

K.-W. Rim

Department of Computer Science and Engineering, Sunmoon University,
Galsan-ri, Tangjeong-myeon, Asan-si, Chungcheongnam-do, Korea
e-mail: rim@sunmoon.ac.kr

J.-H. Lee

Department of Computer Science and Engineering, Inha University,
Yong-Hyun Dong, Incheon, South Korea
e-mail: jhlee@inha.ac.kr

in dynamic and uneven environments. Our proposed system adopts the concepts of biological context-awareness with evolutionary computations where the working environments are modeled and identified as bio-environmental contexts. We have used an unsupervised learning algorithm for bio-context modeling, and a supervised learning algorithm for context identification. A genetic algorithm, for its adaptive criteria, is used to explore action configuration for each identified bio context to implement our concept. This framework has been used to reduce noise in ECG signals that have been gathered in routine remote healthcare monitoring. Experimental results showed that the proposed algorithm effectively removes baseline wander noise and muscle noise, and feature extraction results showed a significant improvement of T duration extraction values.

Keywords Ubiquitous healthcare systems · Context awareness · Interactive healthcare

1 Introduction

Computer scientists are showing more interest in the research area of adaptation under dynamically changing environments. As a result, many studies are carried out and many algorithms and techniques are developed. Recently, adaptation capability under dynamically changing environments has become more important, since advanced applications need service-oriented, pervasive, and ubiquitous computing [1, 2]. In this paper, we discuss the framework of a self-growing system that can behave in an adaptive and robust manner under dynamic variations of application environments. The knowledge of individual environmental contexts and their associated chromosomes is stored in the context knowledge base. The most effective action configuration of the system is determined for the current environment by employing the popular evolutionary computing method, e.g., a genetic algorithm (GA). Evolutionary computing is an efficient search and adaptation method that simulates the natural evolutionary mechanism [3]. GA guides the system adaptive to varying environments. Adaptation to changing environments is an important application of GAs [4, 5].

In this paper, a new method is proposed for real-time adaptation called Evolutionary Bio-inspired Knowledge Accumulation (EBKA). EBKA can adapt to changing situations by identifying its environment in real-time if the environments recur. The main difference of the proposed method from other evolutionary computing methods is that it can optimize action configurations in accordance with an identified context as well as store its knowledge in a knowledge context. Hence, the proposed method can add self-growing and real-time adaptation capabilities to a system. That is, once the context knowledge is constructed, the system can react to changing environments in real-time.

We achieved encouraging experimental results showing that the performance of the proposed algorithm-based filter is superior to those of conventional standard filters in uneven environments.

2 Evolutionary Bio-Inspired Knowledge Accumulation

In this session, we discuss the model of EBKA with the capabilities of real-time adaptation and context knowledge accumulation.

2.1 Goal of Evolutionary Bio-Inspired Knowledge Accumulation

The goal of EBKA can be described as the provision of optimal services using given system resources by identifying its working environments. The major functionalities of EBKA can be formalized as follows (see Fig. 1):

- Identify working environment of the system,
- Configure its structure using an autonomous computing method, and
- Accumulate knowledge from its experience, and therefore grow itself. (Autonomous knowledge accumulation).

Two types of data inputs, context data and action data, are used as inputs in EBKA. The action data, denoted by x , is normal data being processed. The context data, denoted by y , is used to identify an environmental context of EBKA and to

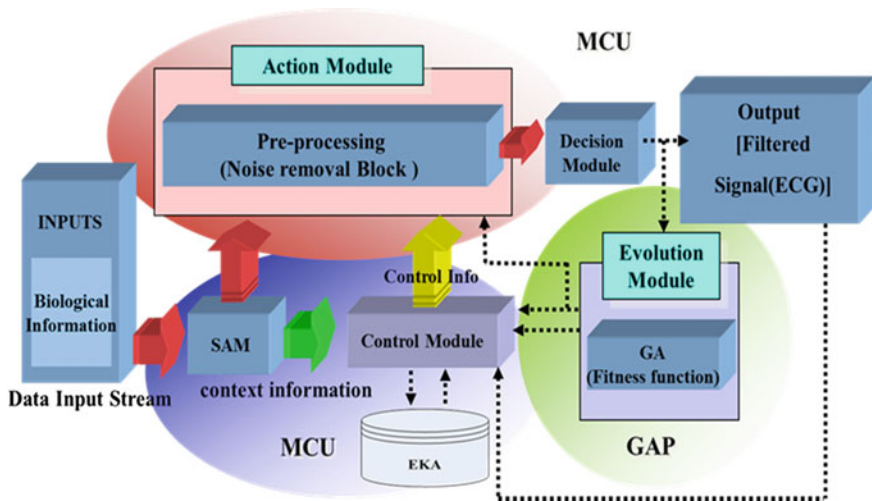


Fig. 1 Major functionality of EBKA

construct a proper action based on the identified context. In many cases, the action data itself can be used as the context data. We assume that EBKA context data can be modeled in association with the input action data [6].

2.2 Proposed EBKA

In this section, we will discuss the EBKA framework. There are two important design issues. The first is how to learn and identify a context category in association with application working environments. And the second is how to decide on an action configuration for an identified context from all possible combinations of action configurations (shown in Fig. 2). The most effective action configurations for an identified context are combined to produce a final output of the scheme using some aggregation method [7]. We need to devise a method of calculating a context profile in order to derive a context action profile. We assume that environmental data can be modeled as being clustered into several discredited environmental contexts in association with distinguishable application actions. The proposed scheme operates in two modes, Learning Mode and Action Mode, using context awareness. The knowledge of the most effective subset of action configurations for identified contexts is accumulated and stored in the context knowledge base (CKB) with associated artificial chromosomes in Learning Mode [8, 15]. The knowledge is used to select a most effective subset of classifiers for an identified environmental context, and the produced effective action configurations are aggregated to produce an optimal output of the scheme in Action Mode. After categorizing the environmental context, it is further reconfigured for a ubiquitous or mobile environment to produce and/or select the ubiquitous service.

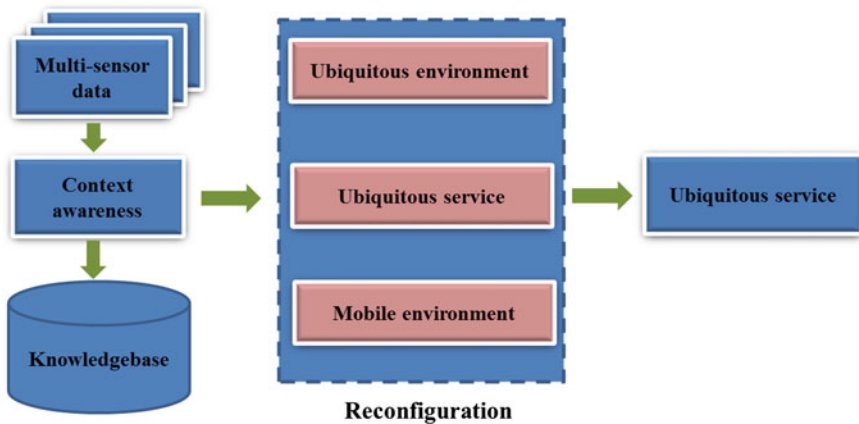


Fig. 2 Dataflow diagram of EBKA

2.2.1 Autonomous Knowledge Accumulation

Action configuration is carried out using action primitive set and accumulated knowledge in the CKB. Action configuration can be described using the artificial chromosome of GA. The configured action produces a response or output of the scheme [9]. Initially, the scheme learns application's environmental contexts. It accumulates the knowledge of context-action association, and stores them in the CKB. The knowledge of context-action association denotes that of most effective action configurations for the context. The detail of context knowledge accumulation process is given in the followings.

Algorithm. (Evolution mode)

Input: An input action data and associated context data.

Output: The result of CKB evolution

Step 1. Train the CAM using the input context data.

Step 2. Start to search for an optimal action configuration for each environmental context category until a predefined criterion is met, where the criterion is the fitness does not improve anymore or the predefined maximum trial limitation is encountered as follows.

- 2.1) Generate initial chromosome population of action configuration.
- 2.2) Evaluate the fitness function of the scheme using the newly derived population of the action configurations. If the criterion is met, go to Step 3.
- 2.3) Search for the population of the action configuration that maximize the fitness function and keep those as the best chromosomes.
- 2.4) Applying GA's genetic operators to generate new population from the current action configuration. Go to Step 2.2.

Step 3. Update the CKB for the identified environmental category and the derived classifier structure.

2.3 Ubiquitous Computing Based on Interactive Healthcare

An IHC (Interactive Healthcare) integration system that is based on bio-interaction was born from the combination of a traditional bio-interaction model and healthcare technology [10]. Now it is establishing a new ideal of healthcare, because it models the molecular and cell levels of system biology, using sensors that accommodate changes over time based on their environmental context.

The sensor is designed to accommodate changes in its state and environment and also changes in the biological systems that it is monitoring via its evolutionary adaptability. We decided the basis for the design of the IHC system by analyzing measurement data's special qualities, and when designing we paid attention to what different area do to the data through evolution adaptation in practice [11].

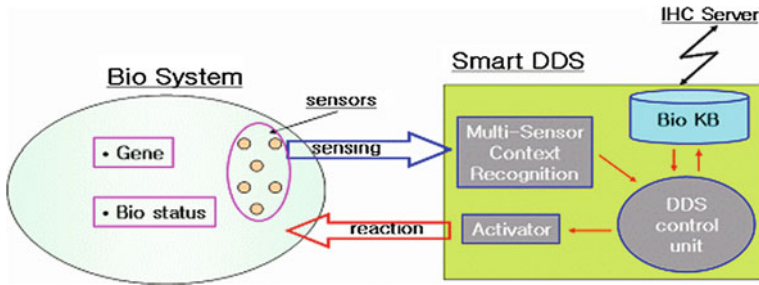


Fig. 3 U-IHC diagram based on real-time distributed bio-interaction

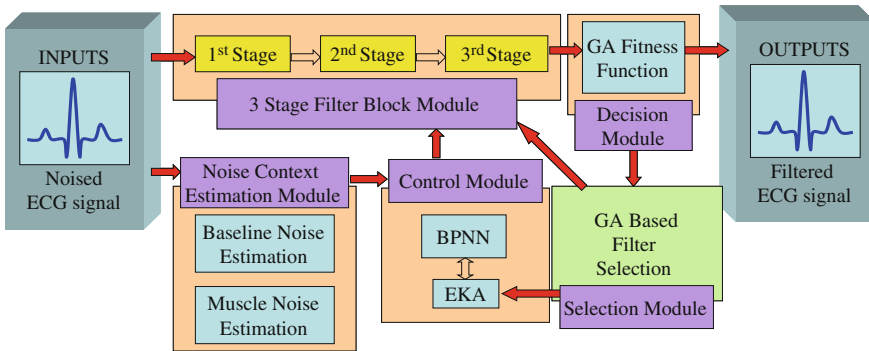


Fig. 4 Conceptual diagram of proposed adaptable noise reduction algorithm

Figure 3 shows IHC concept based on real time Bio-interaction. When it is caused a change of state from the Biology system(user) who receives the services of IHC system by disease or other causes, Smart DDS (Drug Delivery System) of IHC system associated with these incidents take appropriate action after sensing the user’s state. These reactions alter the state of Biology system once again. And, Smart DDS system undergoes the reaction steps based on the information that get feedback from Biology system. This process can maximize the effect of the medication to their patients. Prerequisites to maximize for these effects the reliability for the analysis on the state must be adequately secured. The development of sensing and actuation devices to satisfy this, it can be succeeded only by the close convergence of the existing measurement methods by hand and intelligent evolutionary adaptation technology [12, 14].

We proposed a noise reduction method that uses context estimations, such as that illustrated in Fig. 4, based on the above system. The proposed noise reduction algorithm mainly consists of five modules, including a noise context estimation module, a control module for filter block design at a running mode, a 3-stage filter block module, a selection module for filter design in evolution mode, and a decision module for fitness calculation in evolution mode as shown in Figure 4.

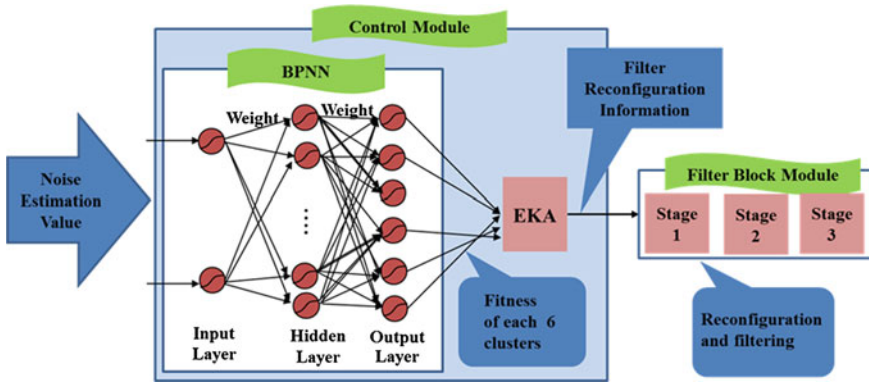


Fig. 5 Control module of neural network base and filter block module

We use an Error Back-Propagation Neural Network for the Control module, and the noise context estimation module’s output as the control module’s input. Then, we use a pre-determined 6 filter combination that is stored in the Evolutionary Knowledge Accumulator (EKA) as output [13]. Fitness passes filter combination information to the 3-step filter block module to reconstruct the greatest combination [16]. Then, the 3-step filter block module composes a suitable filter for the combination information and removes the noise of the electrocardiogram input (Fig. 5).

After removing the noise, the noise removal fitness is calculated by the decision module’s fitness function. If it is satisfactory, it displays the altered electrocardiogram signals and finishes the algorithm achievement. Otherwise, after looking for an optimal filter combination via the filter combination selection module of the genetic algorithm base, it passes the combination information to the 3-step filter block module tries to remove the noise again.

3 Conclusions

This paper proposes a distributed self-growing framework that can be used for adaptive systems under dynamic environments taking advantage of context-awareness, self-growing architecture, and context knowledge base. The EBKA framework separates the Learning Mode from the Action Mode to solve the time-consuming problem that is the intrinsic weak points of the GA. The evolutionary mode accumulates the context knowledge from a varying ubiquitous environment. The Action Mode executes the task of identification using the chromosome knowledge accumulated in the evolutionary mode. The main difference of EBKA from other popular adaptive systems is that it can optimize an action configuration in accordance with an identified context, and store its knowledge in the context knowledge. Hence, EBKA can provide self-growing and real-time adaptation

capability to the system. Once the context knowledge is constructed, EBKA can react to changing environments in real time. We show the feasibility of the EBKA framework in the area of object recognition where most popular approaches show vulnerability under a dynamically-changing environment.

Acknowledgments This work was supported by the R&D Program of MKE/KEIT.

References

1. Ong KG, Dreschel WR, Grimes CA (2003) Detection of human respiration using square-wave modulated electromagnetic impulses. *Microw Opt Technol Lett* 35:339–343, 5 Mar 2003
2. Slay H, Thomas B, Vernik R et al (2004) A rapidly adaptive collaborative ubiquitous computing environment to allow passive detection of marked objects. *Lecture notes in computer science*, pp 420–430
3. Gomez A, Fernandez M, Corch O (2004) *Ontological engineering*, 2nd edn. Springer-Verlag, Berlin Heidelberg New York
4. Goldberg D (1989) *Genetic algorithm in search, optimization, and machine learning*. Addison-Wesley, Reading
5. Mori N et al (2000) Adaptation to a dynamic environment by means of the environment identifying genetic algorithm. *Ind Electr Soc IECON 2000 26th Ann Conf IEEE* 4:2953–2958
6. Liu C, Wechsler H (2000) Evolutionary pursuit and its application to face recognition. *IEEE Trans Pattern Anal Mach Intell* 22(6):570–582
7. Abowd GD (1999) Classroom 2000: an experiment with the instrumentation of a living educational environment. *IBM Syst J* 38(4):508–530
8. Celentano A, Gaggi O (2006) Context-aware design of adaptable multimodal documents. *Multimedia Tools Appl* 29:7–28
9. Gonzalez RC, Woods RE (1993) *Digital image processing*. Addison Wesley, Reading, pp 161–218
10. Kuncheva LI, Jain LC (2000) Designing classifier fusion systems by genetic algorithms. *IEEE Trans Evol Comput* 4(4):327–336
11. Moghaddam B, Nastar C, Pentland A (1996) A Bayesian similarity measure for direct image matching. *Proceeding of international conference on pattern recognition*
12. Pancer TP (2004) A suppression of an impulsive noise in ECG signal processing. *Proceeding 26th annual international conference IEEE EMBS*, pp 596–599
13. Turk M, Pentland A (1991) Eigenfaces for recognition. *J Cong Neurosci* 13(1):71–86
14. <http://www.physionet.org/physiobank/database/#ecg>
15. Yau SS, Wang Y, Huang D, In H (2003) A middleware situation-aware contract specification language for ubiquitous computing. *Proceeding of 9th international workshop on future trends of distributed computing systems (FTDCS2003)*, Puerto Rico, pp 93–99
16. Yau S, Wang Y, Karim F (2002) Developing situation-awareness in middleware for ubicomp environments. *Proceeding 26th international computer software and applications conference (COMPSAC 2002)*, pp 233–238

A Non-Volatile Buffered Main Memory Using Phase-Change RAM

Do-Heon Lee, Chung-Pyo Hong and Shin-Dug Kim

Abstract The new trends of memory semi-conductor technology are changing and developing. Phase-Change RAM (PRAM), Ferroelectric RAM (FeRAM), Magnetic RAM (MRAM) and Resistive RAM (RRAM) are going to take center stage of main memory material of new computer systems in next decade. PRAM also has higher dense, it can keep data about four times more than DRAM. But some problems caused when PRAM uses as a main memory directly. So we suggest Pre-load cache and Assistant buffer. It reduces main memory access and overcome low read speed of PRAM consequently. To reduce write operation also, we propose Assistant buffer. Assistant buffer keeps evicted data and impedes write operation, and facilitates more rapid response about required data when cache misses. As a result of our experimentation, overall performance is decrement of main memory accesses approximately 50 %.

Keywords Phase-change memory · Main memory · Cache · Buffer · DRAM · Non-volatility · Memory capacity · Memory access latency

D.-H. Lee (✉) · C.-P. Hong · S.-D. Kim
Department of Engineering, Yonsei University, 5-4, Sinchon-dong, Seodaemun-gu,
Seoul, South Korea
e-mail: intovortex@yonsei.ac.kr

C.-P. Hong
e-mail: hulkboy@yonsei.ac.kr

S.-D. Kim
e-mail: sdkim@yonsei.ac.kr

1 Introduction

DRAM is main memory material of modern computer system. Requirement of applications and operating systems are increasing. They need larger size and lower latency of main memory. To satisfy their demands, a main memory is increasing its capacity. But it causes more energy consumption and heat problem. To cool off the system, air and temperature conditioning system's energy consumption is increasing. These problems are huddle of computer system growth [1]. Memory semi-conductor technologies are developing and growing to resolve. Phase-Change RAM (PRAM), Ferroelectric RAM (FeRAM), Magnetic RAM (MRAM) and Resistive RAM (RRAM) are going to take center stage of main memory material of new computer systems in next decade. These non-volatile memories have some advantage. First, the memories consume lower power than DRAM. Mobile device like smart-phone is used widely and it involves DRAM main memory also. But mobile devices should not consumes much energy, so it is to be considered low power. Second, density of the memories is higher than DRAM. The memories can save same quantity of data use of fewer chips. It effects power consumption. Third, their major characteristic is non-volatility, which means it doesn't need to be saved to disk when system power is turned off. It can reduce recovery after boot-up time significantly.

But the memories have several issues to apply to use as a main memory. PRAM and any other RAMs are lower performance compared to DRAM. Read latency of PRAM is about 200–300 ns. Write speed is slower than DRAM also. If computer system uses PRAM as a main memory directly, it causes serious performance decreasing. Another handicap of non-volatile memories is write limitations. The memories have limited write number, which has the possibility of cell worn-out. Write operation should be carefully managed.

To solve these problems, this research proposes a new cache and buffer hierarchy. This new hierarchy helps PRAM main memory as an assistant. One of hierarchy components is Pre-load cache. Pre-load cache pre-fetches from main memory. Another component is assistant buffer. When last level cache of processors evicts, it impedes writing on PRAM main memory. The proposed structure is implemented and evaluated by using a trace-based simulator with SPEC CPU 2006 traces and Splash-2 traces [2]. Our experiment shows the number of memory access is halved and decrement of write access. It is directly connected with performance improvement. The less write operation occurs, the less energy is consumed, and we gain the profit from that. Finally, we can store more data by using PRAM main memory and consume less power.

In Sect. 2 introduces background of our research and characteristics of PRAM, structure motivation. Section 3 shows new computer systems that added our suggested structure. Next, Sect. 4 explains our experimental methodology and shows results, analysis. Section 5 is our conclusion.

Table 1 Comparison of measured roughness data, machining center

Parameter	DRAM	NAND flash	NOR flash	PRAM
Density	1X	4X	0.25X	2X-4X
Read latency	60 ns	25 us	300 ns	200-300 ns
Write speed	-1 Gbps	2.4 MB/s	0.5 MB/s	-100 MB/s
Endurance	N/A	10 ⁴	10 ⁴	10 ⁶ -10 ⁸
Retention	Refresh	10 years	10 years	10 years

2 Background and Motivation

With increasing number of processors in the computer system, the pressure on the memory system to satisfy the demand of all concurrently executing applications (threads or processes) has increased as well. PRAM is a one of non-volatile memories that exploits the property of chalcogenide glass to switch between two states, amorphous and crystalline, with the application of heat using electrical pulses. Additionally, these chemical and physical characteristics PRAM shows better performance than flash memories (NAND flash and NOR flash) [3]. Table 1 summarizes the properties of different memory technologies based on the data obtained from the literature. Write endurance is the maximum number of writes for each cell. Data retention is the duration for which the non-volatile technologies can retain data.

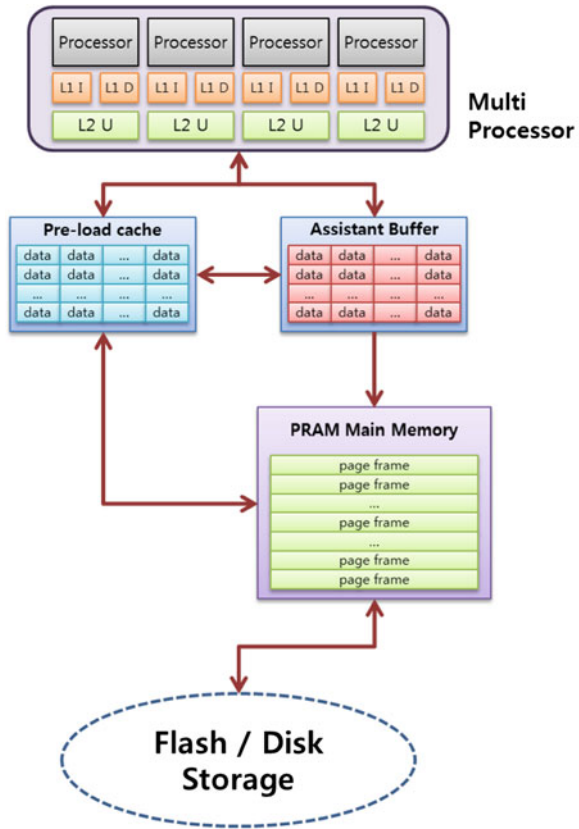
From Table 1, flash memories show worse performance than PRAM which cannot apply as a main memory. Only PRAM has similar performance with DRAM.

Qureshi et al. proposed a DRAM buffer as a hardware cache structure for PRAM main memory [4]. It improves asymmetrical read/write access latencies of PRAM, compared to DRAM. A large amount of DRAM buffer space is used to hide read latency and a write queue is used to hide write latency. Our proposed architecture based on Qureshi's, we changed DRAM buffer to fetch cache and write queue is improved [5]. Prefetching is well known technology, which can reduce cache misses. To minimize initialization cache miss, we set prefetch cache up to off-chip cache. Write queue is just delay writing, assistant buffer can see as a last level cache and can fetch from it. We used simple principles what partial locality and temporal locality [3]. It will be discuss detail next section.

3 New Main Memory Hierarchy System

Figure 1 shows overall architecture that we proposed. Basic architecture is similar to present model; we added Pre-load cache and Assistant Buffer. Main memory does not connect directly with processors anymore, when processors last level cache miss occurs, Pre-load cache or Assistant buffer services received request. Overall detail flow sequences are shown below.

Fig. 1 New hierarchy architecture with Pre-load cache and assistant buffer



First, when request comes to on-chip last level cache, if Pre-load cache has requested data, just serve it and wait for next request. Next, if Pre-load cache has not, request finds Assistant buffer. If Assistant buffer has it, serve requested data to processor and fetch it to Pre-load buffer. When Assistant buffer misses, finally request brings from PRAM main memory.

3.1 Pre-Load Cache

Pre-load cache operates as an off-chip cache. If on-chip last level cache evicts, Pre-load cache receives it. It can pre-fetch main memory data that can serves before memory access very quickly. It operates like small main memory. Applications tends to be concentrated main memory’s special region, accuracy of Pre-load cache is not low. If Pre-load cache eviction occurs and request is need to write back, Pre-load cache evicts request to Assistant Buffer, otherwise bypass the request.

3.2 Assistant Buffer

Assistant Buffer operates like write queue. It retards write operation and keep data from evicted Pre-load cache, not only write queue, but also off-chip last level cache. If Pre-load cache miss occurs and Assistant buffer has data, Assistant buffer can serve directly to processor, the data fetched to Pre-load cache. Lastly Assistant buffer impedes read/write access of PRAM main memory.

4 Experimental Setup

The simulator for the proposed model is developed to evaluate the miss rate and access number. We used Gem5 simulator, SPEC CPU 2006 and Splash-2 benchmarks to extract memory access trace information [6–8]. We implemented Pre-load cache and Assistant buffer, connected each other with processor caches. Proposed architecture has two first level cache, which instruction cache and data cache, which consist of 64 KB, no set associativity, with 256 Byte of block size and 256 lines. Second level cache is on-chip last level unified cache that composed 512 KB, 8 set associativity, with 256 Byte of block size and 256 lines. Pre-load cache is made up 1 MB, with 4 set associativity, 256 Byte of block size and 1,024 lines. Assistant buffer is composed 1 MB, with no set associativity, 256 Byte of block size and 4,096 lines.

Simulation chases each cache or buffer status and number of miss or access. Each benchmark programs has different number of operations, we normalized results.

5 Result and Analysis

5.1 Total Accesses

Figure 2 shows total access to main memory. Existing architecture is normalized to 100 %, proposed model is reduced access percentage approximately 50 %. Benchmark gcc shows lowest access to main memory and h264ref is highest access, because its data is huge and worse locality than any other benchmarks.

5.2 Write Accesses

Assistant buffer impedes writing data to main memory, Fig. 3 shows how Assistant buffer keeps well, finally it can reduce total number of write to main memory. Benchmark fft and lu shows significant decrement of number of writes.

Fig. 2 Total access to main memory

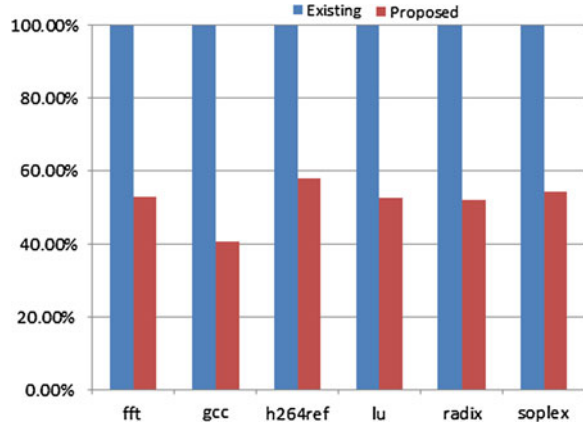
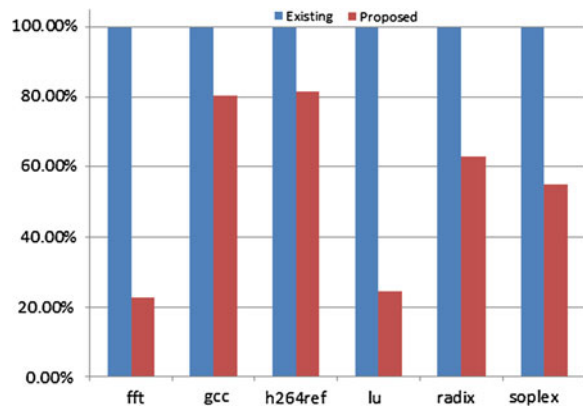


Fig. 3 Write access to main memory



Benchmark h264ref shows still worse locality. Number of writes depends on application’s characteristic.

From two experimental results, Pre-load cache and Assistant buffer can reduce considerable access to main memory. From this result, it can save more energy than DRAM uses. Write access chart shows number of write depends on application, but total access depends on application’s characteristic less than write operation.

6 Conclusion

To solve problems of DRAM, about replacement of main memory material, but new technology is not ready to use directly. So we proposed Pre-load cache and Assistant buffer, which can pre-fetch data and impede writing operation. Experiment shows the performance of cache and buffer is good, but still some problems

remain. First, PRAM has limited lifetime that we should use wear-leveling technique, it causes performance decrease. Second, when access to PRAM main memory, it give penalties a lot of time. To reduce original time to access to PRAM, the research have to continue. Finally, pre-fetch algorithm is incomplete, should develop effective pre-fetch algorithm and eviction algorithm. And should consider tradeoff between uses of pre-fetch mechanism or not.

References

1. Freitas RF, Wilcke WW (2008) Storage class memory: the next storage system technology. *IBM J Res Dev* 52(4.5):439–447
2. Colmenar JM, Risco Martin JL, Lanchares J (2011) An overview of computer architecture and system simulation. *SCS M&S Mag* 2(02):74–81
3. NcPRAM. http://www.samsung.com/global/business/semiconductor/products/fusionmemory/Products_NcPRAM.html
4. Qureshi MK, Srinivassan V, Rivers JA (2009) Scalable high performance main memory system using phase-change memory technology. In: *Proceedings of the 36th annual international symposium on computer architecture*
5. Jung KS, Park JW, Weems CC, Kim SD (2011) A superblock based memory adapter using decoupled dual buffers for hiding the access latency of nonvolatile memory, *Proceedings of the world congress on engineering and computer science*
6. Gem5 simulator system. <http://www.gem5.org/>
7. SPEC CPU 2006. <http://www.spec.org/cpu2006/>
8. The Modified SPLASH-2. <http://www.capsl.udel.edu/splash/index.html>

A Study on the Real-Time Location Tracking Systems Using Passive RFID

Min-Su Kim, Dong-Hwi Lee and Kui-Nam J Kim

Abstract The location awareness technology is a core technology to be expanded to include objects from human-oriented informatization, and active support actions have been performed in developed countries such as the United States and Japan to implement the location awareness technology and Real Time Location System (RTLS) tag and antenna technology for real-time location tracking through a variety of projects for years. However, problems have been posed by Global Positioning System (GPS) based on the location awareness technology and active bat system using sound waves in terms of the space and construction costs. In this regard, this study attempted to suggest passive RFIF-based Indoor Positioning System (IPS) for tracking the location of the moving objects (humans and assets) in real time.

Keywords Passive RFID · IPS · RTLS · GPS · Ultrasound · Infrared

M.-S. Kim (✉) · D.-H. Lee
Department of Industry Security, Kyonggi University, Chungjeongno 2-ga,
Seodaemun-gu, Seoul, South Korea
e-mail: fortcom@hanmail.net

D.-H. Lee
e-mail: dhclub@naver.com

K.-N. JKim
Department of Convergence Security Kyonggi, University, Iui-dong,
Yeongtong-gu, Suwon-si, Gyeonggi, South Korea
e-mail: harap123@daum.net

1 Introduction

Emerged as a new paradigm in computerization, ubiquitous technology is also referred to as fusion technology to provide additional services required by users and businesses through processing and combining the location, time and space information with other information after its storage and management. As technologies to support the ubiquitous era, there exist object recognition, location awareness and behavior analysis technologies. Among these technologies, the location awareness technology is aimed at tracking, monitoring and reporting the location of objects for ubiquitous computing as a core technology to implement invisible technologies recognized as characteristics of ubiquitous society. In this connection, many systems have been formulated for years focusing on automatic position awareness issues. The location awareness technology is a core technology to be expanded to include objects from human-oriented informatization, and active support actions have been performed in developed countries such as the United States and Japan to implement the location awareness technology and Real Time Location System (RTLS) tag and antenna technology for real-time location tracking through a variety of projects for years. However, problems have been posed by Global Positioning System (GPS) based on the location awareness technology and active bat system using sound waves in terms of the space and construction costs. In this regard, this study attempted to suggest passive RFIF-based Indoor Positioning System (IPS) for tracking the location of the moving objects (humans and assets) in real time.

2 Related Study

In this chapter, the concept of location awareness technologies such as GPS, active bat system and passive RFID is summarized as follows.

2.1 Location Awareness Technology

The location awareness technology is considered applied technology required to implement the recognition systems, and it includes triangulation, scene analysis and proximity approach as typical location awareness technologies [1].

The system implemented based on the location awareness technologies can be divided into Macro approaches including GPS, Micro approaches such as ultrasonic and infrared methods and multi-hop location awareness method in accordance with coverage areas [2].

Table 1 Advantages and disadvantages of ultrasound-based indoor positioning [4]

<i>Advantages</i>
Very precise positioning accuracy
3-dimensional positioning available
<i>Disadvantages</i>
Expensive infrastructure installation costs
Transmitter location information needs to be known in advance, and interference problems may occur depending on the placement of transmitter.

The most commonly used location awareness systems include methods using ultra sonic waves and infrared rays, ZigBee, RFID and wireless LAN-based method (WLAN) [3].

2.2 *Ultrasound-Based Positioning Technology*

As a method using ultrasound, it detects the location of objects by using the transfer rate difference between fast RF signals and relatively lower ultrasound [4]. As a method using ultrasound, it detects the location of objects by using the transfer rate difference between fast RF signals and relatively lower ultrasound [4]. The advantages and disadvantages of ultrasound-based positioning technology are summarized in Table 1.

2.3 *Infrared-Based Positioning Technology*

For the location detection system using infrared rays, infrared sensors are installed on the ceiling of office and active badge, a infrared generator with the form of badge is attached to people. The active badge is a system that has each unique identification number and finds the location of specific users through transmitting the identification number periodically (approximately once per second) and detecting infrared signals from infrared sensors on the ceilings [4].

2.4 *Radio Frequency Identification System*

Radio Frequency Identification (RFID) is a next-generation recognition technology to manage the information of various objects such as food, animals and things through wireless [5].

RFID system can be largely divided into positive and passive forms, and the positive system is characterized by the fact that self-RF signal transmission is

possible in the tag, and power supply is provided by batteries. In addition, it has its advantages in that long-distance (more than 3 M) transmission and combination with sensors can be achieved, but the disadvantage is that it has restrictions on the costs and operating time due to the use of batteries. On the other hand, passive system can be implemented at low costs without using batteries since it reflects signals from readers and is operated by power supply from radio signals of readers, but it poses its disadvantage of limited long-term transmission [6–10].

2.5 Real Time Location Systems

RTLS is a technology that detects the location of objects in real time as shown in Fig. 1. As a similar technology, there is a Global Positioning System (GPS), but it cannot be used in the shaded area. Accordingly, RTLS using short-range communications technologies such as i-Fi, Zigbee, Bluetooth and RFID is recommended instead [11].

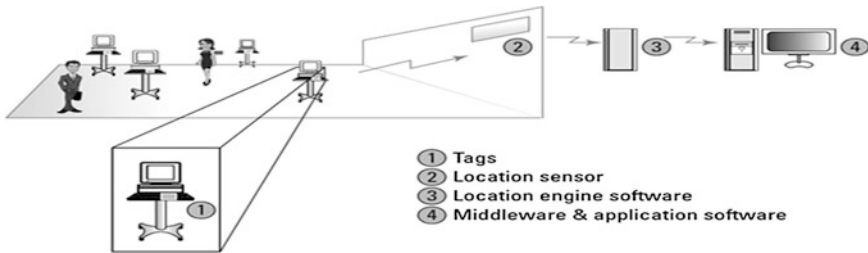


Fig. 1 Components of RTLS

3 Proposed Method

To detect access stages of moving objects and trace the moving path, areas of the interior are divided in accordance with security standards, and RFID readers and antennas are installed in each area.

3.1 Real-Time Location Tracking System

Figure 2 shows the definition of service structure for real-time location tracking, which is composed of a console program to control system operation, a monitor program that displays the movement of moving objects on the screen in real-time

and IPS middle ware to find mobility between security zones through data refining and computing process after collecting tag data of moving objects from passive reader installed in the security zones.

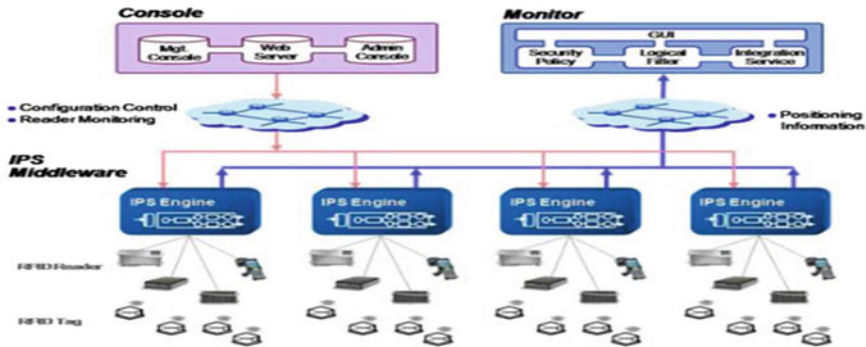


Fig. 2 Conceptual diagram of real-time location tracking system

IPS control RFID reader directly as a software module to figure out the path information and real-time mobility information between security zones using the pattern of stage data stream collected from the moving objects. Through the collection of various events occurred from RFIF readers applicable to end-point in each IPS middle ware, an analysis on the event information and individual monitoring is supported.

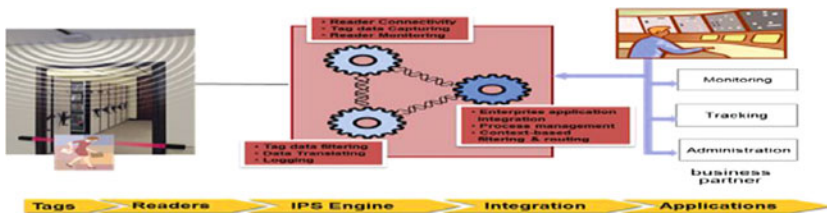


Fig. 3 Operation principles of real-time location tracking systems

3.2 Operation Principles of Real-Time Location Tracking System

Figure 3 represents the operation principle of real-time location tracking systems, which makes it possible to monitor the monitoring events through imagination of the history of moving objects in IPS engine once tag information is recognized by readers. IPS Engine carries out functions that process and refine monitoring events connected with a series of moving objects occurred from readers linked in the network installed through choke point method, and deliver the events to

management systems in the network taking the events in the form of XML. Once is remote powered server is operated, the monitoring events delivered from IPS Engine is monitored through the monitoring screen.



Fig. 4 IPS monitoring screen

Through the IPS monitoring screen shown in Fig. 4, it is possible to identify the moving path and location within the security zone of specific users through the configuration of the indoor space map for the location tracking data monitoring using Building Information Modeling (BIM) Tool.



Fig. 5 Screen that shows the moving path of specific users

As shown in (Fig. 5), the history of the users suspected in the occurrence of internal security threats can be finally identified through the screen that shows the moving path of specific users using IPS system.

4 Location Tracking System Test Results

For the application of the location tracking systems, an actual test was carried out by setting certain space of K institution as security zone.

The test was conducted by dividing the form of moving objects' wearing the tag into different types of necklace type and accessory type in the shirt pocket and pants pocket.

Execution procedures were implemented 50 times respectively in the order of entry and exit from each room as shown in (Fig. 4).

Table 2 Test results

	necklace type		shirt pocket		pants pockets	
	In	OUT	In	OUT	In	OUT
1	O		O		O	
2	O		O		O	
3	O		O		O	
4	O		O		O	
5	O		O		O	
6	O		O		O	
7	O		O		O	
8	O		O		O	
9	O		O		O	
10	O	X	X		O	X
11	O	O	O		X	O
12	O		O		X	
13	O		O		O	
14	O		O		O	
15	O		O		O	
16	O		O		O	
17	O		O		X	
18	O		O		O	
19	O		O		O	
20	O		O		O	
21	O		O		O	
22	O		X		O	X
23	O		O		X	O
24	O		O		O	X
25	O		O		O	
26	O		X		O	X
27	O		O		O	
28	O		O		O	
29	O		X		O	X
30	O		O		O	
31	O		O		O	
32	O		O		O	
33	O		O		O	
34	O		O		O	
35	O		O		O	
36	O		O		O	
37	O		O		O	X
38	O		O		O	X
39	O		O		O	X
40	O		O		O	X
41	O	X	O		O	X
42	O	O	O		O	X
43	O		O		O	X
44	O		O		X	X
45	O		O		X	X
46	O		X		X	X
47	O		O		X	X
48	O		O		X	X
49	O		O		X	X
50	O		O		X	X

From the test results, the necklace type was found to be IN(50/50), OUT(48/50), and accessory type in the shirt pocket turned out to be IN(44/50), OUT(50/50), and pants pockets IN(41/50), OUT(35/50). The test results of Table 2 showed that performance degradation occurred in accordance with the location of the tag won by the moving objects.

5 Conclusion

Since the active bat systems using the ultrasound and GPS based on the location awareness technology have problems of space and construction costs. It is required to establish more efficient system to monitor the location recognition in real-time. In this connection, this study proposed the IPS to trace the location of moving objects(people, assets) in real time using passive RFID. For an actual test, RFID readers and antennas were installed within the established security zone in a certain space of K institution, and the actual test was carried out by dividing the form of moving objects wearing the tag into different types of necklace type and accessory type in the shirt pocket and pants pocket. As a result of the test, it turned

out that the performance degradation occurred in accordance the location of the tag won by subjects, and the best result was found in the necklace-type tag.

Acknowledgments This work was supported by a grant from Kyonggi university advanced Industrial Security Center of Korea Ministry of Knowledge Economy.

References

1. Baek SK, Park MG, Lee KS (2003) Location tracking of an object in a room using the passive tag of an RFID system. *The Korean Railway*, p 569
2. Song YS, Park SJ (2011) IT convergence location aware technology. *The Korea Inst Commun Inf Sci* 28(5):5
3. Kim SJ, Kim YM (2011) Situation and implications of ATM in Korea and foreign countries. *Payment settlement and information technology*, 44th edn. pp 76
4. Cho YS, Cho SY, Kim BD, Lee SH, Kim JC, Choi WS (2007) Technical trend of indoor/outdoor seamless positioning. *ETRI J* 22(3):22
5. Kim BM, Shim MJ, Lee JE, Choi SH (2007) Ubiquitous sensor network location detection technologies and trends. *NIPA* 1291:28
6. Song TS, Kim TY, Lyou J (2006) Conformance evaluation method by successively applying self-running test mode in active RFID tag. *The Inst Electron Eng Korea* 31(11A):1160
7. Rhee SH, Chun JH, Park JA (2008) Performance improvement in passive tag based RFID reader. *The Korea Inst Commun Inf Sci* 45(6):807
8. Cole PH (2003) *Fundamentals in radio frequency identification*
9. MCRF355/360 Reader Reference Design, Microchip Technology Inc (2001)
10. De Vira G, Iannaccone G (2005) Design criteria for the RF section of UHF and microwave passive RFID transponders. *IEEE Trans Microw Theory Tech* 53(9):2278–2290
11. Kim JJ, Son SH, Choi H, Baek YJ (2010) Mobile reader selection for improving precision of location estimation in RTLS. *Korean Inst Sci Eng* 16(1):45

SmartKeyboard for the Disabled in Smartwork

Juyoung Park, Seunghan Choi and Do-Young Kim

Abstract This paper proposes a user customizable UI mechanism (SmartKeyboard) for people with disabilities in smartwork environment. In which, a user can utilize customized input UI through network assistance (such as network server, cloud service) anywhere, anytime. Currently we have developed user customizable input UI as a keyboard form, but our research result can be evolved into various types such as ATM, kiosk, even TV remote controllers.

Keywords Smartwork · Smart UI · Accessibility

1 Introduction

Recently a newly developed keyword ‘Smartwork’ is booming because of several social issues such as prohibiting CO2 emission, facing low-birth rate and aging society, and improving inefficient work style. Though ‘Smartwork’ is a newly generated keyword, it is not a new work style. Because the work style already has been described as telework or telecommute is common in other advanced countries [1].

J. Park (✉) · S. Choi · D.-Y. Kim
Smartwork Research Team, ETRI, 161, Gajeong-dong, Yusong-gu, Daejeon, South Korea
e-mail: jypark@etri.re.kr

S. Choi
e-mail: shchoi@etri.re.kr

D.-Y. Kim
e-mail: dyk@etri.re.kr

The Smartwork in Korea [2] is a nationwide strategy for future-oriented work style; currently some of large-sized companies have deployed Smartwork service such as mobile office to improve their work style. However, providing accessibility methods for the people with disabilities are not seriously touched yet. However, providing accessibility is deeply considered in other advanced countries; the Rehabilitation Act. 508 (US) [3] is a good example of how other country takes effort in.

This paper proposes the smart UI mechanism for the people with disabilities in smartwork environment. In which, a user can utilize customized input UI through network servers at anywhere, anytime, and with any devices. Currently we have developed user customizable input UI as a keyboard form, but it can be evolved into various types such as ATM, kiosk, even TV remote controllers.

This paper starts with the importance of Smart UI by touching other counties' smartwork services and then proposes the smart UI framework and finally shows our research results (SmartKeyboard device and its platform) and its use cases.

2 Related Works

2.1 Keyboards for the People with Disabilities

There have been many researches and products on accessible IT devices which can help people who feel difficulties in using typical IT devices, such as keyboard, mouse etc.

Some of notable efforts are: (1) the reinforcement input screen, (2) full-size keyboard for the elderly, (3) symbol keyboard for the people with disabilities, (4) a simple keyboard for the students in learning disabilities. Figure 1 shows some specially designed keyboards for the people with disabilities.

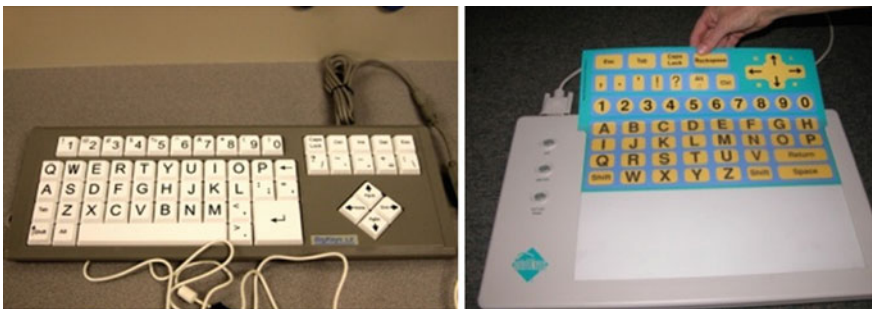


Fig. 1 Specially designed keyboards for the disable

2.2 Germ-Free Touch Screen Keyboard

Recently a touch screen based keyboard is getting attention because of its technology-based infection prevention and control solutions. As an example, a solution from Cleankeys Inc., based in Canada, presents computer keyboards which are easy to clean and disinfect [4].

Also there is a very interesting virtual keyboard which for touch screens and surfaces that adapts to the user’s natural finger positions and allows users to touch-type on smooth surfaces [5].

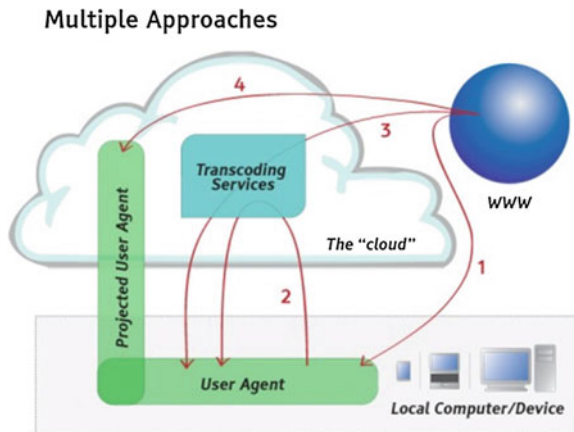
2.3 Global Public Inclusive Infrastructure

The purpose of the Global Public Inclusive Infrastructure (GPII) [6] is to ensure that everyone who faces accessibility barriers due to disability, literacy, or aging, regardless of economic resources, can access and use the Internet and all its information, communities, and services for education, employment, daily living, civic participation, health, and safety.

The GPII would not create new access technologies or services, but would create the infrastructure for making their development, identification, delivery, and use easier, less expensive, and more effective. Like building a road system does not provide transportation but greatly enhances the ability of car companies and others to do so—and provides an infrastructure that car companies themselves cannot do. The Internet is the infrastructure for general information and commerce. The GPII enhancements to the Internet would provide the infrastructure to enable the Internet to be truly inclusive for the first time.

GPII is a paradigm shift. The GPII will, for the first time, introduce automatic personalization of user interfaces and user context adaptation based on user

Fig. 2 The GPII supports a number different delivery models: (1) downloaded user agents, (2) on-demand web service, (3) proxy-based transcoding, and (4) web-based user agents delivery models



preferences. Each information and communication technology (ICT) device will be able to instantly change to fit users as they encounter the device, rather than requiring users to figure out how to adapt, configure or install access features they need. It also introduces a system of shared components and services to reduce cost, increase interoperability, and foster innovation (Fig. 2).

3 User-Customizable SmartKeyboard

With this paper, we propose a user-customizable keyboard (SmartKeyboard) device and its platform which can provide dynamic keyboard layout. The keyboard layout can be selected and can be designed; even it can be re-designed by any user. As shown in Fig. 3, the propose architecture consists of (1) SmartKeyboard device, (2) PC agent, (3) user-customizable UI distribution platform server and (4) UI development toolkits.

Each part will be discussed in detail in the following paragraph.

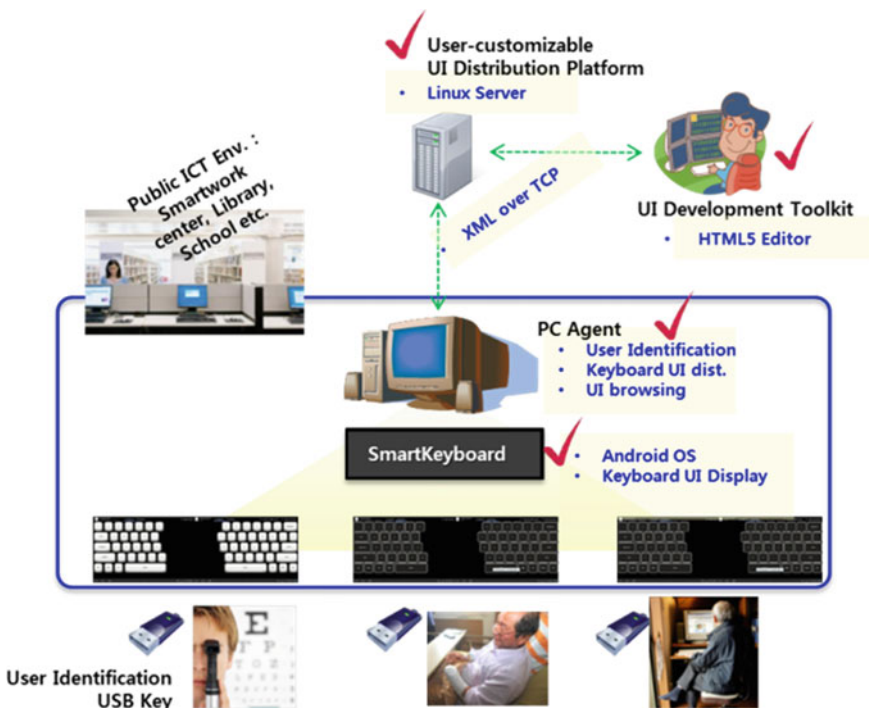


Fig. 3 A sample line graph using colors which contrast well both on screen and on a black-and-white hardcopy

3.1 SmartKeyboard Device

The SmartKeyboard is a ‘common’ virtual keyboard rather than a specially devised physical dummy type: the S/W driven intelligent keyboard with touch panels and processors. Inside of the SmartKeyboard, it consists of following entities:

- Embedded terminal operating system to control hardware For the PC and connect the PC interface,
- UI agent who displays the keyboard’s UI, and
- Space to hold basic keyboard UI and one or more downloaded keyboard UIs from network server.

3.2 PC Agent

PC agent is a very important part to connect SmartKeyboard with PC itself. PC agent communicates with manager installed in SmartKeyboard to read keyboard user’s intention and provide requested keyboard layouts (UIs) by a user.

The keyboard layout (UI) of SmartKeyboard is stored in UI distribution platform server and is chosen by the user; PC agent delivers requested keyboard layout to SmartKeyboard. To identify a user, PC agent may ask a user with a various way of identification methods such as ID/passwd, fingerprint, face/eye patterns, and so on; but in this paper, we have chosen USB memory stick to show its feasibility.

3.3 User-Customizable UI Distribution Platform

The keyboard layouts (UI) of SmartKeyboard are provided by UI distribution platform server (UI server); the UI server can be deployed in the form of a single-server system somewhere in the network or cloud, or it can be deployed in the form of a home gateway.

The key elements of UI distribution platform are user profile DB and keyboard UI DB; user profile DB is to manage registered users’ preference and keyboard UI DB is to store, up-/download UI from the users or pro-designers. This platform can map each user with preferred keyboard UI.

3.4 User-Customizable UI Toolkits

To make it possible to design keyboard layout, the user-customizable UI toolkit is very important. We defined keyboard layout schema in XML, and the toolkit

should generate keyboard pattern according to pre-defined keyboard layout schema.

Also this toolkit can easily upload the designed keyboard layout to UI distribution platform.

4 Implementation

To show the feasibility of SmartKeyboard suits, we implemented SmartKeyboard prototype as follows.

4.1 Smart-Keyboard Device

To hold various types of keyboard layout, the size of SmartKeyboard device should be large enough. Therefore we coupled two smart pads together to emulate SmartKeyboard.

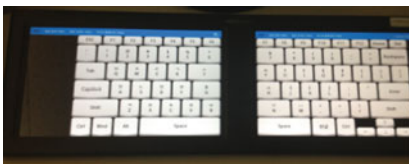
Figure 4 shows how we have coupled two smart pads to make SmartKeyboard.

4.2 SmartKeyboard UI distribution platform and UI Editor

Because the capacity of SmartKeyboard UI distribution platform is not quite important at this stage, we have implemented this platform on Linux server. We also provide UI editor inside of UI platform.

After connecting UI platform server, a user can browse existing SmartKeyboard UIs and then can select one or more keyboard UIs. Figure 5 shows UI browsing method.

Also a user can design one's own keyboard layout or modify any existing keyboard layout. After complete keyboard layout UI, a user can register and publish his UI. Figure 6 shows UI editing method.



SmartKeyboard	
OS	Android 2.3 (Gingerbread)
	CPU Dual Core T20
H/W	RAM 1GB RAM , 16GB Memory
	Size 200 x 550 x 10 (mm)

Fig. 4 SmartKeyboard device



Fig. 5 UI browsing screenshot

5 Use Cases

The SmartKeyboard can influence the following areas can be:

5.1 Cost Reduction of Specially Devised Keyboard

By providing a common keyboard suite instead of a separate keyboard solution, mass production can be possible and then lower the cost of manufacturing of the terminal.



Fig. 6 UI editor

5.2 Free from Bring One’s Own Keyboard

Because SmartKeyboard can adapt one’s disabilities, it can be placed in public work environment such as Internet café, library; a user with disabilities do not have to bring his own keyboard but can use SmartKeyboard of one’s preferred keyboard layout to use Internet browser.

5.3 Provide Optimized Keyboard UI According to the Degree of Disability

Changes in the degree of disability (worse or better) may cause one to replace his keyboard with a new H/W product. But with SmartKeyboard a user can redesign SmartKeyboard layout for him.

5.4 Language Impairment Support: Foreign Assistance

In addition to physical disability, this SmartKeyboard even can support ‘foreigners’ to use keyboard comfortably by using their own language keyboard layouts in public environment.

6 Conclusions

Until now, we discussed the architecture of SmartKeyboard and its use cases. The notable merit of SmartKeyboard will be as follows:

- Cost reduction through common terminal platform
- Optimized UI through S/W based keyboard layout editing
- International Language Support for foreigner
- Educational Usage in kindergarten
- Adaptable to Alternative Augmentative and Alternative Communication (AAC)
- Future-oriented Universal UI for TV remote controller, car dashboards

Although we emphasize that SmartKeyboard can help people with disabilities through this paper, but we expect that this SmartKeyboard can make a new eco-system consisted of users, telecommunications and service providers, handset manufacturers, UI developers, and government.

Acknowledgments This research was supported by the Korea Communications Commission (KCC), Korea, under the ETRI R&D program supervised by the Korea Communications Agency (KCA) (KCA-2012-11921-05001).

References

1. Park J (2011) Trend of smartwork technology and future. RAPA, March 2011
2. Sung S (2011) Strategy of smartwork in Korea. KCC, Sep 2011
3. The Rehabilitation Act Amendments (Section 508). <http://www.access-board.gov/sec508/guide/act.htm>
4. Cleankeys keyboard. <http://cleankeysinc.com/>
5. LiquidKeyboard. <http://www.liquidkeyboard.org/>
6. GPII. <http://gpil.net/>

Addressing the Out-of-date Problem for Efficient Load Balancing Algorithm in P2P Systems

Khaled Ragab and Moawia Elfaki Yahia

Abstract Load-balancing is of major significance for large-scale decentralized networks such as Peer-to-Peer (P2P) networks in terms of enhanced scalability and performance. P2P networks are considered to be the most important development for content distribution and sharing infrastructures. Load balancing among peers in P2P networks is critical and a key challenge. This paper addresses the out-of-date problem as a result of node's state changes during loads movement among nodes. Consequently, this work proposes a load balancing algorithm that is based on extensive stochastic analysis and virtual server concept in P2P System. Finally, this work is complemented with extensive simulations and experiments.

Keywords Out-of-date problem • Peer-to-peer networks • Virtual servers • Load balancing

1 Introduction

Recently **Peer-to-Peer (P2P)** paradigm is an increasingly popular approach for developing various decentralized systems especially the internet applications. P2P is a class of applications that takes advantage of resources e.g. storage, cycles, content, human presence, available at the edges of the Internet [1]. P2P systems

K. Ragab (✉)
Computer Science Division, Mathematics Department, College
of Science, Ain Shams University, Cairo, Egypt
e-mail: kabdultawab@kfu.edu.sa

M. E. Yahia
Computer Science Department, College of Computer Science
and Information Technology, Hofuf, Saudi Arabia

offer an alternative to such traditional client–server systems for several application domains. They have emerged as an interesting solution for sharing and locating resources over the Internet. Moreover, P2P systems do not have a single point of failure and can easily scale by adding further computing resources. They are seen as economical as well as practical solutions in distributed computing. In P2P systems, every node (peer) of the system acts as both client and server (**servant**) and provides part of the overall resources/information available from the system. Each node often has different resource capabilities (e.g. processor, storage, and bandwidth) [2]. Thus, it is required that each node has a load proportional to its resources capabilities. On account of the dynamism natures of the P2P systems, it is difficult to ensure that the load is uniformly distributed across the system. In particular, this paper considers a P2P system of M nodes in which nodes join/leave, and data entity inserted/deleted continuously. Similarly to [3–5] this paper assumes node and data entity have been assigned identifiers that chosen randomly. Thus, there is a $\Theta(\log M)$ imbalance factor in the number of data entities stored at a node. Additionally, the imbalance factor becomes more worse if the P2P applications associate semantics with data entity IDs since IDs will not be uniformly distributed.

Consequently, it is important to design mechanisms that balance the system load. There are two distinct strategies to distribute the system workload [6]. First, load balancing algorithms that strive to equalize the workload among nodes. Second, load sharing algorithms which simply attempt to assure that no node is idle while jobs at other nodes are waiting for service. Load balancing techniques in P2P systems should be scalable and cope with its large size. They should place or re-place shared data entities optimally among nodes while maintaining an efficient overlay routing tables to redirect queries to the right node.

The communication delays among peers significantly alter the expected performance of the load balancing schemes. Due to such delay, the information that a particular peer has about other peers at any time is dated and may not accurately represent the current state of the other peers. For the same reason, a load sent to a recipient peer arrives at a delayed instant. In the mean time, however, the load state of the recipient peer may have considerably changed from what was known to the transmitting peer at the time of load transfer. This paper proposes a stochastic dynamic load balancing algorithm that tackles the out-of-date problem.

The remainder of this paper is organized as follows. [Section 2](#) introduces a survey for the load balancing algorithms. [Section 3](#) exposes the proposed stochastic load balancing model and algorithm. Evaluation of the proposed has been discussed in [Sect. 4](#). [Section 5](#) draws a conclusion of this paper.

2 Load Balancing Survey

Load balancing is the problem of mapping and remapping workload in the distributed system.

2.1 Load Balancing Design

Load balancing design determines how nodes communicate and migrate loads for the purpose of load balancing. It moves workload from heavily loaded nodes (senders) to lightly loaded nodes (receivers) to improve the system overall performance [15]. Load balancing design includes four components that can be classified as follows [16, 17].

- *Transfer policy*: It decides whether a node is in a suitable state to participate in a load transfer; either receiver or sender.
- *Location policy*: Once the transfer policy decides that a node is a receiver or sender. The location policy takes the responsibility to find a suitable sender or receiver.
- *Selection policy*: Once the transfer policy decides that a node is a sender, the selection policy specifies which load should be transferred. It should take into account several factors such as load transfer cost, and life time of the process that should be larger than load transfer time.
- *Information policy*: It decides when and how to collect system state information.

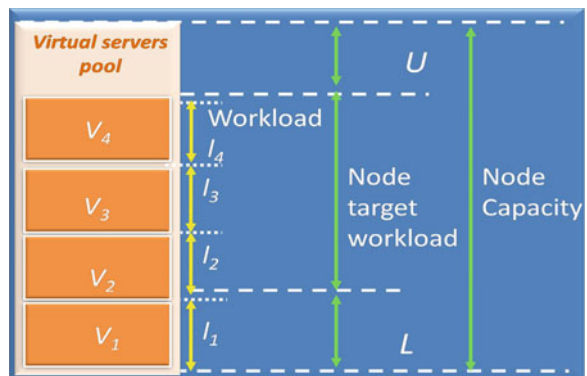
Load balancing designs are categorized into static and dynamic. With a static load balancing scheme, loads are scattered from sender to receiver through deterministic splits. Static schemes are simple to implement and easy to achieve with slight overhead [13]. They perform perfectly in homogenous systems, where all nodes are almost the same, and all loads are same as well. On the other hand, the dynamic load balancing schemes make decisions based on the current status information [15]. Accordingly, the transfer policy at certain node decides to be a sender or receiver, the selection policy selects the load to be transferred. The dynamic load balancing schemes perform efficiently when its nodes have heterogeneous loads, and resources. The typical architectures of dynamic load balancing schemes can be classified into centralized, distributed, and topological. In a centralized scheme, a central server “*coordinator*” receives load report from the other nodes, while overloaded nodes request the coordinator to find underloaded nodes [17]. In distributed architecture, each node has a global or partial view of the system status. Consequently, the transfer policy at each node can locally decide to transfer a load either out from it (*sender-initiated*) or into it (*receiver-initiated*) [18]. Then, the location policy at each node probes a limited number of nodes to find a suitable receiver or sender. *Kruger* [19], proposed symmetrically-initiated adaptive location policy that uses information gathered during its previous searches in order to keep track of the recent state of each node in the system. It finds a suitable receiver when a heavily-loaded node wishes to send a load out, and finds a suitable sender when a lightly-loaded node wishes to receive a load. Finally, in a system with large number of nodes, a topological scheme should be used [20]. It partitions nodes into groups. The load balance is performed in each group first, then, a global load balance among groups will be performed. However nodes in the

hierarchical architecture [21] are organized into a tree. Inner nodes gather the status information of its sub-trees. Then, load balancing is performed the leaves to the roots of the tree.

2.2 P2P Load Balancing

Load balancing is a critical issue for the efficient operation of the P2P systems. Recently, it attracted much attention in the research community especially in the *distributed hashed table (DHT)* based P2P systems. Namespace balancing is struggling to balance the load across nodes by ensuring that each node is responsible for a balanced namespace. This is valid only under the assumption of uniform workload and uniform node capacity. Otherwise, there is a $\Theta(\log M)$ imbalance factor in the number of objects stored at a node. To mitigate this imbalance two categories, *node placement* and *object-placement* load balancing techniques were proposed. In the node placement technique, nodes can be placed or replaced in locations with heavy loads. For example, a node in the *Mercury* load balancing mechanisms [7] is able to detect a lightly loaded range, and move there if it is overloaded. In object placement technique, objects are placed at lightly loaded nodes either when they are inserted into the system [11] or through dynamic load balancing schemes based on the virtual servers (VSs) concept [8], whose explicit definition and use for load balance was proposed by *Godfrey* et. al. [9] and Rao et. al. [11]. In [8], a virtual server represents a peer in the *DHT*; that is the storage of data items. In addition, routing takes place at the virtual server level rather than at the physical node level. In this paper, we assume virtual server as a virtual machine that is able to process a set of jobs like physical machine. Each physical node creates a pool of VSs as seen in Fig. 1. Load balancing could be achieved by migrating VSs from heavily loaded physical node to lightly loaded physical node. One main advantage of using VSs for balancing the load is that approach does not require any changes to the underlying network. In fact, the

Fig. 1 Node’s load specification



transfer of a virtual server can be implemented simply as a peer leaving and peer joining the system. In [11], Rao et. al. proposed three simple and static load balancing schemes: “one-to-one”, “one-to-many” and “many-to-many”. Godfrey et. al. combines both “one-to-many” and “many-to-many” schemes and uses them in different scenarios [9]. Clustered VSs scheme is presented in [12] that optimized the basic VS framework to reduce the overhead involved in the VS framework. However, VSs cannot be moved, and therefore, the scheme cannot respond to dynamic changes in network conditions.

This paper focuses on the design and analysis of P2P load balancing algorithm based on stochastic analysis [23, 24] and based on the VSs concept [8].

2.3 Challenges: P2P Load Balancing

Load balancing techniques in P2P systems are facing challenges coming from the characteristics of these systems. First, the size of the P2P system is large that means a scalable load balancing technique is required. Second, dissimilar to the traditional systems, nodes of a P2P system are not replicas and requests cannot be executed in any node. If nodes have dated, inaccurate information about the state of other nodes, due to random communication delays between nodes, then this could result in unnecessary periodic exchange of loads among them. For example, an overloaded node removes some of its virtual servers. However, such simple deletion will cause the problem of “load thrashing¹”, for the removed virtual servers may make other nodes overloaded. Consequently, this paper proposes a stochastic P2P load balancing algorithm that approximately determines the minimum amount of time to change the node’s state from overloaded to underloaded and vice versa. Comparing that time with the required time to migrate virtual servers enable us to come to a careful decision. Accordingly, the proposed algorithm undoubtedly avoids the load thrashing. To the best of the author’s knowledge, there is no any load balancing algorithm for the P2P system based upon the following stochastic analysis.

3 Load Sharing Algorithm

3.1 Model

This paper considers a P2P system consisting of M physical nodes (peers), denoted by P_i , $1 \leq i \leq M$. Each peer can be modeled as a queuing system, such as $M/M/1$,

¹ *Load thrashing* is a condition when the load balancing algorithm is engaged in moving virtual servers back and forth between nodes. .

M/D/1, etc. Each physical node P_i has a capacity C_i that corresponds to the maximum amount of load that it can process per unit of time. Nodes create virtual servers (VSs), which join the P2P network. Therefore, it can own multiple non-contiguous portions of the DHT’s identifier space. Each virtual server participates in the DHT as a single entry (e.g. routing table). Moreover, each virtual server stores data items whose IDs fall into its responsible region of the DHT’s identifier space. As seen in Fig. 1, a node P_i might have n VSs v_1, v_2, \dots, v_n ; where $n = VSset.size$. Each v_j has load l_j ; (for $j = 1, \dots, n$). The load of peer P_i in a unit of time is $L_i = l_1 + l_2 + \dots + l_n$. The utilization of a node’s P_i is L_i/C_i . From the perspective of load balancing, a virtual server represents certain amount of load (e.g. the load generated by serving the requests of the data items whose IDs fall into its responsible region) [25]. To avoid fluctuations in workload nodes should operate below their capacity. If a node finds itself receiving more load L_i than the upper target load U (i.e. $(L_i/C_i) > U$), it considers itself overloaded. A node P_i also has load L_i less than L is considered to be underloaded. An overloaded node is not able to store objects given to it route packets, or carry out computation, depending on the application.

Definition 1 A node P_i is in one of the following state as follows

$$S_i = \begin{cases} \text{Underloaded} & \text{if } Q_i < L \\ \text{Normal} & \text{if } L \leq Q_i \leq U \\ \text{Overloaded} & \text{if } U < Q_i \end{cases}$$

Clearly the state space Q_i consists of non-negative integers sub-divided into three disjoint regions $[0, L)$, $[L, U]$, and (U, ∞) corresponding to underloaded, normal, and overloaded state respectively.

A P2P system is defined to be balanced if the sum of the load L_i of a physical node P_i is smaller than or equal to the target load of the node for every node P_i , $1 \leq i \leq M$ in the system. When the system is imbalanced, the goal of a load balancing algorithm is to move VSs from overloaded node to underloaded one with minimum load transfer overheads.

The amount of overload to be transferred from the overloaded node P_i ; $1 \leq i \leq M$ is a random variable denoted by A is given by

$$A(p_i) = \max(0, Q_i - U) = \begin{cases} Q_i - U & \text{if } Q_i > U \\ 0 & \text{Otherwise} \end{cases}$$

Similarly, the amount of underload that can be accepted at the underloaded peer P_i ; $1 \leq i \leq M$ is a random variable denoted by B is given by

$$B(p_i) = \max(0, L - Q_i) = \begin{cases} L - Q_i & \text{if } Q_i < L \\ 0 & \text{Otherwise} \end{cases}$$

Definition 2 Let $\{Q(t); t \geq 0\}$ be a stationary² stochastic process with state space consisting of non-negative integers. Let S_i and S_j be two distinct non-negative numbers. The First Passage Time (FPT) between states S_i and S_j is denoted by $FPT(S_i, S_j)$, is given by

$$FPT(S_i, S_j) = \begin{cases} \inf \{t; Q(t) = S_j, Q(0) = S_i\} & \text{if } S_i \neq S_j \\ 0 & \text{if } S_i = S_j \end{cases}$$

It is a random variable which measures the minimum amount of time needed to reach state S_j from state S_i . We note that because the same stochastic process $Q(t)$ is stationary, translating the above events by a fixed amount of time has no effect upon the probability distribution of $FPT(S_i, S_j)$. In fact, a first passage time from state S_i to state S_j can be divided into two parts, namely the first transition out of state S_i (say S_k) followed by the first passage from S_k to S_j .

Assume that, $i < j$; since changes of state have unit magnitude in a birth and death of load, then

$$FPT_{ij} = FPT_{ik} + FPT_{kj} \quad i < k < j; \quad k = i + 1, i + 2, \dots, j - 1 \text{ thus}$$

$$FPT_{ij} = \sum_{k=i}^{j-1} FPT_{k,k+1} \quad i < j \tag{1}$$

Similarly, if $i > j$

$$FPT_{ij} = \sum_{k=j}^{i-1} FPT_{k+1,k} \quad i > j \tag{2}$$

Let $H_{ij}(t) = P\{FPT_{ij} \leq t\}$ and considering that the summands in both Eqs. (1) and (2) are independent. Thus, if we apply Laplace transformer $\tilde{H}_{ij} = \prod_{k=i}^{j-1} \tilde{H}_{k,k+1}(s)$ for $H_{ij}(t)$, then we can show that:

$$\tilde{H}_{ij}(s) = \prod_{k=i}^{j-1} \tilde{H}(s, k, k + 1) \quad ; i < j \quad (\text{Upward}) \tag{3}$$

$$\tilde{H}_{ij}(s) = \prod_{k=j}^{i-1} \tilde{H}(s, k + 1, k) \quad ; i > j \quad (\text{Downward}) \tag{4}$$

² A **stationary stochastic process** has the property that the joint distribution don not depend on the time origin. The stochastic process $\{Q(t); t \in \mathbb{S}\}$ is called stationary if $t_i \in \mathbb{S}$ and $t_i + s \in \mathbb{S}$, $i = 1, 2, \dots, k$ (k is any positive integer), then $\{Q(t_1), \dots, Q(t_k)\}$ and $\{Q(t_1 + s), \dots, Q(t_k + s)\}$ have the same joint distribution [22].

Clearly the distribution of the first passage time of unit downward is independent of starting state while the distribution of the first passage time of unit upward depends upon starting state.

3.2 Load Sharing Edge

The aim of this section is to study the *FPT* of the transition from normal state to overloaded, overloaded to normal, underloaded to normal, etc. For each transfer pair, *FPT* will be computed to predicate the future behavior of the transfer pair before the load transfer (i.e. virtual server migration) decision is taken.

Definition 3 Let $[Q(t), R(t)]$ be a transfer pair with $Q(t) = X$ and $R(t) = Y$, where $X > U$ and $Y < L$. **The Load Sharing Edge (LSE)** between Q and R is a random variable $E(X, Y)$ which is defined as follows:

$$E(X, Y) = \min\{FPT(X, U), FPT(Y, L)\}$$

Where, $FPT(X, U)$ is the first passage time to move from state X to state U and $FPT(Y, L)$ is the first passage time to move from state Y to state L .

LSE is the period of time within which the overloaded node must complete transferring load to the underloaded node before the overloaded node identifies that is unnecessary to transfer load or the underloaded node becomes ineligible to receive a transferred load. Assume the load transfer time is denoted by Δ . It is the time needed to package and send the load (i.e. the least loaded virtual server that will release overload) to sink R . Thus, the load transfer must be initiated only if $LSE > \Delta$. Since *LSE* is a random variable we need to formulate the transfer criterion in terms of probabilities. Assume the probability that *LSE* exceeds Δ is $P\{E(X, Y) > \Delta\}$. Therefore, the load transfer must be initiated if $P\{E(X, Y) > \Delta\}$ is large. These considerations led to the formulation of a class of rules so called *Quantile rules*. The quantile of a probability distribution function is defined as follows:

Definition 4 Let $F(t)$, $t \geq 0$ be the probability distribution function of a non-negative random variable X . Let $0 < \beta < 1$. The β -quantile of F is a non-negative real number q_β satisfying

$$1 - F(q_\beta) = \beta, \quad P\{X \geq q_\beta\} = \beta.$$

From Definition 4, the β -quantile rule for load transfer was introduced.

Definition 5 Given a transfer pair $[Q(t), R(t)]$ and a load transfer time is Δ . Also, for $0 < \beta < 1$, let q_β be the β -quantile of the probability distribution of the LSE between $Q(t)$ and $R(t)$. Then the load transfer is initiated only if $q_\beta > \Delta$.

The proposed algorithm in this paper uses a β -quantile rule before transferring load and ensures that $I \geq P\{E(X,Y) \geq \Delta\} \geq \beta$. In general β can be taken 0.90 or large.

The probability distribution of the random value LSE is given as follows

$$P_e(t; X, Y) = P\{E(X, Y) \leq t\} \text{ for } t \geq 0.$$

Thus, the probability distribution function of the load sharing edge LSE between pair $Q(t)$ and $R(t)$ is given by

$$P_e(t; X, Y) = 1 - [1 - F(t; X, U)] \times [1 - G(t; Y, L)].$$

Where $F(t; X,U)$ and $G(t; Y,L)$ are the probability distribution of the first passage time from X to U and from Y to L in the queues $\{Q(t)\}$, $\{R(t)\}$, respectively. Each node is modeled as $M/M/1$ queue, in which processes arrive according to a Poisson process with mean arrival rate λ , then processed with exponential service time [23, 24] and with mean service rate μ .

Lemma 1 Assume constant birth rates $\lambda = \lambda_0 = \lambda_1 = \dots$ and death rates $\mu = \mu_0 = \mu_1 = \dots$, then the probability distribution function is

$$H_{k+1,k}(\cdot) = H_{1,0}(\cdot) \quad k = 0, 1, \dots$$

Proof

$FPT_{k+1,k}$ is the time that elapses before the cumulative number of deaths first exceeds the cumulative number of births when $X(0) = k + 1$. Also, the value of $H_{k+1,k}(\cdot)$ do not depend on $X(0)$, [23, 24]. □

Lemma 2 The Laplace–Stieljes transform of the probability distribution function of the first passage time from state k to state 0 in an $M/M/1$ queue is

$$\tilde{H}_{k,0}(s) = \left[\frac{s + \lambda + \mu - \sqrt{(s + \lambda + \mu)^2 - 4\lambda\mu}}{2\lambda} \right]^k \tag{5}$$

Proof

Assume that, the first passage time FPT_{ij} can be expressed as

$$FPT_{ij} = S_1 + \begin{cases} FPT_{i+1,j} & \text{if } X(s_1) = i + 1 \\ FPT_{i-1,j} & \text{if } X(s_1) = i - 1 \end{cases}$$

Where S_1 is the time of the first transition. Assume that $H_{ij}(t) = P\{FPT_{ij} \leq t\}$, as well FPT_{ij} can be upward or downward after the first transition S_1 . Using Theorem 4–7, [23, 24], $H_{ij}(t)$ can be expressed as follows

$$H_{ij}(t) = \lambda_i \int_0^t H_{i+1,j}(t-x)e^{-(\lambda_i+\mu_i)x} dx + \mu_i \int_0^t H_{i-1,j}(t-x)e^{-(\lambda_i+\mu_i)x} dx \quad (i)$$

Taking *Laplace–Stieltjes* transform on both side of Eq. (i) and use the convolution property, then the following equation can be obtained:

$$\tilde{H}_{i,j}(s) = \frac{\lambda_i \tilde{H}_{i+1,j}(s) + \mu_i \tilde{H}_{i-1,j}(s)}{s + \lambda_j + \mu_j} \quad (ii)$$

Thus,

Set $i = 1$, and $j = 0$

$$\tilde{H}_{1,0}(s) = \frac{\lambda_0 \tilde{H}_{2,0}(s) + \mu_0 \tilde{H}_{0,0}(s)}{s + \lambda_0 + \mu_0} \quad (iii)$$

From lemma 1, $\lambda = \lambda_0$, $\mu = \mu_0$ and from Eq. (4)

$$\tilde{H}_{i,j}(s) = \prod_{k=j}^{i-1} \tilde{H}_{k+1,k}(s) \quad ; i > j. \text{ Thus, } \tilde{H}_{2,0}(s) = \tilde{H}_{1,0}(s) \times \tilde{H}_{2,1}(s) = (\tilde{H}_{1,0}(s))^2$$

from Eq. (iii), we obtain the following quadratic equation

$$\lambda [\tilde{H}_{1,0}(s)]^2 - (s + \lambda + \mu) \tilde{H}_{1,0}(s) + \mu = 0 \quad (iv)$$

Equation (iv) has two solutions, we consider the solution which satisfies that $\tilde{H}_{1,0}(s) \leq 1$

$$\tilde{H}_{k,0}(s) = \left[\frac{s + \lambda + \mu - \sqrt{(s + \lambda + \mu)^2 - 4\lambda\mu}}{2\lambda} \right]; \tilde{H}_{1,0}(s) \leq 1, \text{ for real.}$$

Hence,

$$\tilde{H}_{k,0}(s) = [\tilde{H}_{1,0}(s)]^k = \left[\frac{s + \lambda + \mu - \sqrt{(s + \lambda + \mu)^2 - 4\lambda\mu}}{2\lambda} \right]^k. \quad \square$$

Corollary 1 *The density function of the first passage time from state k to state 0 in an M/M/1 queue is*

$$h_{k,0}(t) = ke^{-(\lambda+\mu)t} I_k \left(2t\sqrt{\lambda\mu} \right) \frac{(\mu/\lambda)^{k/2}}{t} \quad (6)$$

For $t > 0$ m ; where I_k is the Modified Bessel function of order k .

Proof

If Eq. (5) is bona fide Laplace transform, it is the Laplace transform of $h(t)$, [23, and 24]. From Laplace transform,

$$\wp\{e^{-ct}f(t)\} = f(s + c); \text{ set } (c = \lambda + \mu) \text{ and } w = (s + \lambda + \mu)$$

So we can write Eq. (5) as

$$\tilde{H}_{1,0}(w) = \left[\frac{w - \sqrt{w^2 - 4\lambda\mu}}{2\lambda} \right] \square$$

While the Laplace transformation $\wp\{I_1(at)/t\} = \frac{s + \sqrt{s^2 + a^2}}{a}$, [23] then we can say that the numerator is the Laplace transform of $2(\lambda\mu)^{1/2}I_1(2t(\lambda\mu)^{1/2})/t$; thus, $H_{1,0} = (s)$ is the Laplace transform of Eq. (5). So we have

$$h_{1,o}(t) = ke^{-(\lambda+\mu)t}I_1\left(2t\sqrt{\lambda\mu}\right)\frac{\sqrt{(\mu/\lambda)}}{t}. \quad \square$$

Hence,

$$h_{k,o}(t) = ke^{-(\lambda+\mu)t}I_k\left(2t\sqrt{\lambda\mu}\right)\frac{(\mu/\lambda)^{k/2}}{t}; t > 0$$

Lemma 3 (Downward) *The probability distribution of the first passage time from state k to state 0 in an M/M/1 queue is:*

$$H(x; k, 0) = 1.0 - k\mu^k \sum_{n=0}^{\infty} \frac{(\lambda\mu)^n}{(\lambda + \mu)^{2n+k}} \Gamma(2n + k, x) \quad (7)$$

Where $\Gamma(\)$ is the incomplete Gama function.

Proof

Assume that the modified Bessel function of order k is

$$I_k(x) = \left(\frac{x}{2}\right)^k \sum_{n=0}^{\infty} \frac{(x^2/4)^n}{n!(n+k)!}; x \geq 0$$

Set $x = 2t\sqrt{\lambda\mu}$. Thus,

$$I_k(2t\sqrt{\lambda\mu}) = \left(\frac{2t\sqrt{\lambda\mu}}{2}\right)^k \sum_{n=0}^{\infty} \frac{(t^2\lambda\mu)^n}{n!(n+k)!}; x \geq 0$$

By substitute into Eq. (6), we can compute the density function $h_{k,o}(t)$ as follows:

$$h_{k,o}(t) = k\mu e^{-(\lambda+\mu)t} \sum_{n=0}^{\infty} \frac{(\lambda\mu)^n t^{k+2n-1}}{n!(n+k)!}; t \geq 0$$

From the definition of the probability distribution function, we have $H_{k,0}(x) = \int_0^x h_{k,0}(t) dt$ and $H_{k,0}(x) = 1 - \int_x^\infty h_{k,0}(t) dt$.

Thus, $H_{k,0}(x) = 1 - \int_x^\infty k\mu^k e^{-(\lambda+\mu)t} \sum_{n=0}^\infty \frac{(\lambda\mu)^n \mu^{k+2n-1}}{n!(n+k)!} dt$.

If we exchange the infinite sum and the integral we get

$$H_{k,0}(x) = 1 - \sum_{n=0}^\infty \frac{(\lambda\mu)^n}{n!(n+k)!} \int_x^\infty e^{-(\lambda+\mu)t} t^{k+2n-1} dt ; \Gamma(i, x) = \int_x^\infty e^{-t} t^{i-1} dt.$$

Hence,

$$H(x; k, 0) = 1.0 - k\mu^k \sum_{n=0}^\infty \frac{(\lambda\mu)^n}{(\lambda + \mu)^{2n+k}} \Gamma(2n + k, x). \quad \square$$

Lemma 4 *The probability density function of the first passage time of the M/M/1 queue from state 0 to state 1 is $h_{0,1}(t) = \lambda e^{-\lambda t}$.*

Proof

Let $\tilde{h}_{i,j}(s) = \frac{\lambda_i \tilde{h}_{i+1,j}(s) + \mu_j \tilde{h}_{i-1,j}(s)}{s + \lambda_j + \mu_j}$ set $i = 0, j = 1$ (i.e. there is no service $\mu_0 = 0$) and $\tilde{h}_{0,1}(s) = \frac{\lambda_0 \tilde{h}_{1,1}(s)}{s + \lambda_0}$. For simplicity set $\lambda_0 = \lambda$. Hence $\tilde{h}_{0,1}(s) = \frac{\lambda}{s + \lambda}$; $\tilde{h}_{1,1} = 1$ since $FPT_{11} = 0$. By computing the inverse of the Laplace transformation we get:

$$h_{0,1}(t) = \lambda e^{-\lambda t}. \quad \square$$

Lemma 5 (Upward) *The probability distribution function of the first passage time of the M/M/1 queue from state i to state j ; $i < j$ is*

$$H_{i,j}(t) = 1 + \lambda^{j-i} \sum_{k=1}^j C_k e^{(-r_k t)} \tag{8}$$

where $r_k; 1 \leq k \leq j$, are j distinct roots of the polynomial of degree j defined recursively as:

$$D(s-1) = 0, D(s, 0) = 1; D(s, 1) = s + \lambda$$

also, for

$$D(s, j) = (s + \lambda + \mu)D(s, j-1) - \lambda\mu D(s, j-2); j \geq 2$$

$$1 \leq k \leq j, C_k = (s + r_k) \frac{D(s, i)}{sD(s, i)} \Big|_{s=-r_k}$$

Proof

In lemma 4, we have prove $\tilde{h}_{0,1}(s) = \frac{\lambda}{s+\lambda}$ that is the Laplace transformation of the probability distribution of the first passage time of FPT_{0j} . Assume that

$$\tilde{h}_{i,j}(s) = \frac{\lambda_i \tilde{h}_{i+1,j}(s) + \mu_j \tilde{h}_{i-1,j}(s)}{s + \lambda_j + \mu_j} \quad \text{Set}$$

$$i = k, j = k + 1, \quad \lambda_k = \lambda_{k+1} = \lambda, \quad \mu_k = \mu_{k+1} = \mu$$

$$\tilde{h}_{k,k+1}(s) = \frac{\lambda \tilde{h}_{k+1,k+1}(s) + \mu \tilde{h}_{k-1,k+1}(s)}{s + \lambda + \mu} \quad (i)$$

But, $\tilde{h}_{k+1,k+1}(s) = 1$ since $FPT_{k+1,k+1} = 0$,

$$\tilde{h}_{k-1,k+1}(s) = \tilde{h}_{k-1,k}(s) \tilde{h}_{k,k+1}(s) \quad .$$

Hence,

$$\tilde{h}_{k,k+1}(s) = \frac{\lambda + \mu \tilde{h}_{k-1,k}(s) \tilde{h}_{k,k+1}(s)}{s + \lambda + \mu}$$

$$\tilde{h}_{k,k+1}(s) = \frac{\lambda}{s + \lambda + [\mu(1 - \tilde{h}_{k-1,k}(s))]} \quad (ii)$$

Using mathematical induction, we can prove that Eq. (ii) is satisfied for all $k \geq 0$. It can be rewritten as the ratio of two functions $N(s, k)$ and $D(s, k)$. These functions can be defined as follows:

$$L(1) = -\lambda, \quad L(k) = -(\lambda + \mu), \quad k \geq 2$$

$$M(k) = \lambda\mu \quad k \geq 1$$

Thus,

$$N(s, k) = \lambda D(s, k - 1), \quad k > 1; D(s, 0) = 1 \quad \text{Where } D(s, k)$$

$$D(s, k) = [s - L(k)] \times D(s, k - 1) - M(k) \times D(s, k - 2); \quad k \geq 1$$

is a polynomial of degree $k; k \geq j$.

Thus, Eq. (ii) can be rewritten as follows:

$$\tilde{h}_{k-1,k}(s) = \frac{N(s, k)}{D(s, k)} = \frac{\lambda D(s, k - 1)}{s + \lambda + \mu D(s, k - 1) - \lambda\mu D(s, k - 2)}, \quad k \geq 2$$

But for general transitions from state i to state $j: 0 \leq i \leq j$, is

$$\tilde{h}_{i,j}(s) = \frac{\lambda D(s, i)}{D(s, i + 1)} \times \frac{\lambda D(s, i + 1)}{D(s, i + 2)} \times \dots \times \frac{\lambda D(s, j - 2)}{D(s, j - 1)} \times \frac{\lambda D(s, j - 1)}{D(s, j)}$$

Thus, we can obtain the Laplace transform of the density of the upward transition from state i to state $j, \tilde{h}_{i,j}(s) = \frac{\lambda^{j-i} D(s,i)}{s D(s,j)}$ by canceling the common terms from the denominator and the numerator from the above equation.

From the definition $LT \left[\int_0^x h(t) dt \right] (s) = \frac{LT[h](s)}{s}$, thus

$$\tilde{H}_{i,j}(s) = \frac{\lambda^{j-i} D(s, i)}{sD(s, j)}$$

It is a relation function in which the numerator polynomial has degree I while the denominator polynomial has degree $(j + I)$ and $(i < j)$. Thus, we can expand $\tilde{H}_{i,j}(s)$ into a finite sum of partial fractions as follows: If $D(0, j) = \lambda^j$ for all $j \geq 1$ and $\lambda > 0$ then zero cannot be a root of $D(s, j)$; $j \geq 1$ and it can be written in the following form:

$D(s, j) = (s + r_1)(s + r_2) \dots (s + r_j)$ where r_k for all $k = 1$ to j are roots of $D(s, j)$. $\tilde{H}_{i,j}(s) = \lambda^{j-i} \left(\frac{C_0}{s} + \sum_{k=1}^j \frac{C_k}{(s+r_k)} \right)$ Hence, it can be written as follows:

where $C_k = (s + r_k) \frac{D(s, i-1)}{sD(s, j)} \Big|_{s=-r_k}$; $1 \leq j \leq k$ and

$$C_0 = \frac{D(0, i)}{D(0, j)} = \frac{\lambda^i}{\lambda^j} = \lambda^{i-j}$$

$$\tilde{H}_{i,j}(s) = \frac{1}{s} + \lambda^{j-i} \sum_{k=1}^j \frac{C_k}{(s + r_k)}$$

Accordingly, we can invert the Laplace Transform of the above equation. But, each term in the right hand side is in the form $\frac{\alpha}{(s+\beta)}$ α, β Where α, β are constants. Each term has inverse Laplace transformation $\alpha e^{\beta t}$. Hence,

$$H_{i,j}(t) = 1 + \lambda^{j-i} \sum_{k=1}^j C_k e^{(-r_k t)}$$

Theorem Let $[Q(t), R(t)]$ be a transfer pair that consists of M/M/1 queues. Let m be the amount of overload and n the amount of underload. Then the probability distribution function of the Load Sharing Edge (LSE) is

$P_e(t; m, n) = 1 - m\lambda^n \mu^m \left[\sum_{k=1}^L C_k e^{-r_k t} \right] \times \left[\sum_{k=0}^{\infty} \frac{(\lambda\mu)^k}{(\lambda+\mu)^{2k+m}} \Gamma(2n + m, x) \right]$ where, $r_k, 1 \leq k \leq L$; are the roots of the polynomial defined recursively as

$$D(s, -1) = 0; D(s, 0) = 1; D(s, 1) = s + \lambda$$

$$D(s, L) = (s + \lambda + \mu)D(s, L - 1) - \lambda\mu D(s, L - 2); L \geq 2; 1 \leq k \leq L$$

Also, for

$$1 \leq k \leq L \quad C_k = (s + r_k) \frac{D(s, L - n)}{sD(s, L)} \Big|_{s=-r_k}$$

Proof

The transfer pair has the probability distribution function $P_e(t; i, j)$ of the LSE; $i > U, j < L$ which is defined by

$P_e(t; i, j) = 1 - [1 - F(t; i, U)] * [1 - G(t; j, L)]$ where m is the amount of overload ($m = i - U$), and n is the amount of underload ($n = L - j$) then

$P_e(t; m, n) = 1 - [1 - F(t; m, 0)] * [1 - G(t; L - n, L)]$ for an M/M/1 queue case $F(t, m, 0)$ and $G(t, L - n, L)$ have been derived from 3 and 5 respectively. Hence,

$$P_e(t; m, n) = 1 - m\lambda^n \mu^m \left[\sum_{k=1}^L C_k e^{-r_k t} \right] \times \left[\sum_{k=0}^{\infty} \frac{(\lambda\mu)^k}{(\lambda + \mu)^{2k+m}} \Gamma(2n + m, x) \right]. \square$$

Due to the infinite number of terms in the probability distribution $P_e(t; m, n)$ of the LSE in M/M/1, the following lemma will drive a formula for LSE as finite number of terms as follows.

Lemma 6 For a transfer pair $[Q(t), R(t)]$ with an amount of overload m and an amount of underload n , the Mean Load Sharing Edge $MLSE(m, n)$ is

$$MLSE(m, n) = -\lambda^n \sum_{k=1}^L C_k \left(\frac{1}{r_k} - \tilde{F}(r_k, m, 0) \right)$$

where $r_1, r_2, \dots, r_k, C_k$ are constants defined in the previous theorem. Also,

$$\tilde{F}_{m,0}(s) = \left[\frac{s + \lambda + \mu - \sqrt{(s + \lambda + \mu)^2 - 4\lambda\mu}}{2\lambda} \right]^m$$

Proof

Since $P_e(t; m, n) = 1 - [1 - F(t; m, 0)] \times [1 - G(t; L - n, L)]$ we obtain,

$$MLSE(m, n) = \int_0^{\infty} P_e(t; m, n) dt = \int_0^{\infty} [1 - F(t; m, 0)] \times [1 - G(t; L - n, L)] dt$$

From lemma 5, we get

$$MLSE(m, n) = - \int_0^{\infty} \left[\lambda^n \sum_{k=1}^L C_k e^{-r_k t} \right] \times [1 - F(t; m, 0)] dt$$

But the Laplace transform $LT[q](s)|_{s=0} = \int_0^{\infty} q(t) dt$ and $LT[e^{-\alpha t} q](s)|_{s=0} = LT[q](s + \alpha)$. Set $\alpha = r_k$, then

$$MLSE(m, n) = -\lambda^n \sum_{k=1}^L C_k \left(\frac{1}{r_k} - \tilde{F}(r_k, m, 0) \right)$$

Consequently, it has been observed that for queuing models in which job arrival and processing rates are independent of queue size, such as $M/M/1$ queues. The distribution of LSE depends only on the amounts of underload and overload. The following algorithm will use the numeric value of the given mean load sharing edge formula that is based upon the following parameters $(m, n, \lambda_i, \mu_i, \lambda_j, \mu_j, L, q_\beta)$.

3.3 Algorithm

This section introduces the proposed load balancing algorithm based upon the above analysis. Periodically every T seconds, each overloaded peer transfers the exceeds load to the underlaoded peers (i.e. sender-initiated algorithm). This algorithm imposes a β -quantile rule for transferring load. For each pair (λ, μ) a corresponding β -quantile should be determined while β must be taken 0.90 or large. The algorithm is shown in the following scenario:

1. The overloaded node S_i creates a suitable domain (group) D_i from neighbor nodes to the peer S_i . Each node blongs to D_i satisfies $D_i = \{S_j; P(FPT_i > - t_{ij}) \geq 0.90 \text{ and } i \neq j\}$. Where t_{ij} is the required communication delay to send a message form node S_i to S_j plus the required time to reply with load transfer of certain virtual server from S_i to S_j . Also, FPT_i is the first passage time of node S_i to tranfer from overloaded state to normal or underlaoded state. D_i is an ordered set with respect to the communication t_{ij} . It is implemented as an order linked list.
2. Thus, S_i sends a broadcast messages to all nodes belonging to the doamin D_i . Node S_i must receive a reply from all nodes belonging to D_i within the FPT_i time.
3. Node S_i selects an underlaoded node $S_j \in D_i$ where the mean load sharing edge $MLSE$ between S_i and S_j . if $q_\beta > \Delta$ then transfers load (virtual server) from S_i to S_j . Where pair λ and μ are given, Δ is the time needed to transfer load less than or equal to $A(S_j)$, β is 0.90 or large.
4. Repeate step 3 for each underlaoded node S_j belonging to D_i whenever FPT_i period doesn't run out yet.

```

Load_balance( $S_i, T$ )
// Every period  $T$  seconds  $S_i$  checks its load
// jumps above upper load  $U$ .
// It should do the following.
Create( $D_i, S_i$ ); // create domain of neighbors
While ( $q_\beta > 0$ ) do
    // repeat the following within a period  $q_\beta$ 
    Select  $S_j \in D_i$ ; // select from order set  $D_i$ 
     $D_i = D_i \setminus \{S_j\}$ 
    If ( $q_\beta > \Delta$ ) then transfer_load( $S_i, S_j$ ); }
}

transfer_load( $S_i, S_j$ )
{ If !(Overloaded) then return; //Sender-initiated
  If ( $S_i \rightarrow VSset.size > 1$ ) then
    Choose  $v \in S_i \rightarrow VSset$  such that:
    a. Transfer  $v$  to  $S_j$  will not overload  $S_j$ 
    b.  $v$  is the least loaded virtual server that will release
       overload.
    Failing that, let  $v$  be the most loaded VS.
    Return the virtual servers reassignment.
}

```

4 Evaluation

This paper implements an event-based simulation to evaluate the proposed load balancing algorithm. It uses several parameters as follows: default number of virtual servers per node (12), number of nodes (4096), system utilization (0.8), Object arrival rate (Poisson with mean arrival time 0.01 s), average number of objects (1 million), and periodic load balancing period ($T = 60$ s). This simulation evaluates the following metric. Load Movement Ratio (*LMR*), defined as the total movement cost incurred due to load balancing divided by the total cost of moving all objects in the system at once. In case the value of the *LMR* is 0.1, it infers that the balancer consumes about 10 % of its bandwidth to insert objects. The node arrival rate is modeled by a Poisson process, and the lifetime of a node is drawn from an exponential distribution. This simulation ran with two inter-arrival times 10 and 60 s, it fixes the steady-state number of nodes in the system to 4096 nodes.

Figure 2 plots the *LMR* metric as a function of system utilization, to study the load moved by the proposed load balancing algorithm as a fraction of the load moved by the underlying DHT due to node arrivals and departures. Figure 2 demonstrates that the load moved by the proposed load balancing algorithm is significantly smaller than the load moved by the underlying DHT especially for small system utilization. In addition, Fig. 2 shows that the *LMR* with node inter-arrival time 10 s is larger than with node inter-arrival time 60 s. Figure 3 verifies

Fig. 2 LMR versus system utilization with two node arrival times

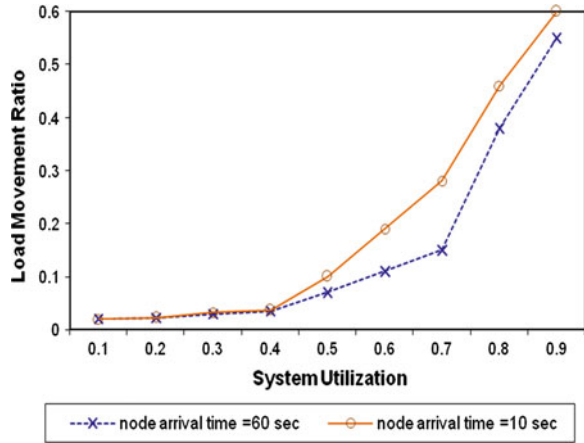


Fig. 3 LMR versus number of virtual servers with two node arrival

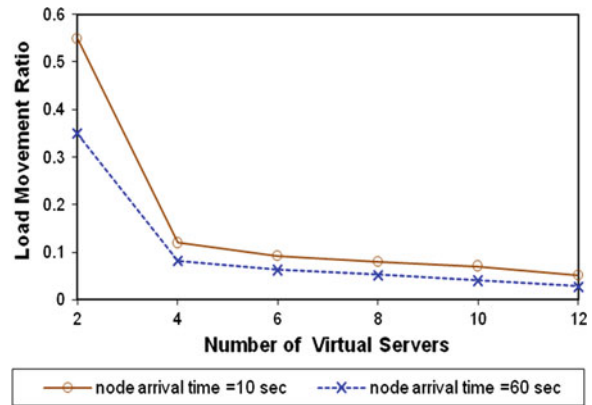


Fig. 4 Bandwidth lost versus system utilization with different number of virtual servers per node

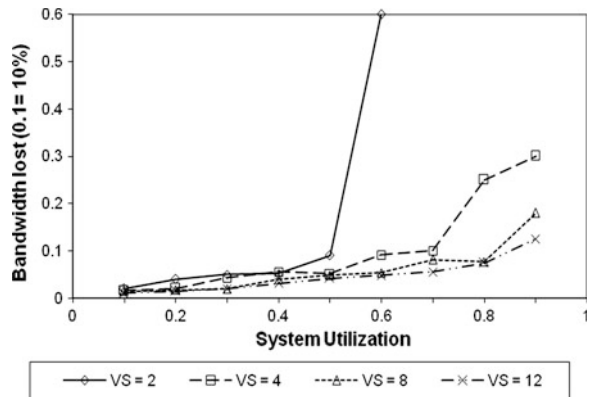
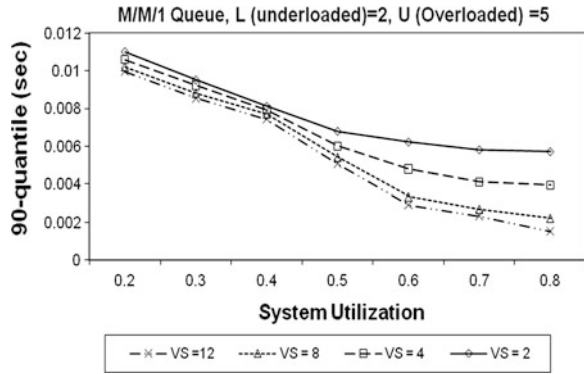


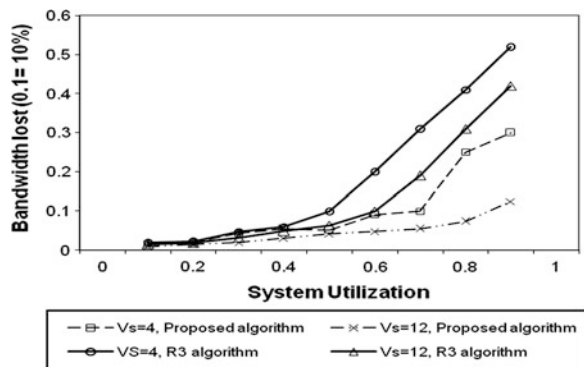
Fig. 5 90-quantile of the LSE versus system utilization with different number of virtual server per node



the perception that increasing the number of virtual servers decreases considerably the fraction of load moved by the underlying DHT. Figure 4 demonstrates that increasing number of virtual servers per node assists load balance at high system utilizations and grants efficient load movements due to low bandwidth lost. Figure 5 plots the 90-quantile of the load sharing edge (LSE) with system utilization when overload is 5 at source node and underload is 2 at the destination node. It demonstrates that the 90-quantile of the LSE tends to be smaller as system utilization increases. As seen from Fig. 5, the 90-quantile of the LSE is 9.94 ms thus load can be transferred only if $\Delta < 9.94$ ms. In addition, increasing the number of virtual servers reduces significantly the 90-quantile that helps in avoiding load thrashing.

In this paper we compare our results with simple load balancing algorithms such as *Random/Round Robin (R3) load distribution algorithm* [26]. The *R3* algorithm pushes load from an overloaded virtual server to a randomly or in a round robin fashion chosen neighbor that may absorb that load if it has the capacity, or pushes the load further on to another virtual server chosen in the same fashion. The advantages of the *R3* algorithm in compare with the proposed algorithm are as follows: simplicity and statelessness. However, the disadvantages of the *R3* algorithm are as follows: unpredictability and insufficient (random)

Fig. 6 Compare proposed algorithm with R3 algorithm



convergence on the chance for load thrashing. Figure 6 shows that the proposed algorithm is more efficient than the *R3* algorithm due to load thrashing in *R3* that increases the bandwidth lost.

5 Conclusion

Load balancing among peers is critical and a key challenge in peer-to-peer systems. This paper demonstrates a stochastic analysis that avoids the load thrashing and tackles the out-of-date problems due to peer's state changes during load movement (virtual servers migration). Then, it proposes a load balancing algorithm based on the stochastic analysis. An efficient simulation has been carried out that demonstrates the effectiveness of the proposed load balancing algorithm. Further research is to design a P2P load-balancing algorithm based on fuzzy logic control

References

1. Shirky C (2000) Modern P2P definition. <http://www.openp2p.com/pub/a/p2p/2000/11/24/shirky1-whatisp2p.html>
2. Saroiu S et al (2002) A measurement study of peer-to-peer sharing systems. In: Proceedings multimedia computing and networking conf (MMCN)
3. Sotica I et al (2001) Chord: a scalable peer-to-peer lookup service for internet applications. In: ACM SIGCOMM'01 pp 149–160
4. Rowstron A, Druschel P (2001) Pastry: Scalable distributed object location and routing for large-scale peer-to-peer systems. In Proceedings middleware
5. Ratnasamy S et al (2001) A scalable content- addressable network. In Proceedings ACM SIGCOMM'01, California
6. Derk I et al (1986) Adaptive load sharing in homogenous distributed systems. IEEE Trans on Soft Eng 12(5)
7. Bharambe AR et al (2004) Mercury: supporting scalable multi-attribute range queries. In Proceedings of the conference on applications, technologies, architectures, and protocols for computer communication. ACM, New York
8. Dabek F et al (2001) Wide-area cooperative storage with CFS. In: Proceedings of the 18th ACM Symposium on Operating Systems Principles (SOSP'01), pp 2020-2215
9. Godfrey et al (2004) Load balancing in dynamic structured P2P systems. In: Proceedings IEEE INFOCOM
10. Byers J (2003) Simple load balancing for distributed hash table. In: Proceedings of the second international workshop on peer-to-peer systems (IPTPS'03)
11. Rao A et al (2003) Load balancing in structured P2P systems. In Proceedings Of the second international . Workshop on peer-to-peer systems (IPTPS'03)
12. Godfrey PB, Stoica I (2005) Heterogeneity and load balance in distributed hash table. In: Proceedings IEEE INFOCOM
13. Li J, Kameda H (1994) A decomposition algorithm for optimal static load balancing in Tree hierarchy network configurations. IEEE Trans on parallel and distributed systems, 5(5)

14. Casavant TL, Kuhl JG (1988) A taxonomy of scheduling in general-purpose distributed computing systems. *IEEE Trans Softw Eng* 14(2):141–154
15. Shivaratri NG et al (1992) Load distributing for locally distributed systems. *Computer* 25(12):33–44
16. Goscinski A (1991) *Distributed operating system: the logical design*, Addison-Wesley
17. Zhou S (1988) A Trace-driven simulation study of dynamic load balancing. *IEEE Trans Softw Eng* 14(9):1327–1341
18. Eager DL et al (1985) A comparison of receiver-initiated and sender-initiated adaptive load sharing. *SIGMETRICS Perform Eval Rev* 13:2
19. Kruger P, Shivaratri NG (1994) Adaptive location policies for global schedule. *IEEE Soft Eng* 20:432–443
20. Zhou S et al (1993) Utopia: a load sharing facility for large, heterogeneous distributed computer systems. *Softw Pract Experience* 1305–1336
21. Dandamudi SP, Lo KC (1997) A hierarchical load sharing policy for distributed systems. In: *Proceedings of the 5th International. Workshop on modeling, analysis, and simulation of computer and telecommunications systems, MASCOTS*. IEEE CS, Washington, DC
22. Heyman DP (1982) *Stochastic models in operation research*, vol I. MxGraw-Hill Inc, New York
23. Kobayshi H (1978) *Modeling and analysis: an introduction to system performance evaluation methodology*. Addison-Wesley
24. Hisashi kobayshi et al (2009) *System modeling and analysis: foundations of system performance evaluation*. Prentice Hall
25. Zhu Y Hu Y (2005) Efficient proximity-aware load balancing for DHT-based P2P systems. *IEEE Trans On Parallel Distributed Syst*, 16(4) 349–361
26. Andrzejak A, Graupner S, Kotov V, Trinks H (2007) Algorithms for self-organization and adaptive service placement in dynamic distributed systems. *Internet systems and storage laboratory*

Unified Performance Authoring Tool for Heterogeneous Devices and Protocols

Youngjae Kim, Seungwon Oh, Dongwook Lee, Haewook Choi,
Jinsul Kim and Minsoo Hahn

Abstract A contents authoring software with device management for interactive shows and performances is proposed. The proposed system allows user to register and to manage various stage-related devices such as lighting system, fog generator, robots or multimedia players. In this paper, we describe software architecture, XML specification for unified protocol description, and user interface of the proposed system. The system can control multiple devices of heterogeneous protocols in timely manner. Once devices are registered, user can browse commands of each device and can organize commands to step-based list according to the performance scenario. Unlike current commercialized products, the system allows to manipulate authored list of commands during the show instantly.

Y. Kim · S. Oh · D. Lee
Digital Media Lab, Korea Advanced Institute of Science and Technology,
F309 KAIST, Moonji-dong, Daejeon, Republic of Korea
e-mail: yj_kim@kaist.ac.kr

S. Oh
e-mail: chiefvictoroh@kaist.ac.kr

D. Lee
e-mail: aalee@kaist.ac.kr

J. Kim
School of Electronics and Computer Engineering, Chonnam National
University, Gwangju, Republic of Korea
e-mail: jsworld@jnu.ac.kr

H. Choi · M. Hahn (✉)
Department of Electrical Engineering, Korea Advanced Institute of Science
and Technology, F309 Moonji-dong, Daejeon, Republic of Korea
e-mail: mshahn@ee.kaist.ac.kr

H. Choi
e-mail: hwchoi2@ee.kaist.ac.kr

The proposed system also supports to control playback of authored commands by sensor inputs so that a player can trigger devices to execute according to events. The contributions of this paper can be summarized to three points; (1) define unified protocol description with XML structures for simplicity (2) give sensor triggering capability to control devices, and (3) implement fast and instant manipulation user interface during the show.

Keywords Show controller · Controller · User interface · Multimedia

1 Introduction

As stages and scale of shows become larger and larger for a decade, both show directors and audiences' expectations also become higher than past. Show directors try to give impression to audiences by utilizing lots of projectors, lighting system, and bold audio effects. Show controllers are developed along with remote control capability and raised requirements of modern shows and performances [1]. Such devices are equipped with additional remote controllers or control ports using commonly known TCP/IP or MIDI protocols in order to operate with other devices. To control devices efficiently, the show control system is developed as a centralized multi-device control system for shows and performances. The main objective of show control system is to manage and to control various devices in given time as a show plays [2]. Early show control system was provided with guaranteed hardware by vendors, however, computer performance and stability are enhanced so that only software installation can give enough capability to run the shows. As the show control system is widely adopted, professionally organized shows are not limited to Olympic Games or national ceremony. They can be performed in small or mid-size stages, or outdoor playing environment with limited resources. For this reason, various scales and forms of shows are played so that new requirements are raised in show control scheme. For example, custom-built electronic sensor system are used to interact with players or audiences in order to support adlib (unplanned and improvised progress of play) or reactive performances. A new requirements are raised because current show control system are hard to provide modern experimental shows.

General show control system consist of three components; device management, scenario organization, and playback control. Device management component registers and manages various devices to the show control system. It checks status of devices in order to control them at any time without trouble during the show. The challenge of device management is to support multiple protocols. Currently, more than seven protocols are commonly used to run a basic show [3]; MIDI, DMX512, Open Sound Control (OSC), Ethernet-based TCP/IP (both UDP and TCP), Time Code, MIDI Show Control (MSC), generic RS232-based serial communication, and etc. Such protocols are not unified yet, because they are

specialized to control their own characteristics and requirements. For example, the DMX512 protocol is based on the RS485 protocol which has bus topology. The DMX512 protocol is specialized in lighting system control because lighting system is required to organize in multiple groups and to control them in a synchronous manner [3]. As the protocol is vary, the complexity becomes higher when protocols are combined or stacked with another protocol. Suppose that three beam projectors are required to display the multimedia. Generic RS-232 protocols are widely used to control beam projectors, but the communication distance is unreachable with the protocol. In this reason, RS-232 is normally used with ‘RS-232 on Ethernet’ communication transformer which can send and receive RS-232 data through TCP protocol. Likewise, new protocol has been proposed [4, 5] and widely adopted as *de facto standard* such as the Art-Net [6] protocol which transfer DMX512 signals on the UDP protocol. In summary, Device management component is required to handle multiple protocols and, moreover, to manage combined protocols. The scenario organization component manages customizable list of commands, so called cue, which contain commands of registered devices along with sequential order of the show. Two types of organization approach are used; step-based and timeline-based. Step-based approach transmits commands in the cue with triggers. The trigger can be activated by a show director’s signal manually or by other commands in the cue as a chain-reaction programmatically. Timeline-based approach is a kind of automatic trigger which is activated as time flows. In general, step-based approach is dominantly used and timeline-based approach is substantially placed only for complex and time-sensitive controls. Lastly, the playback control component of show control system is a runtime control of step-based or timeline-based organized cues.

Based on the show control system commonalities, the proposed system simplifies protocol architectures and user interfaces. We found a limitation of the current show control system that the organized cue cannot be modified during the show. There are two reasons of the restriction; (1) The show control system is too complex to manage various devices so that the cue needs to be compiled before running. (2) Traditional shows do not need to be changed their progresses during the shows, so that the restriction is acceptable in general. But as we described earlier, such a restriction support modern requirements of shows. In this paper, the proposed system allows user to edit and to reorganize commands during the show on the fly. If we can solve protocol complexity, a sensor-triggering capability may take advantage because customized sensors can be easily connected to the system. In our work, we propose new show control system which has simplified and unified XML protocol specification in order to support various combination of protocols, customized sensors and playback control by their triggering signals, and easy-to-use interface in order to allow fast and instant manipulation to user.

The organization of the remainder of this work is as follows. [Chap. 2](#) describes some design consideration and architectures for building unified show control system. [Chap. 3](#) shows implementation detail and its result. Finally, we conclude the work in [Chap. 4](#).

2 Design

2.1 Architecture

Our goal is to provide simplified device management, sensor-triggered control, and intuitive user interface to user. Based on the general show control system that have device management, scenario organization, and playback control as we described in Chap. 1 , we design the architecture of the proposed system to achieve the goal. In device management, the simplicity is achieved by allowing the protocol not to be stacked. In other words, our system does not allow combination of the device connectivity in terms of protocol description. If the RS-232 protocol is used by both native and UDP-to-RS-232 scheme, the protocol descriptions are not treated as identical. Figure 1 describes overall architecture of the proposed system.

As we can see in Fig. 1, the timer runs periodically to check the cue and each cue is triggered to run with its programmed time. When a command in the cue is executed, the Device Manager interprets the command to the device byte signal according to the device XML information. The term packetize stands for the process of the conversion of human readable command into the device byte signal (packet). And finally, the registered devices which established the connection with device can send to the device. The drawback of this architecture is that, the timer has an error so that the cue can be run inaccurately. The sensor input data is interpreted by the Expression Interpreter for sensor-triggering capability. The sensor data can be simple zero-or-one signal, or analog level signal. Once the signal is given, the Interpreter executes to user defined commands.

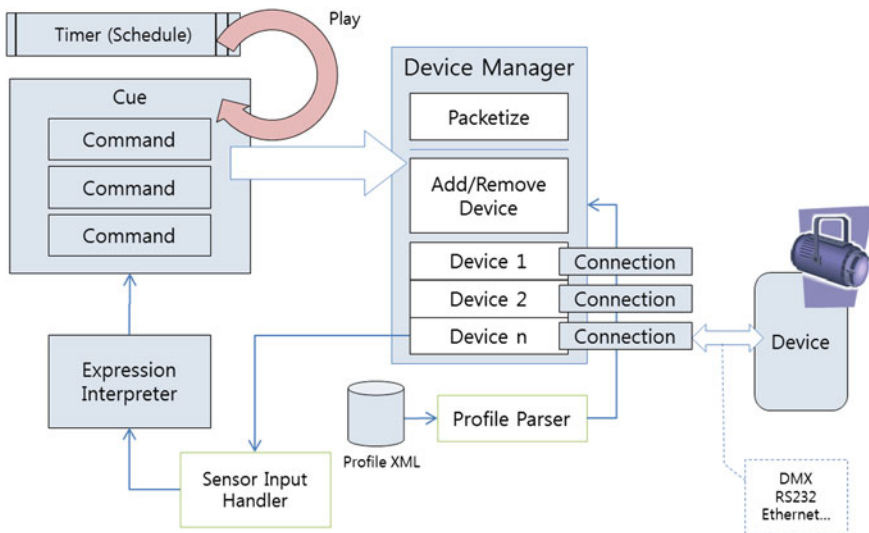


Fig. 1 Block diagram of the proposed system

2.2 Unified Device XML Specification

The proposed system stores each device information including byte-level communication specification in XML format. The proposed XML specification has four parts; device information, protocol description, command list, and the optional converter information. Figure 2 below shows an example of the proposed unified XML device protocols. The *Command* property groups byte signals and the *PacketWord* property has a series of two bytes packets. The XML description consists of three parts; general information, protocol description and commands. General information contains model name, vendor, and unique identifier (GUID). The protocol description contains IP address and port information. The commands can have both receiving and transmit signal data. In other words, it can contain sensor and actuator information in a same structure [7]. The converter is useful when the input or output data gives converted data. For example, if the sensor sends the data with hexadecimal system, the converter helps to convert the data into decimal system which gives more human readable information.

```
<?xml version="1.0" encoding="utf-8" ?>
<DeviceProfile GUID="8836dc79-f550-4deb-9d89-786710f3799a" Name="Sanyo PLC Projector X460" Desc="프로젝터 제어" Version="1.0">
  <Protocols>
    <Protocol Type="TCP">
      <Data Key="IP" Value="127.0.0.1" />
      <Data Key="Port" Value="5000" />
      <Data Key="BindIP" Value="5001" />
    </Protocol>
    <Protocol Type="RS232">
      <Data Key="Port" Value="COM1" />
      <Data Key="BaudRate" Value="9600" />
      <Data Key="FlowControl" Value="" />
    </Protocol>
    <Protocol Type="IP+RS232">
    </Protocol>
  </Protocols>
  <CommandList>
    <Command Name="Power On" Desc="전원 켜기" Direction="TX" ID="0" ReferenceID="">
      <Packet Type="End">02</Packet>
      <Packet Type="Command">4A</Packet>
      <Packet Type="Splitter">01</Packet>
      <Packet Type="Parameter" Start="0" End="255" Converter="byte_to_hex">33</Packet>
      <Packet Type="End" Converter="command+parameter+1">33</Packet>
      <Packet Type="End">13</Packet>
    </Command>
    <Command Name="Volume Check" Desc="볼륨확인" Direction="TX" ID="1" ReferenceID="2">
      <PacketWord>433031</PacketWord>
    </Command>
    <Command Name="Volume CheckIn" Desc="현재볼륨" Direction="RX" ID="2" ReferenceID="1">
      <PacketWord>433031</PacketWord>
    </Command>
    <Command Name="Set Volume" Desc="볼륨설정" Direction="TX" ID="3" ReferenceID="">
      <PacketWord>433031</PacketWord>
    </Command>
  </CommandList>
  <Converters>
    <Converter Name="byte_to_hex" Data="..." />
    <Converter Name="command+parameter+1" Data="..." />
  </Converters>
</DeviceProfile>
```

Fig. 2 An example of XML device description

3 Implementation

We implement a software which utilize the proposed XML device description and gives fast and easy manipulation to user. Figure 3 is the screenshot of the implemented software. The software is implemented by C# 4.0 and Windows. One of the goal of the user interface is to give fast manipulation time. As a result, we use drag-and-drop manipulation scheme as many as possible to the software. User can drag a command to the cue to add the command, or user can drag command list in the cue up and down to reorganize. Several drop-zones are placed to add or to register various data.

As we can see Fig. 3, the left side of the screen is a view of device manager which displays registered devices and their commands. The right side of the screen is cue lists which is organized as a play directed. The bottom of the screen displays logs of the execution of the software. The right-bottom of the screen has the *quickslot* to add a command or a set of commands to execute with a single click. And there is a drop-zone at the bottom of the list of the cue. The user can drag a command and can drop it to add a cue.



Fig. 3 Screenshot of the implemented system

4 Conclusion

The proposed system aims to unify various show related protocols into single XML specification. The system utilize the specification by controlling shows, performances, and various devices. The proposed system defines XML specification in order to unify various stage-related device protocols. The system gives sensor-triggering capability enables to control devices with reactive manner and the user interface which allows to manipulate organized commands list during the show. We plan to develop the implementation software more stable manner in order to deploy to real world examples.

References

1. Miller MR (2010) Show design and control system for live theater. Massachusetts Institute of Technology, Cambridge
2. Miller MR (2010) Show design and control system for live theater, Massachusetts Institute of Technology, Cambridge, p 75
3. Ivanov IA, Plakhov AG (2007) Wireless control of city beautification dynamic lighting systems. Light and Engineering, Znack Publishing House, Moscow, p 63
4. Gang D, Chockler GV, Anker T, Kremer A, Winkler T (1997) TRANSmidi: a system for midi sessions over the network using transis. Citeseer, pp 283–286
5. Goto M, Neyama R, Muraoka Y (1997) RMCP: remote music control protocol. pp 446–449
6. Newton S (2005) Art-net and wireless routers. IEEE, pp 857–861
7. Kim Y, Shin H, Hahn M (2009) A bidirectional haptic communication framework and an authoring tool for an instant messenger. 11th international conference on advanced communication technology. ICACT 2009. IEEE, 2009, pp 2050–2053

Discrimination System of Surround Illumination Using Photo Sensor's Output Voltage Ratio

Eun Su Kim, Hee Dong Park and Kyung Nam Park

Abstract In this paper, we implemented discriminating system for various surround illuminants using photo sensor. To discriminate surround illuminants of display device we implemented discriminating system using photo sensor which has output Y_e and C_y . Experimental results shows that we could discriminate surround illuminants effectively by using the output voltage ratio (Y_e/C_y) of photo sensor in the varying luminous intensity of surround illuminants.

Keywords Surround illuminant · Discriminating system · Photo sensor

1 Introduction

There has been researched about color reproduction for standard visual display devices like TV. One method for color reproduction is using colorimetric color reproduction which is using ratios of the stimulus intensity and the other method is

E. S. Kim
School of Electron Engineering, Sunmoon University, 5-4, Kalsan-ri,
Tangjeong-myeon, Asan-si, Chungnam, South Korea
e-mail: eunsu.kim@sunmoon.ac.kr

H. D. Park
Department of Information and Communication, Korea Nazarene University,
456, Ssangyong-dong, Seobuk-gu, Cheonan-si, Chungnam, South Korea
e-mail: hdpark@kornu.ac.kr

K. N. Park (✉)
Department of Multimedia, Korea Nazarene University, 456, Ssangyong-dong,
Seobuk-gu, Cheonan-si, Chungnam, South Korea
e-mail: knpark@kornu.ac.kr

preference color reproduction to the specific color such as flesh or leaf color. But the environments which use the real color devices are probably different from the standard visual environment. When the color display is used in the home and office, natural light is used in the daytime, but natural light is short of or during the night time artificial lights such as incandescent light and fluorescent light are used.

In this difference of visual environment situations, chromatic adaptation is happened in human visual system because of the difference cone cell's gain [3–9]. And three kinds of cone cells, L(long), M(middle), and S(short), have maximum sensitivity about 575, 330, 455 nm respectively. This chromatic adaptation phenomenon causes that user's feeling of color changes during watching the display device. So in the case of visual environment is different from the standard's one, reproduction color should be the same with the original subject's color.

There are several models [3–10] which reproduced color in consideration by the chromatic adaptation. Von Kries [4] and Breneman [9] supposed that human visual coefficient ratio of the color adaption is linear from the specific illumination to the other illumination. And Bartleson [8], Fairchild [7], CIECAM97 s [10], and modified von Kries Model [11] used nonlinear adaptation, namely, these methods supposed that stimulus changing rate is nonlinear in human visual. In order to implement these models, we have to discriminate surround illuminants. So in this paper, we proposed the method of discriminating surround illuminant using photo sensors.

2 Transformation Relation from Bi-Stimulus of Ye and Cy to Tri-Stimulus of X, Y, Z of CIE for Discriminating System of Surround Illuminants

In this paper, We used two output photo sensors, Ye(yellow) and Cy(cyan) to discriminate surround illuminants. Figure 1 and Table 1 shows that relative sensitivity of Ye and Cy's to the wavelength and x and y color coordinate of CIE1931 respectively.

To measure the color temperature using this photo sensors, we have to obtain x and y coordinate of input lights. To obtain the coordinate, Transformation from

Fig. 1 Relative sensitivity of photo sensor's wavelength

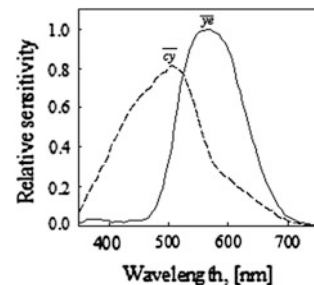


Table 1 CIE 1931 and x and y coordinate of photo sensor's Ye and Cy

	Ye		Cy	
	x	y	x	y
CIE 1931	0.4446	0.5498	0.2186	0.3528
AM-32-CY-02	0.3810	0.6122	0.2270	0.4085

bi-stimulus of Ye and Cy to the tri-stimulus of X, Y, and Z of CIE is shown in Eq. (1).

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = M \begin{bmatrix} Ye \\ Cy \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} Ye \\ Cy \end{bmatrix} \tag{1}$$

Here, wavelength is obtained by sampling of 5 nm interval and wavelength range is from 280 nm to 750 nm.

From Eq. (1), First row represented again by matrix format. Then Eq. (2) is obtained. Eq. (2) have many equations and unknown quantity is fewer, So we can find out the unknown quantity which has minimum error by generalized inverse matrix.

$$\begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \bar{y}_{c1} & \bar{c}_{y1} \\ \bar{y}_{c2} & \bar{c}_{y2} \\ \vdots & \vdots \\ \bar{y}_{cn} & \bar{c}_{yn} \end{bmatrix} \begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} \tag{2}$$

So, unknown quantity a11 and a12 is shown like Eq. (3).

$$\begin{bmatrix} a_{11} \\ a_{12} \end{bmatrix} = \left(\begin{bmatrix} \bar{y}_{c1} & \bar{c}_{y1} \\ \bar{y}_{c2} & \bar{c}_{y2} \\ \vdots & \vdots \\ \bar{y}_{cn} & \bar{c}_{yn} \end{bmatrix}^T \begin{bmatrix} \bar{y}_{c1} & \bar{c}_{y1} \\ \bar{y}_{c2} & \bar{c}_{y2} \\ \vdots & \vdots \\ \bar{y}_{cn} & \bar{c}_{yn} \end{bmatrix} \right)^{-1} \begin{bmatrix} \bar{y}_{c1} & \bar{c}_{y1} \\ \bar{y}_{c2} & \bar{c}_{y2} \\ \vdots & \vdots \\ \bar{y}_{cn} & \bar{c}_{yn} \end{bmatrix}^T \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{3}$$

Like this way, we could find out the transformation form from bistimulus of Ye and Cy to tristimulus of CIE's X, Y, and Z.

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = M \begin{bmatrix} Ye \\ Cy \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} Ye \\ Cy \end{bmatrix}, \quad M = \begin{bmatrix} 0.7619 & -0.0623 \\ 0.8111 & 0.0845 \\ -0.6667 & 1.3505 \end{bmatrix} \tag{4}$$

If apply Eq. (5) to the Eq. (4) and divide both sides by stimulus Y, then Eq. (6) is obtained.

$$x = \frac{X}{X + Y + Z}, \quad y = \frac{Y}{X + Y + Z}, \quad z = \frac{Z}{X + Y + Z} \tag{5}$$

$$\begin{bmatrix} X/Y \\ Y/Y \\ Z/Y \end{bmatrix} = \begin{bmatrix} x/y \\ 1 \\ z/y \end{bmatrix} = \frac{1}{Y} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} Ye \\ Cy \end{bmatrix} \quad (6)$$

And get the first and third row from the Eq. (6) and if perform inverse transform then Eq. (7) could be obtained.

$$\begin{bmatrix} Ye \\ Cy \end{bmatrix} = Y \begin{bmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{bmatrix}^{-1} \begin{bmatrix} x/y \\ z/y \end{bmatrix} \quad (7)$$

Here, stimulus Y could be as Eq. (8).

$$Y = a_{21}Ye + a_{22}Cy \quad (8)$$

If we represent the inverse matrix of Eq. (7) as Eq. (9), then Eq. (7) could be changed to Eq. (10).

$$\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{bmatrix}^{-1} \quad (9)$$

$$\begin{bmatrix} Ye \\ Cy \end{bmatrix} = (a_{21}Ye + a_{22}Cy) \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} x/y \\ z/y \end{bmatrix} \quad (10)$$

So the relation of the photo sensor's output voltage Ye and Cy could be described as Eq. (11).

$$Ye = \frac{b_{11}x + b_{12}z}{b_{21}x + b_{22}z} \cdot Cy \quad (11)$$

For example, Using the Eq. (11) we could find out the photo sensor's output voltage ratio of Ye and Cy as Table 3. And Fig. 2 shows that Ye's output voltage relation to the Cy.

And we could also find out that illumination's coordinates using Eqs. (4) and (5). Table 2 shows display white and various illuminants. In home incandescent lamp and fluorescent lamp are used as illumination, In order to apply TV and PC monitor, we have to classify the illuminants rather than measurement of color temperature. And if we get illuminant's x and y coordinates by Eq. (11), then we classify the illuminants by the photo sensor's output voltage ratio. Namely, if the ratio between Ye output voltage and Cy output voltage is 1.5, then we classify as incandescent lamp, and the ratio is 1.0, then we classify as fluorescent lamp as in Fig. 2 (Table 3).

Table 2 Display white and illuminants

Display white and illuminants	Chromaticity, W_D or W_L			
	u	v	x	y
D_{65}	0.1978	0.3122	0.3127	0.3290
9300 K + 27 MPCD	0.1822	0.3024	0.2810	0.3110
Incandescent lamp	0.2560	0.3495	0.4476	0.4075
Fluorescent lamp	0.1675	0.3360	0.3060	0.4100

Fig. 2 Y_e and C_y 's output voltage of illuminants

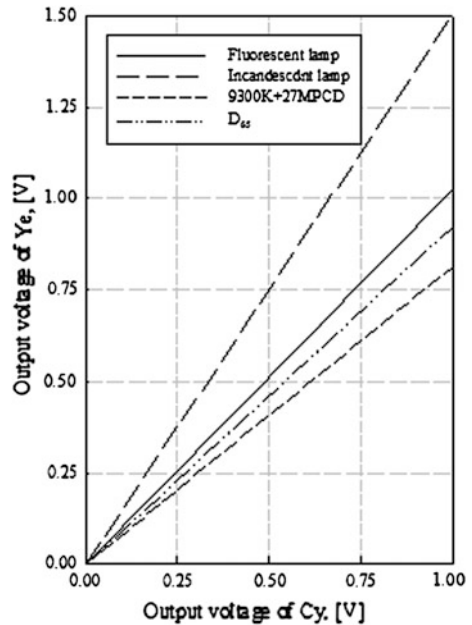


Table 3 Photo sensor's Y_e/C_y of display white and illuminants

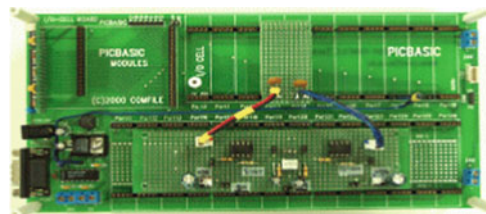
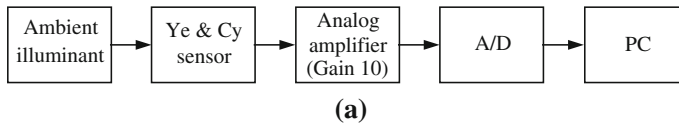
Display white and illuminants	Y_e/C_y
D_{65}	0.924
9300 K + 27 MPCD	0.813
Incandescent lamp	1.501
Fluorescent lamp	1.025

3 Implementation of Discriminating System of Surround Illuminant

To implement the discriminating system of surround illuminant, we experiments with the lights which is installed in viewing box (The Judge 2 model) of GretagMacbeth. In this viewing box A, CWF, D50, D65, and D75 light is possible by

Table 4 Color coordinates and color temperature of viewing box

Illuminant	x	y	Color temp.	Model
A	0.453	0.412	2500 K + 0MPCD	Incandescent lamp
CWF	0.376	0.407	4300 K + 30MPCD	Cool White F20T12/CW
D ₅₀	0.341	0.370	5200 K + 20MPCD	5000 F20T12/50
D ₆₅	0.308	0.336	6700 K + 18MPCD	6500 F20T12/65
D ₇₅	0.300	0.317	7400 K + 7MPCD	7500 F20T12/75

**Fig. 3** a block diagram of discriminating system of surround illuminant and b hardware

pre-installed incandescent lamp and fluorescent lamp and its inner wall is made of non gloss material so as to maintain the characteristic uniformly in the box. We measured light's characteristic and photo sensor's real operation characteristic using chroma meter (MINOLA CS-100, CL-100), LCD color analyzer (MINOLTA CA-210), and digital oscilloscope. Table 4 shows the measured color coordinates and color temperatures.

And Fig. 3 show implemented system's block diagram and its hardware. By the photo sensor Ye and Cy's output voltage is generated, and these output voltage was amplified 10 times in consideration of A/D converter's full range and adaptation for hardware.

And amplified output voltages converted 8 bit digital value which is transferred to PC using RS-232c communication.

4 Experimental Results

Discriminating system of surround illuminant is implemented as hardware and measured photo sensor's output voltage for the viewing box's illuminants. Figure 4 shows photo sensors output voltage characteristic in same condition. But luminosity

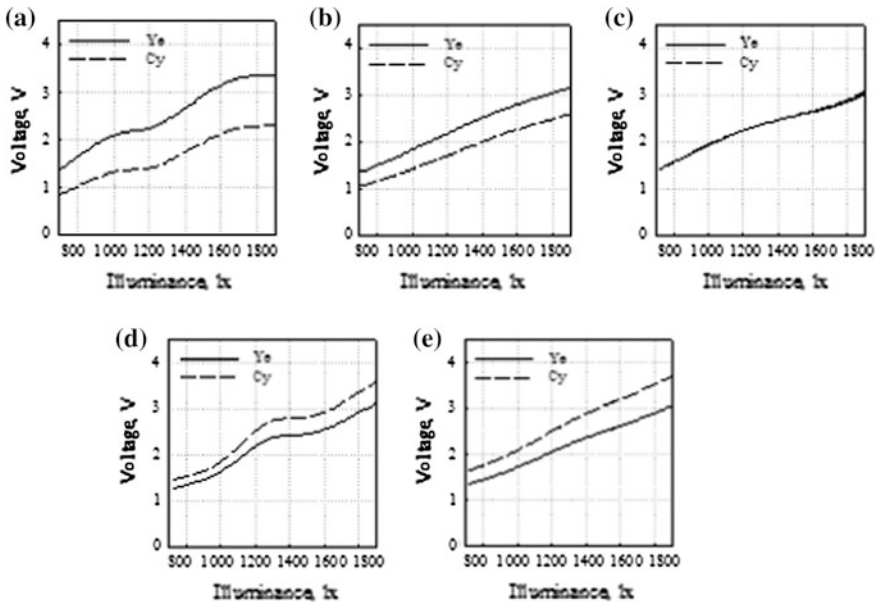


Fig. 4 Photo sensor’s output voltage of **a** A, **b** CWF, **c** D50, **d** D65, and **e** D75

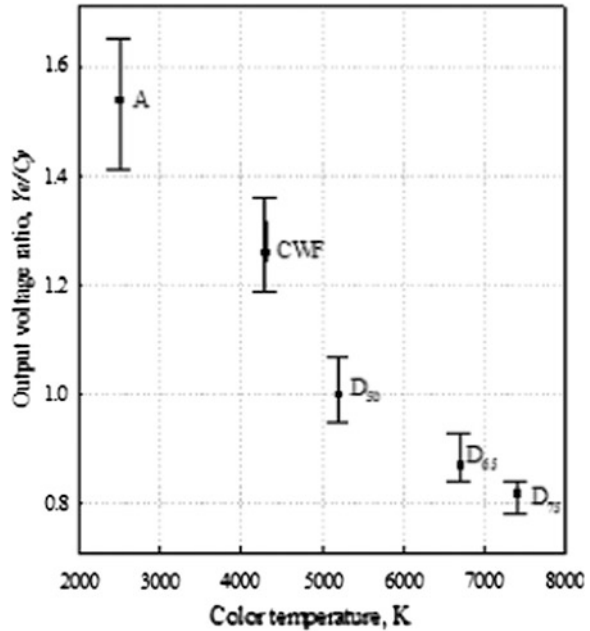
Table 5 Photo sensor’s output voltage ratio of various illuminants

Illuminant	Ye/Cy (before amplification)			Ye/Cy (after amplification)		
	Max.	Min.	Avg.	Max.	Min.	Avg.
A	1.63	1.47	1.54	1.61	1.46	1.54
CWF	1.28	1.19	1.24	1.30	1.22	1.26
D ₅₀	0.99	0.96	0.97	1.00	0.98	1.00
D ₆₅	0.86	0.84	0.85	0.88	0.86	0.87
D ₇₅	0.82	0.79	0.80	0.83	0.82	0.82

varies 650, 950, 1250, 1550, and 2000 lx respectively. As you see in Fig. 4, the output voltage ratio is constant regardless of luminosity.

And Table 5 shows measured maximum, minimum, and average value of the photo sensor output voltage ratio for each illuminants. In the case of A light, photo sensor’s output voltage ratio is 1.54 and D65 is 0.86. So we could classify the surround illuminants according to the photo sensor’s output voltage ratios. In Fig. 5, we verified that illuminants have not overlapped the output voltage ratios and this means illuminants could be classified in spite of various luminosities.

Fig. 5 Output voltage ratios of discriminating system for the surround illuminants



5 Conclusions

In this paper we proposed discriminating method of surround illuminants using photo sensor's output voltage ratio which is precede color reproduction process. To determine the surround illuminant during watching display device, we use two photo sensors which generated output voltage of Ye (yellow) and Cy (cyan). And according to the experimental results by implemented system, we verified that photo sensor's output voltage ratio could classify various illuminants effectively regardless of varying luminosity. Various illuminants could be discriminated using proposed systems. Furthermore, this illuminant discriminating system could be used multiple applications for the illuminant discrimination area.

References

1. Michael Stokes and Matthew Anderson (1996) A standard default color space for the internet-sRGB.0 <http://www.w3.org/Graphics/Color/sRGB.html>
2. Hunt RWG (1987) The reproduction of colour in photography. Printing and Television, Fountain Press, England, pp 177–196
3. Wyszecki G, Stiles WS (1982) Color science. Wiley, New York, pp 117–451
4. MacAdam DL (1981) Color measurement. Springer, New York, pp 200–208
5. Bartleson CJ (1978) Comparison of chromatic-adaptation trans-FORMS. Color Res Appl 3:129–136

6. Bartleson CJ (1979) Changes in color appearance with variations in chromatic adaptation. *Color Res Appl* 4:119–138
7. Mark DF (1998) Color appearance models. Addison-Wesley, New York, pp 173–214
8. Bartleson CJ (1979) Predicting corresponding colors with changes in adaptation. *Color Res Appl* 4:143–155
9. Breneman Edwin J (1987) Corresponding chromaticities for different states of adaptation to complex visual fields. *J Opt Soc Am* 4:1115–1129
10. CIE TC1-34 Final Report (1998) The CIE 1997 interim colour appearance model (Simple Version). CIECAM97 s
11. Kim ES, Jang SW, Kwon YD, Han CH, Sohng KI (2004) Corresponding-color reproduction model according to surround viewing conditions. *IEICE Trans E87-A*:1514–1519
12. Video Electronics Standards Association (2001) Flat panel display measurements standard version 2.0, VESA, p 115

The Development of Korea: Computer Access Assessment System (K-CAAS) for Persons with Physical Disabilities

Jinsul Kim and Juhye Yook

Abstract The purpose of the study was to develop a computer access assessment system for improving computer access of individuals with physical disabilities for the first time in South Korea. Korea-Computer Access Assessment (K-CAAS) presented in this article tests three user skills in aim, drag, and menu. Skill levels of the three test areas could be set as primary, intermediate, proficient, and individualized according to individual's needs and conditions. All tests have their default set in each level, and the skill levels can be selected and changed by the abilities and goals of a user. Tests could be selected for users' training and, their results could be traced and shown. Therefore, the K-CAAS is a training program to improve computer access skills as well as an assessment program. It would help users with physical disabilities operate a computer by themselves as improving their computer access skills.

Keywords Physical disabilities · Computer access · Assistive technology assessment · Digital divide · Assessment of computer accessibility

J. Kim

Department of Electronics and Computer Engineering, Chonnam National University,
Gwangju 550-757, South Korea
e-mail: jsworld@jnu.ac.kr

J. Yook (✉)

Department of Rehabilitation Technology, Korea Nazarene University,
Chungnam 331-718, South Korea
e-mail: jhyook@kornu.ac.kr

1 Introduction

In Korea, 1,599,468 individuals with physical disabilities (63.5 %) were registered out of 2,517,312 persons with all types of disabilities in 2010. Those with physical disabilities anticipated to need assistive technology hardware and software for information access are 488,186 people (30.5 %) for their severe disability from level 1 (the severest) to level 3 [1]. There are 6 levels for disability severity by The Welfare for the Disabled Act in South Korea and level 6 is the mildest. These people were supported with Korean Me-Too Keyboard, Roller Trackballs, One-Hand User Keyboard, Head Z Mouse, Touch Monitor, Rakuraku Joystick, Ultimater 8, Keyguard, Pocket-Go-Talk, Rakuraku Mouse, etc. from the National Information Society Agency (NIA) or the Korea Employment Agency for the Disabled (KEAD) in 2007. About 35 % of those with physical disabilities said that they experienced manipulation difficulty and about 20 % of them reported problems in usage of supported assistive technology for information and communication access [2].

According to the 2011 Status Survey on Digital Divide, 23.2 % of people with cerebral palsy and 15.7 % of those with physical disabilities did not use Internet because of their disability condition. The most number of persons with visual disabilities, hearing and language disabilities, and cerebral palsy and physical disabilities replied that they did not use the Internet because of not knowing how to use it. Moreover, much fewer individuals with disabilities used smart phones and could access to mobile devices comparing to that the total population did in recent days [3].

Persons with physical disabilities need a training and education process based on accessibility assessment to use assistive technology for information and communication and for the activities of their lives. Then, the use of assistive technology should be analyzed if it results in positive effects later on [4]. However, Korea government and public agencies distribute assistive technology devices when individuals with disabilities apply for the devices without objective and systematic verification of needs and appropriateness. This does not secure the active use of devices received and the improvement of the quality of their lives [5].

First of all, interaction performance such as accuracy and speed between users and assistive technology devices for information access used or suggested needs to be measured objectively. These measurement tools and programs of assistive technology use for information access are not commercialized in South Korea yet. Therefore, this study was conducted to develop a computer access assessment system for improving computer access of individuals with physical disabilities to solve this issue. The software named Korea—Computer Access Assessment System (K-CAAS) was developed and would be distributed to improve the computer access and assessment of persons with physical disabilities for the first time in South Korea. K-CAAS could be used for computer training of people with physical disabilities as a rehabilitation program.

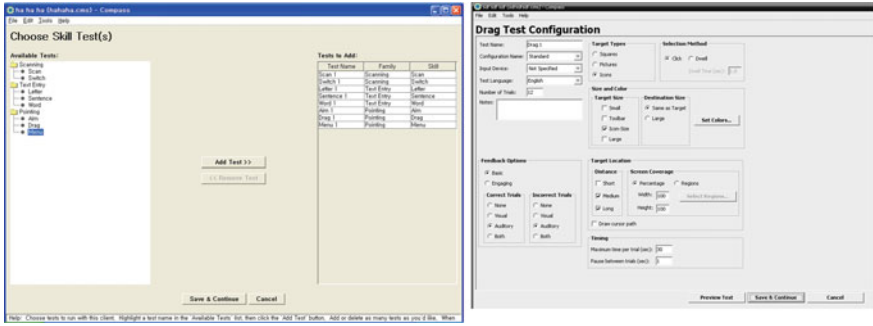


Fig. 1 Compass: ‘Choose skill test’ screen to choose test skills (left), ‘drag test configuration’ (right)

The article includes an analysis of existing evaluation programs for computer access in Chap. 2, design and implementation of K-CAAS in Chap. 3, development results in Chap. 4, and conclusion in Chap. 5.

2 An Analysis of Existing Evaluation Programs for Computer Access

2.1 Compass

Compass is computer access assessment software to measure functions of computer use in the areas of ‘Pointing Devices’, ‘Text Entry’, and ‘Scanning’ after individual users register and login (see in Fig. 1). ‘Pointing Devices’ measures the accuracy rates (percentages of correctness) and speed (response time in seconds) in subareas of ‘Aim’, ‘Drag’, and ‘Menu’. ‘Text Entry’ tests input accuracy rates and speed in subareas of ‘Letter’, ‘Word’, and ‘Sentence’ and ‘Scanning’ tests the skills in ‘Scan’ and ‘Switch’ [6]. The scope of the configuration options, however, is too extensive to select every time evaluators use the program. The results of the skills in each areas and in each sessions are not be analyzed in an integrated way while the results of the users’ skills are presented in graphs, tables, and user tracks on screen in each area and in each session.

2.2 EvaluWare

EvaluWare is an informal and computerized program for computer access evaluation (see Fig. 1). Users explore and perform activities in ‘Explore Your Looking Skills’, ‘Explore Your Listening Skills’, ‘Explore Your Motor Skills’, and



Fig. 2 EvaluWare main page

'Explore Related Skills' while an evaluator observes the users' skills. Functions to test can be individually customized and played like a fun game by setting up test activities in 'Building a Screen' options [7]. The limitation of the program would be that it does not measure or record the results of the assessment by itself and relies on evaluators' observation and recording (Fig. 2).

3 Module Design and Implementation of K-CAAS

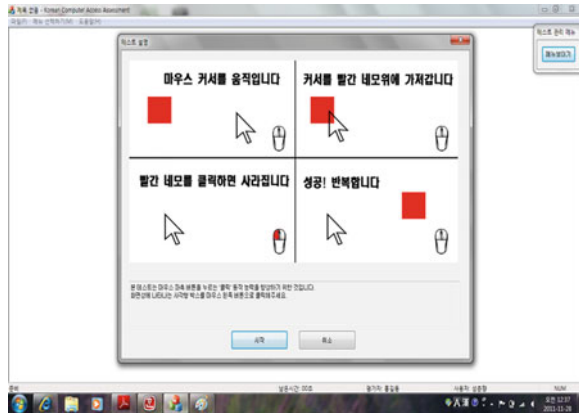
3.1 User-Registration Functions

Two types of login are for a new user and existing users. An evaluator registers her/his name and a user's name and selects test items for the user in the module of user registration functions. Existing users' previous test results saved in the system can be retrieved to analyze their improvement of computer accessibility by login as an existing user name.

3.2 Environmental Creation Functions

Environmental creation functions make possible that users take the test in the optimum conditions. First of all, test levels can be set as primary, intermediate, proficient, and individualized. Each level of primary, intermediate and proficient has its own preset test options in a number of trials, time limit, icon sizes, and scanning cycles. These options can be selected all differently according to the needs of the users for each session in the individualized level.

Fig. 3 Aim



3.3 Assessment List Set and Controls Functions

Then, test and subtest items can be chosen. There are ‘All choose’ and ‘All cancel’ buttons for 5 assessment items of ‘Aim’, ‘Drag’, and ‘Menu’. Each test item at a time also can be selected according to the users’ needs. In the assessments on the computer, users repeatedly operate and control a mouse to perform a particular task to show their abilities and skills with task directions. These repeated tasks in the assessments allow users with disabilities to improve accessibility to a computer and, the saved data are used as test results of their computer use skills.

3.4 User-Saving & Management Functions

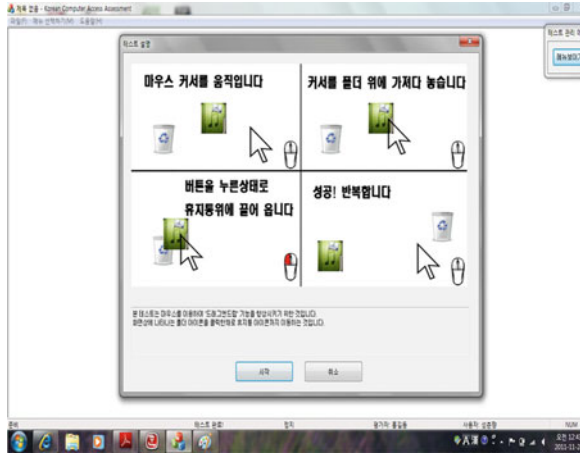
User-saving and management functions provide success rates, failure rates, and response time by each test item after completing a test. The results of the test indicate users’ ability of computer use and their improvements later when the results of repeated tests are stored in the system. The test results of each user can be printed by exporting function.

4 Development Results

4.1 Aim

The test, ‘Aim’, is to improve the left click ability of a mouse. A user moves a cursor and clicks on a target when it appears on a computer screen. The test is assessing and helping users to perform a mouse click proficiently (see Fig. 3).

Fig. 4 Drag



The scenario in detail is as the following. A user moves a mouse cursor and clicks on the left when a square target appears at a random spot on a computer for certain period of time. Then, the same performance of the user is required when a square target appears continually at other spots. The target images appear repeatedly until the test set is completed. ‘Failure’ message pops up and is counted as failure in the system when the user doesn’t click in time or clicks in error. ‘Failure’ or ‘Success’ messages call the user’s attention and thus, improve the accuracy when the user clicks a target.

4.2 Drag

The test, ‘Drag’, is to improve the ‘drag and drop’ ability of a mouse. This is to improve the ability to click and maintain until the clicked target icon is moved to an intended spot and, the click action is released finally. A target icon to move and the icon for destination (trash can icon) appear on screen. Then, the target icon is moved with a mouse while clicked until the mouse click is released on the icon for destination. A new target icon to move and a new icon for destination appear on screen when one task of dragging is completed. The tasks are repeated until icons do not appear anymore. It is counted as failure if the target icon is not moved to the icon for destination in time and then, new icons appear (see in Fig. 4).

4.3 Menu

The test, ‘Menu’ is to improve the ability of clicking a choice from a Windows menu bar on the top of the screen. The direction of a menu and a submenu that a user should click is presented on the center of the screen, and the user clicks the

Fig. 5 Menu

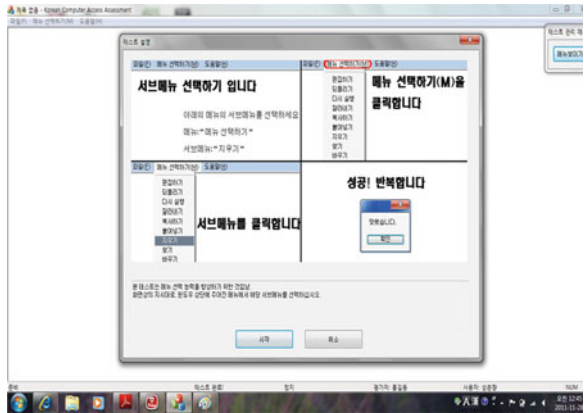
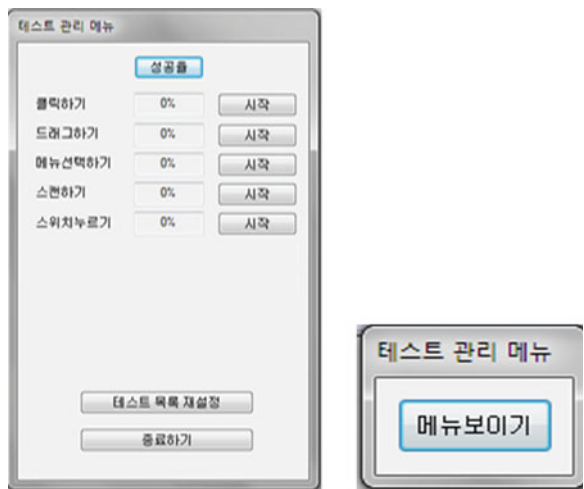


Fig. 6 Assessment: Success rates



menu and the submenu following the direction. A new direction of clicking a menu and a submenu is presented after the user clicks correctly, clicks incorrectly, or does not click in time. The test of clicking a menu and a submenu is repeated until the direction does not appear anymore (see in Fig. 5).

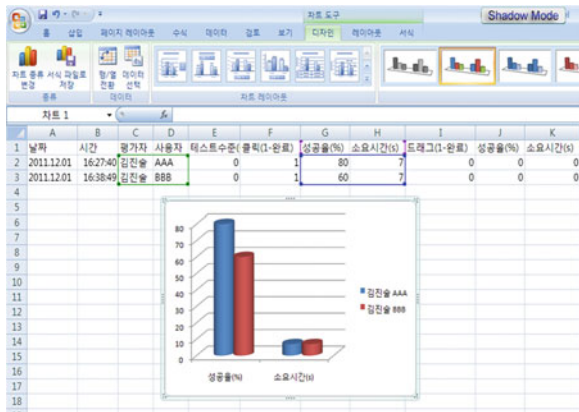
4.4 Assessment Use and Result Analysis

The assessment is designed to start and stop at any time and is able to add a new test item and restart in the middle of the test. The assessment is easy to control while performing by selecting a ‘Test Management’ button placing on the top-right of the screen (see in Fig. 6).

Fig. 7 Assessment: Failure rates (left) and response times (right)



Fig. 8 Assessment results in statistical description and graphs in excel file



Assessment results are presented as success rates, failure rates, and response times by clicking the buttons. The results indicate the ability of the user tested for computer access and can be analyzed to see the user’s progress (see in Fig. 7).

Assessment results can be exported to save as CVS file in text so they can be retrieved by various types of files later. Figure 8 shows the assessment results easily in statistical description and graphs to see the user’s ability improvement in excel file.

5 Conclusion

The study developed and implemented the Korea—Computer Access Assessment System (K-CAAS) for individuals with physical disabilities and, the system is intended to distribute for the first time in South Korea. The system is designed to

improve and assess computer accessibility in 3 areas for persons with physical disabilities. The program also can test in individualized format by the number of sessions, time limit, icon sizes, distance of cursor movements, and scanning cycles with 'Correct' or 'Incorrect' feedback. A user can choose the level of tests with preset tasks in each level and also can set the elements of the tests at her/his choice for the individual's ability and needs. The results can be tracked and analyzed statistically after the user completes the assessment. So, the K-CAAS is efficient to save assessment time as an assessment tool of computer use for individuals with physical disabilities and can be used as a rehabilitation program of improving their ability of computer operation skills at the same time.

The system will be upgraded as adding to test scroll skills and typing skills and thus, improved its quality. The system also will be converted into mobile format with smart devices so that persons with disabilities can use it as a rehabilitation program at anytime and anywhere in the future study.

Acknowledgments This work was supported by the Technology Innovation Program (100036459, Development of center to support QoLT industry and infrastructures) funded by MKE/KEIT, Korea, 2012.

References

1. Ministry of Health and Welfare, Statistics Portal. <http://stat.mw.gokr/>
2. Go D, Park K, Yook J, Yoo J (2007) A field study on access and use electronic & information technology assistive device of person with disabilities. *J Special Educ Theory Pract* 8(2):319–343
3. National Information Society Agency (2011) Status survey on digital divide 2012
4. Yook J (2011) Designing computer workstation for a university student with cerebral palsy. *J Rehabil Technol* 2(7):41–58
5. Park G, Yun H, Park S, Yook J (2011) Status of assistive technology devices use of teachers for students with physical disabilities. *J Rehabil Technol* 2(3):17–39
6. Koester HH (2004) Compass Koester performance research
7. Assistive Technology Inc. (2002) EvaluWare™: assessment activities for AAC and computer access

Color Coding for Massive Bicycle Trajectories

Dongwook Lee, Jinsul Kim, Haewook Choi and Minsoo Hahn

Abstract As the smartphone market grows, tracking a person's own positions become easier and popular. Especially for the bicycling, file based GPS data make it easier to manage and access personal trajectories. In this paper, we propose an effective color coding method for massive bicycle trajectories visualization. The motivation of the method is based on characteristics of the bicycle trajectories which have different spatial aspects compare to the automobiles. The proposed method modifies the color of the line segments based on the direction and flow, and provides visually enhanced trajectories. GPS data collected from Han riverside bicycle tracks were applied to the proposed visualization methods, and shown the potential possibilities for trajectory analysis.

Keywords Visualization · Bicycle trajectory · GPS

D. Lee (✉) · M. Hahn
Digital Media Lab, Korea Advanced Institute of Science and Technology,
119, Munji-Dong, Yuseong-Gu, Daejeon, South Korea
e-mail: aalee@kaist.ac.kr

M. Hahn
e-mail: mshahn@ee.kaist.ac.kr

J. Kim
School of Electronics and Computer Engineering, Chonnam National University,
Gwangju 550-757, South Korea
e-mail: jsworld@chonnam.ac.kr

H. Choi
Department of Electrical Engineering, Korea Advanced Institute of Science
and Technology, 119, Munji-Dong, Yuseong-Gu, Daejeon, South Korea
e-mail: hwchoi2@ee.kaist.ac.kr

1 Introduction

Before smartphones got popular, bicyclists needed dedicated GPS devices [1] for position tracking, and the recording and sharing of the users' GPS trajectory were limited. Currently, the smartphones make it possible for them to easily collect and share the user's daily position logs with built-in GPS module and tracking applications. The tracking applications of the smartphones record positions in two- (latitude and longitude), or three-dimensional (latitude, longitude and elevation) coordinates and overlays the trajectories on the map. They also provide additional activity information such as average speed, acceleration, pace and calories burned which were retrieved from the trajectories.

The bicyclists have tendency to share their riding experiences with others [2]. Recently, GPS trajectories sharing on the website is one of the most popular activities in the online bicycle communities, so the number of the shared bicycle GPS data is steadily increasing. The route choice of bicyclists differs from the motorists [3, 4]. While the latter tends to choose the shortest path, former prefers to ride on the segregated bicycle facilities such as bicycle tracks or lanes. Therefore, the distribution of the bicycle GPS trajectories is tends to concentrate on the specific locations.

The aggregated trajectories have been studied in various fields. In the fields of intelligent transportation system, massive GPS trajectories are used to create and modify the road maps in real-time [5]. They also have been applied to the map-matching techniques to enhance accuracy [6]. For the visualization of the spatial data, the GPS trajectories provide the base elements for drawing: vertices and lines. With these primitive elements, the trajectories can simply be displayed as sequences of line segments.

Various methods and techniques were proposed to provide more insight into the visualized trajectories. Visualizing GPS trajectories based on the kernel density estimation provides intensity of the scattered data with proper range of the distribution [7]. Scheepens et al. [8] proposed a visualization system which consider the moving objects as multivariate time series and shown the architecture of the system. The system is based on the density map and applied multiple density fields to the visualization.

Up to date, most of the trajectories visualization methods are focused on the high-speed transportations such as automobiles, airplanes and vessels [7, 8]. The visualization methods for bicycle trajectories or pedestrians [9] are relatively hard to find. In this paper we propose a simple color coding method for the visualization of bicycle trajectories. The motivation of the method is based on the behavior of the bicycling. The bicycle trajectories on the segregated bicycle lanes are less concentrated on one side of the road compare to the trajectories of automobiles because bicyclists cross the centerline of the road much more often than motorists. This symptom creates overlapping of the trajectories and makes display hard to understand.

Fig. 1 Trajectories of bicycles collected from two-way bicycle path

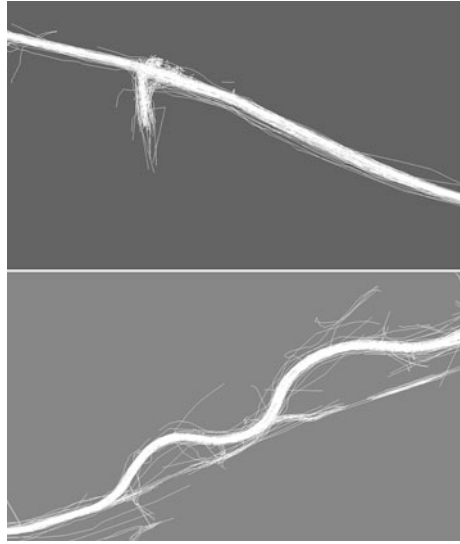


Figure 1 shows the bicycle trajectories collected from segregated bicycle tracks alongside Han River in Seoul. The bicycle track consists of two-way bicycle path, thus the trajectories were collected from both side of the roads. The overlapping of the trajectories can be found from the trajectories of automobiles. However, while the automobile trajectories overlap among their own ways, the bicycle trajectories overlap in both ways. To solve this problem, the proposed method changes color of the trajectory based on its vector angle and provides the direction and flow.

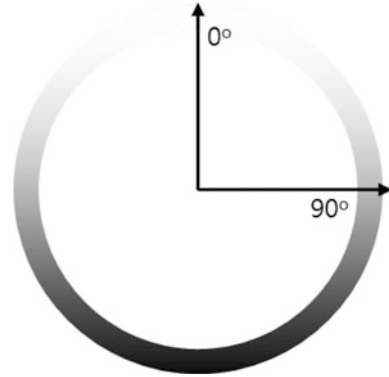
In Sect. 2, the detailed design of the color coding method will be explained. The experimental results of the proposed method will be shown in Sect. 3. In Sect. 4, we draw conclusions and suggest future work.

2 Method

The trajectory of a moving object is a gathering of the position sequence according to the timeline. In the aspect of visualization, it can be considered as the sequence of line segments. Each line segment consists of two neighboring points, and by connecting adjacent line segments, a trajectory can be visually created. The proposed color coding method maps each point of the trajectories with different colors to provide effective visual information.

The color of the point is based on the angle between the vector of the line segment and a vector directed to the north. Since the angle ranges from 0° to 360° , the color of the point can be gradually changed according to the direction of the line segment. Based on the vector angle of the line segment, elements of the color such as chroma and hue were controlled. Basic idea of color change can be simply

Fig. 2 Chroma transition and angle of the line segment



described by using an edge of a circle. Figure 2 shows the chroma transition of the point according to the angle. As the angle closes to 0° or 360° color gets brighter, and it gets darker when the angle closes to 180° .

By applying the chroma transition, the line segments around 0° vector angle and 180° can be divided. However, the line segments having 90° vector angle has same chroma value with the line segments having 270° . To differentiate these line segments, each half of the circle has to have different hue. Figure 3 shows the hue variations pairs. Each pair consists of complementary colors except number (6) which consists of two primary colors.

3 Experimental Results

The proposed color coding method was applied to the bicycle trajectories collected from abovementioned bicycle tracks. Over one thousand GPS trajectories were recorded by using smartphone applications and the dedicated GPS devices. Invalid and erroneous data were excluded during preprocessing.

Figure 4 shows the visualized bicycle trajectories. The hue variation pairs in Fig. 3 were applied to visualizations in Fig. 4 according to the number shown in the figures. The visualized trajectories in the figure show their direction and transition of flow with gradually changing color of the line segments.

4 Conclusions and Future Work

In this paper, we proposed a color coding method for visualization of bicycle trajectories. The method was motivated from the bicyclists' behaviors which differ from the motorists' counterparts. The purposes of the method are to differentiate opposing trajectories and provide visual cues to the user. To evaluate the color

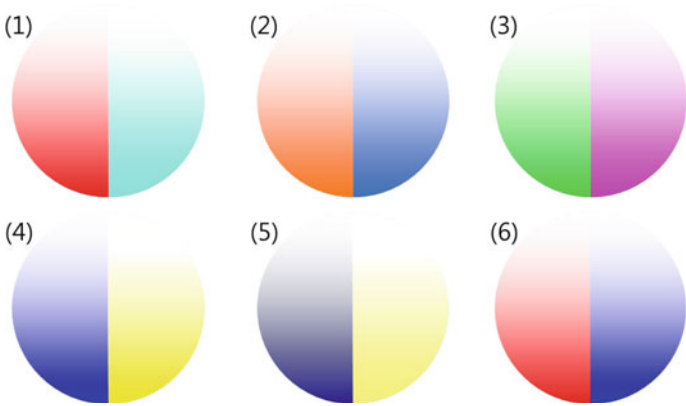


Fig. 3 Hue variation pairs for line segments

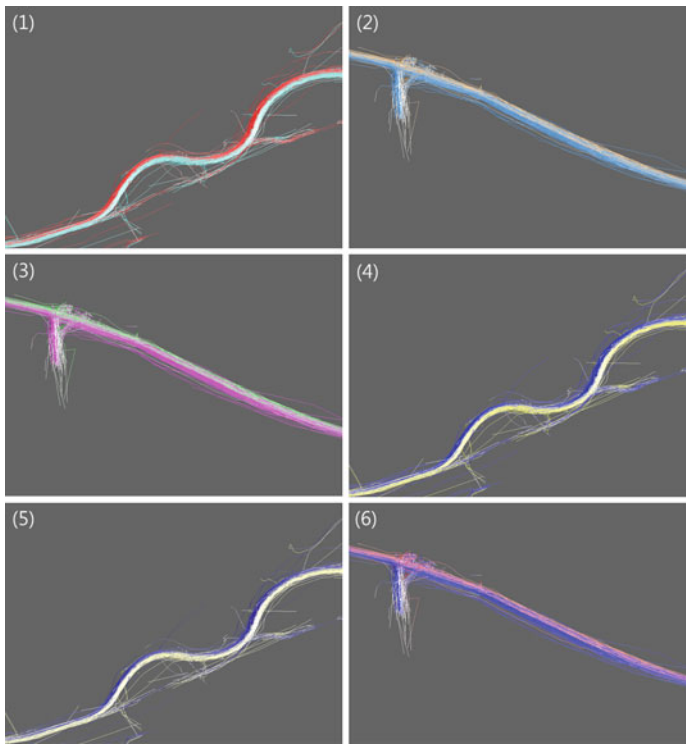


Fig. 4 Bicycle trajectories visualized with color coding

coding, trajectories collected from segregated bicycle paths were visualized. As a result, applying color coding provided insights among the numerous overlapping trajectories.

In future research, we are planning to analyze principal component of the bicycle trajectories. The principal components such as principal curves can extract the spatial summary of the data. With the principal components, density estimation will be applied to enhance the visualization of the trajectory. We expect applying these methods to visualization will reveal the characteristics of bicycle trajectories.

Acknowledgments This research is supported by Ministry of Culture, Sports and Tourism(MCST) and Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Research & Development Program

References

1. Garmin. <http://www.garmin.com/us/>
2. Dill J, Gliebe J (2008) Understanding and measuring bicycling behavior: a focus on travel time and route choice. OTREC-RR-08-03 final report
3. Caulfield B, Brick E, McCarthy OT (2012) Determining bicycle infrastructure preferences—a case study of Dublin. *Transp Res Part D-Transp Environ* 17(5):413–417
4. Buehler R, Pucher J (2011) Cycling to work in 90 large American cities: new evidence on the role of bike paths and lanes. *Transportation* 39(2):409–432
5. Agamennoni G, Nieto JI, Nebot EM (2011) Robust inference of principal road paths for intelligent transportation systems. *IEEE Trans Intell Transp Syst* 12(1):298–308
6. Brunsdon C (2007) Path estimation from Gps tracks. In: The 9th international conference on geocomputation. National University of Ireland, Maynooth
7. Hauser H, Lampe OD (2011) Interactive visualization of streaming data with kernel density estimation. In: *Proceedings of IEEE pacific visualization symposium*
8. Scheepens R, Willems N, van de Wetering H, Andrienko G, Andrienko N, van Wijk JJ (2011) Composite density maps for multivariate trajectories. *IEEE Trans Vis Comput Graphics* 17(12):2518–2527
9. Zheng Y, Wang L, Zhang R, Xie X, Ma WY (2008) GeoLife: managing and understanding your past life over maps. In: *Proceedings of the 9th international conference on mobile data management*. IEEE Press, Beijing pp 211–212

User-Oriented Load Balancing Scheme for MMORPG

Hye-Young Kim

Abstract MMORPGs (Massively Multi-player Online Role-Playing Games) follows a client–server model that has the numerous gaming users with many interactions at the same virtual world, massive loading that result in delays, resource shortages, and other such problems occur. To solve this, many developers devote research to load-balancing servers, yet due to steady and dynamic map divisions, such research is unreliable. Many developers propose algorithms to distribute the load on the server nodes, but the load is usually defined as the number of players on each server, what is not an ideal results. So, we propose a gaming user-oriented load balancing scheme for the load balancing of MMORPG servers in this paper. This scheme shows effectiveness at dealing with hot-spots and other gatherings of gaming users at specific servers compared to previous methods.

Keywords MMORPG · Load balancing · User oriented · Seamless · Gaming user

1 Introduction

A Massively Multi-player Online Role-Playing Games (MMORPG) is immensely popular with several commercial games reporting millions of subscribes. Most of them require huge virtual worlds, significant hardware requirements, and dedicated support staffs [1]. The main characteristic of MMORPGs is the large number of

H.-Y. Kim (✉)

Major in Game Software, School of Games, Hongik University, Shinan-ri,
Jochiwon-eup, Yoengi-gun, Chungnam 339-701, South Korea
e-mail: hykim@hongik.ac.kr

players, having dozens, or even hundreds, of thousands of participants simultaneously. This large number of players interacting with one another generates traffic on the support network which may grow quadratic compared to the number of players in the worst case [2].

MMORPGs follow a client-server model. When using client-server architecture, it is necessary that the server intermediates the communication between each pair of players that game is intended to provide guarantees of consistency and resistance to cheating. Each of the players remotely controls one or several in-game characters called avatars. So, this server will have a large communication load, thus, it must have enough resources to meet the demand of the game.

The problem is that it must be delegated to each server node a load proportional to its power when using a distributed server. No matter to which server each player is connected, their game experience will be similar and the time it takes to be notified of actions from other players as well as of state change in the virtual environment of the game [3–5].

The Players can freely move their avatars through the game world. This makes possible the formation of hot-spots around which the players are more connected than in other regions of the virtual environment. MMORPGs not only permit but also stimulate the formation of these points of interest [6–9]. However, if these avatars are close to each other, each player should be updated not only of his own actions, but also of the actions of the other player [10, 11].

So, it is not enough just to divide the players between servers, even if this division is proportional to the resources of each one of them. Also, the overhead of the distribution is an important issue.

Therefore, we propose a scheme that equally distribute of the load using a gaming users-oriented load balancing rather than a partition of maps on gaming server and compared to the previous studies. In addition, we designed this to facilitate interactions between gaming users, neighbor gaming users on the map as much as possible on a single gaming server, MMORPG gaming user-oriented load balancing scheme to map segmentation by combining the efficiency of the load balancing of the game in progress when raised of gaming users.

The rest of the paper is structured as follows. [Section 2](#) describes our proposed architecture and provides the detailed gaming user-oriented load balancing scheme. We present a performance evaluation and the mathematical analysis, and a comparison of performance between the previous studies and the proposed scheme in [Sect. 3](#). And [Sect. 4](#) constitutes a summary of the proposal.

2 Gaming User-Oriented Load Balancing Scheme

MMORPGs places importance on interaction between gaming users and this interaction usually takes place in visual range in the game.

Therefore, when gaming users are within visual range of each other, it is best to locate them in one server. To manage gaming users like the above, both the scope

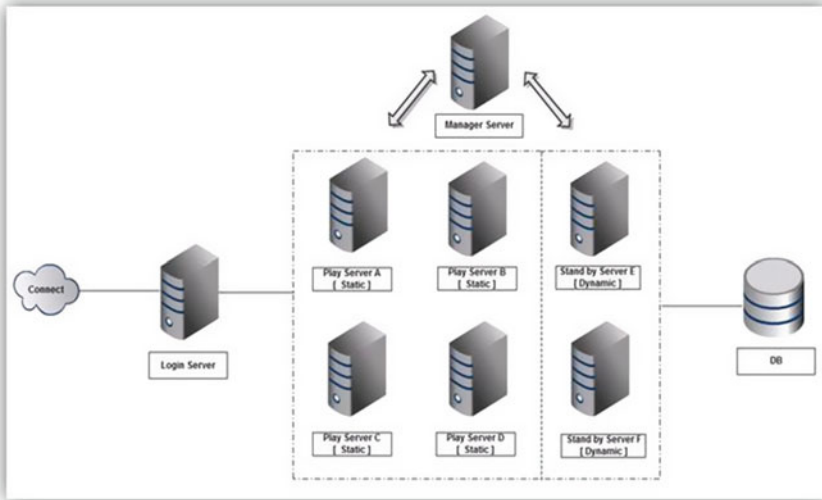


Fig. 1 Proposed system architecture

of interactions and probable interactions have to be taken into account and then set the neighboring gaming users area to about 1.5–2 times the visual range with the gaming users at the center. When new gaming users process to log-in, first check whether the existing gaming users are in the nearest gaming user area and if they are, then distribute the new gaming users to the nearest gaming user areas. We show the architecture of the proposed system configuration for load balancing of gaming users in Fig. 1.

Figure 2 shows an example of the nearest gaming user distribution scheme. Server A and Server B each has 4 and 3 gaming users and gaming user 1 and gaming user 2 are logged-in. The load balancing server compares the location of the currently logged-in gaming users 1 and 2 and then allocates gaming users 1 and 2 to either server A and B depending on which one has the nearest gaming user.

For load-balancing, each the center of each gaming server gets updated at the load-balancing server and this point is equated by the average coordinates of the game users. Load-balancing servers distribute gaming users to the server which is closest to the center of the server when new game users log-in and when all gaming users of the game servers are not included in the nearest section.

3 Performance Analysis of Proposed Scheme

MMORPGs do not distribute equally accordingly to the gaming user preference of location and problems in the geography of the game maps, and problems also occur in hot-spots [12, 13].

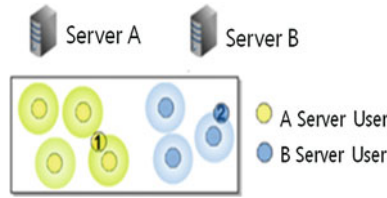


Fig. 2 Neighbor gaming user-oriented load balancing

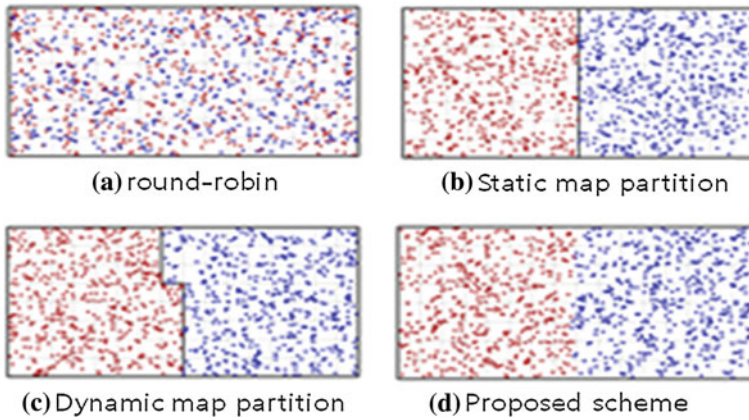


Fig. 3 Comparison of server distribution methods

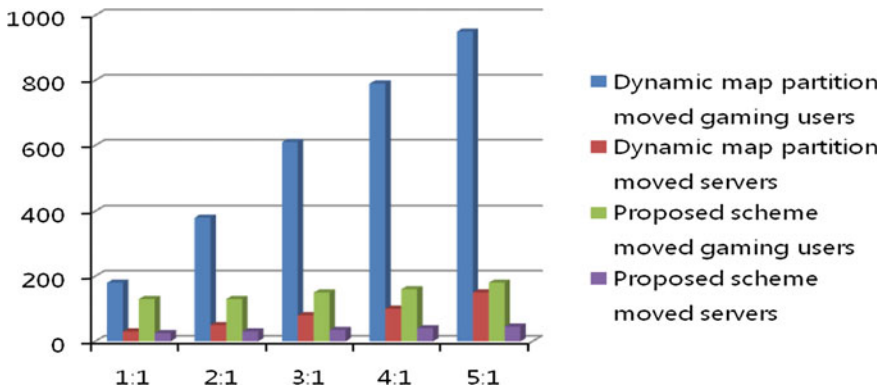


Fig. 4 Comparison of gaming users by user's congestion

Figure 3 shows a figure of a situation in which the server relocates when a 1024 * 512 sized map's coordinates are (0.0)–(512, 512) and the ratio of the accessing of the gaming users are increased by 1:1, 2:1, 3:1, 4:1, and 5:1. It also shows that in the case dynamic map partitioning load-balancing, the rate of the

gaming user's relocation and the relocation of the server itself is proportionate to the increase of the condensation of the gaming users.

However, game user-oriented load-balancing shows that there is no difference in the relocation of the server and gaming users despite the increase of the difference in the log-in rates of areas. Hence gaming user-oriented load-balancing is a much more flexible gaming user location distribution method and more efficient at relocating servers than dynamic map partitioning load-balancing. In the Fig. 4 present the above descriptions.

4 Conclusions

We proposed two schemes, taking care of neighboring gaming users in one server by searching for nearby gaming users when one logs-in a server, and lessening the load that occurs by server relocation by using gradients between servers to relocate the servers resulting in a common critical section. Relocation of servers with load-balancing showed an efficiency of about 40 % with a number of total gaming users at 1000, map distribution at an average of 156 persons, and gaming user-oriented distribution server had 99 gaming users relocated.

Also, according to gaming user preferences, dynamic map partitioning load-balancing had server relocation and relocated gaming users grow proportionally to the condensing of the gaming users in a specific location. Compared to that, our scheme showed no large difference in the rate of server relocation despite the increase differences in the amount of gaming users in a game area.

Acknowledgments This work was supported by 2012 Hongik University Research Fund.

References

1. Ahmed D, Shirmohammadi S (2008) A microcell oriented load balancing model for collaborative virtual environments. In: Proceeding of the IEEE conference on virtual environments, human computer interfaces and measurement systems, VECIMS, pp 86–91
2. Hye-Young K (2012) An efficient access control scheme for online gaming server. In: Proceeding of the computer science and convergence, lecture notes in electrical engineering, vol 114, part 1, pp 259–267
3. Bezerra CEB, Geyer CFR (2009) A load balancing scheme for massively multiplayer online games, multimedia tools and applications, vol 45, no. 1, pp 263–289
4. Zamboni MA, Ferretti S (2005) Interactivity maintenance event synchronization in massive multiplayer online games, Technical report UBLCs-2005-05
5. Nae V, Losup A (2011) Dynamic resource provisioning in massively multiplayer online games, Parallel and distributed systems, IEEE transactions on, vol 22, Issue 3, pp 380–395
6. De Grande RE, Boukerche A (2009) Dynamic partitioning of distributed virtual simulations for reducing communication load, Haptic audio visual environments and games, IEEE international workshop on, pp 176–181

7. Duong TNB, Zhou S (2003) A dynamic load sharing algorithm for massively multiplayer online games, network ICON 2003, the 11th IEEE international conference on, pp 131–136
8. Grosu D, Chronopoulos AT (2004) Algorithmic mechanism design for load balancing in distributed systems, systems Man and Cybernetics, part B: cybernetic, IEEE transactions on, vol 34, Issue 1, pp 77–84
9. Andrade G, Corruble V (2005) Challenge-sensitive action selection an application to game balancing, intelligent agent technology, IEEE/WIC/ACM international conference on, pp 194–200
10. Wang J, Yue Z (2010) A finding less-load server algorithm based on MMOG and analysis, intelligent computation technology and automation (ICICTA), international conference on, vol #1. pp 96–99
11. Huang G, Ye M, Cheng L (2004) Modeling system performance in MMORPG, global telecommunications conference workshops, IEEE, pp 512–518
12. Nae V, Prodan R, Fahringer T (2010) Cost-efficient hosting and load balancing of massively multiplayer online games. GridComputing CGRID, 2010, 11th IEEE/ACM international conference on, pp 9–16
13. Quax P, Cleuren J, Vanmontfort W, Lamotte W (2011) Empirical evaluation of the efficiency of spatial subdivision schemes and load balancing strategies for networked games, computer communications and networks (ICCN), proceeding of 20th international conference on, pp 1–6

Smart Desk: A Single Interface Based File Sharing System

Naveed Ejaz, Sung Wook Baik and Ran Baik

Abstract File sharing is the practice of distributing or providing access to digitally stored information. The advantages of the existing file sharing repositories are limited by the factors of the cost and usage of multiple repositories for different purposes. This paper presents the development of an intelligent web-enabled facilitation system called Smart Desk which provides a single file sharing interface to the users, with the capability of virtually unlimited space. Smart Desk uses the services of other repositories by storing the data on other networks, and thus providing virtually unlimited space to the users for storing data. The drag & drop based single interface allows the users to easily view, edit, delete, upload, and copy data. A usability study has been performed to evaluate the efficacy of the system. The evaluation shows that Smart Desk is a useful interface for sharing files which is time effective and user friendly.

Keywords File sharing · User interface design · Model view controller · Data sharing

N. Ejaz · S. W. Baik
College of Electronics and Information Engineering, Sejong University,
Seoul, Republic of Korea
e-mail: naveed@sju.ac.kr

S. W. Baik
e-mail: sbaik@sejong.ac.kr

R. Baik (✉)
College of Business, Honam University, Gwangju, Republic of Korea
e-mail: baik@honam.ac.kr

1 Introduction

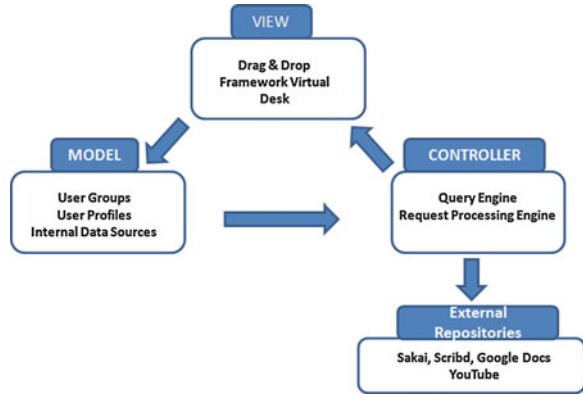
Modern technology is continuously changing all the spheres of life. The digitization of data and its easy accessibility on the internet is prevailing [1, 2]. The data in digital format is now preferred everywhere. The file sharing is the practice of distributing or providing access to the digitally stored information, such as computer programs, multimedia data, documents, and e-books, etc. [3]. The file sharing may be implemented using a variety of storage, transmission, and distribution models. The demand for managing contents like audio, video, and images over the internet is on the rise. With the advent of large repositories on the internet (such as Yahoo and Google groups, Sakai, Moodle, Scribd, and Flickr etc.), the tasks of data uploading and storage have become quite easy [4, 5]. However, these repositories do have their own limitations when it comes to free sharing, due to which many users are unable to make full use of these file sharing facilities. Moreover, the users have to use multiple file sharing repositories for different purposes. This situation demands for the development of an efficient, effective, and robust single interface system for storing and managing data of all the repositories.

This paper presents the framework and development of a robust file sharing interface called 'Smart Desk'. It is an intelligent web-enabled facilitation system which provides a single file sharing interface. Smart Desk presents solutions for two major problems of file sharing by providing: (1) Unlimited data storage, (2) One click access to all repositories. The system uses the services of other network repositories, keeping all the data on their networks, and thus provides unlimited space for the users for data storage. Smart Desk uses different application programming interface (APIs) like Google documents list API, Google spreadsheets data API, YouTube data API, Picasa web albums data API, Scribd API, and Flickr API [5]. In this way, content created elsewhere can be dynamically posted and updated on the system. The users can easily view, edit, delete, upload, and copy their data through a drag & drop single interface. The system also provides notifications for group users on their respective virtual desks. Therefore, the proposed interface is time efficient for every individual who wants to avoid typical hassles in uploading files individually through different accounts on file sharing websites. The implementation of the system has been performed on the .Net framework because of its ability to support multiple programming languages and other features such as interoperability, common runtime engine, language independence, simplified deployment, security, and portability.

2 Framework

The architecture of the proposed system, shown in Fig. 1, is based on a famous design pattern Model-View-Controller (MVC) [6]. Each of the model, view, and controller components are discussed in subsequent sub-sections.

Fig. 1 Architecture of the proposed system



2.1 The Model

The model is the backbone of the whole system. It includes all the databases in which the complete data of the system is to be stored. Due to the complexity of the system, it is divided into different parts such as User Groups, User Profiles, Internal Data Sources, User Types, and User Articles. Figure 2a shows the main components of the model. The user groups store all the information of the groups. They store the references of the files placed on the groups, group messages, and the group profiles. The files include the group files of the group. The group messages refer to the messages automatically generated or manually sent by the group owners, coordinators, and users, which are then broadcasted to every group member. The group profile includes the information about the owners, coordinators, and the users of the group. The internal data sources eventually map to the external repositories. They save the references of the files placed on external repositories and the information regarding those files, such as owners of the files. The user articles include all the information about authors of the article, the ratings of the articles and which users have verified the articles.

The user type of the account holders refers to the different types of users which include administrators and account holders into the respective data sources of the user types. The administrators include the users which run the whole system. This information is stored in account holders' type along with the roles assigned to the account holders including owners, coordinators, users, etc.

2.2 The View

The views have to deal with the users' experience of the system. A view basically is the presentation layer in which the users can interact with the system. The view of Smart Desk has three main parts: drag & drop framework, virtual desk, and

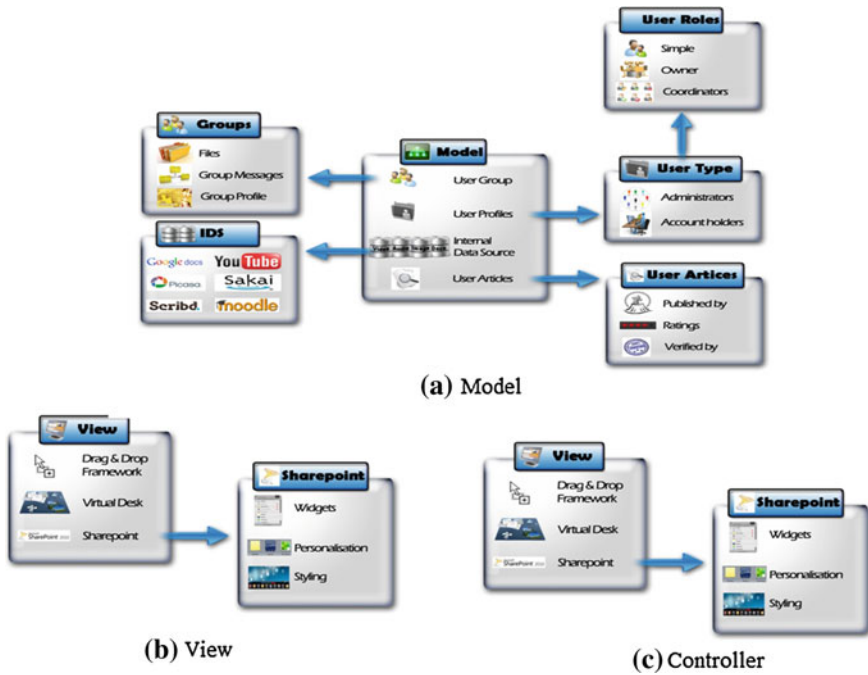


Fig. 2 The detailed model, view and controller of the proposed framework

SharePoint. Figure 2b shows the main components of the view. Smart Desk is based on the drag & drop framework which allows the users to drag-and-drop data for enabling real-time moving (or copying) of data on the page. The users can copy an element from a source to a target, or just move it to its new location with the help of the drag & drop framework. This framework is a very important feature of our system because it endows a simple copy and paste facility and thus can easily be used by a naive computer user. The drag & drop framework provides the users with all of the client-side scripting to handle the mouse drag & drop process portably across browsers. ASP.NET and AJAX controls leverage the patterns implemented by the drag & drop framework. The framework is perfectly interoperable with any kind of HTML element content and requires absolutely no software other than a Web browser for viewing web pages. The drag & drop facility is provided by the help of the drag and drop framework of Infragistics [7]. The users are able to access their virtual desks. In this case, they can make virtual shelves and place their old files in their virtual shelves. They can place their important files on their desk through the drag & drop framework. The interface is similar to a physical desk. Microsoft SharePoint 2010 [8] is used to give personalized features to the users. The users can perform styling, personalization, and addition of widgets to their desks. Because of this, the users will have a more personalized feeling about their desks.

2.3 *The Controller*

The main aim of the controller is to connect the model with the view. The controller is handling all the communication between the view and the model. Figure 2c shows the main components of the controller. There are two main components in controller: (1) Query analyzer engine, and (2) Request processing engine. The function of the query analyzer engine is to get the user query from the view and then analyze the type of query. It then forwards the query to the appropriate request processing engine. The request processing engine consists of “internal request processing” and the “external request processing”. The responsibility of the internal request processing engine is to process the queries which lie in the internal domain of the system. The internal request processing engine has two components to deal separately with the user and the group queries. The responsibility of the external request processing engine is to process the queries which require communication with the external repositories. The external repositories include Google Docs, Picassa, Youtube, Scribd, Moodle, and Sakai.

3 Evaluation of Smart Desk

The evaluation of the Smart Desk environment was carried forward by using a usability assessment. The usability data [9, 10] can be used to know about the functional effectiveness, efficiency, ease in learning and use, motivational influence, and quality assurance of a program. The usability data was obtained by surveying 20 students of Sejong University, Korea. All of the 20 students completed the usability questionnaire that included 20 items using a five-point Likert scale of 1–5. On the scale, 1 means strongly disagree and 5 means strongly agree. The questionnaire was titled “Survey on the usability of Smart Desk—A file sharing interface”. A total of 20 questions were posed, including 3 questions related to overall perception of the system, 8 questions related to the ease in use and comfort, 4 questions related to the efficacy of the system in saving time, and 3 questions related to the design of the graphical user interface. In addition, there were 2 questions pertaining to the subject’s background, which are helpful in describing the characteristics of the sample.

Table 1 shows the summary of each question and the mean values of all the scores awarded by users for each of the questions. As can be seen, most of the users are above average level computer users with reasonable experience in file sharing software systems. It can be observed that the respondents present positive response on most criteria of the evaluation. The results also demonstrate that the proposed system is useful and convenient for sharing files. The users also find the interface of the system to be user friendly and easy to use. Moreover, in the opinion of the respondents, it is easy to learn the system options and the environment is comfortable. The system is time effective as compared to the traditional

Table 1 Results of usability study

Dimension	No.	Items	Score
Subject background	1	Expertise in usage of computer	3.75
	2	Experience in usage of file sharing systems	3.79
Overall perception	3	User friendliness	4.15
	4	Preference over other Applications	3.95
	5	Comfort level	4.2
Ease in usage	6	Ease in understanding functions	4.33
	7	Ease in movement of items	4.12
	8	Ease in rectifying errors	4.33
	9	Ease in knowing what to do next	4.32
	10	Ease in remembering how to do things	3.75
	11	Enjoyability	4.16
	12	Ease in switching among different tasks	4.35
	13	Ease in determining the available options	4.52
Efficacy of the system in saving time	14	Overall speed of the system	4.51
	15	Using the system may increase productivity	4
	16	Using the system may enhance effectiveness	4.26
	17	Time effectiveness being a single file sharing interface	4.77
GUI	18	Aesthetics	4.12
	19	Organization of information	2.85
	20	Design for all levels of users	4.35

file sharing systems. On the whole, the users are particularly satisfied by the interface of the system. A concern however, is the organization of information which gets a sufficiently low score. This may be because of accumulating too much information on one display. We intend to rectify this issue in the future.

4 Conclusion

In this paper, we presented a system for providing a single file sharing interface by using the services of all data repositories thus endowing unlimited storage free of charge. The proposed interface provides an easy capability to each individual for avoiding typical hassles during file sharing using different accounts on various file sharing websites. A usability based study shows that the users find it easier to view, edit, delete, upload and copy their data through this one single interface. In overall rating, the participants rated Smart Desk as a useful interface for sharing files regarding the amalgamation of data from different repositories.

Acknowledgments This research was supported by, (1) Research fund from Honam University, 2012, (2) The Industrial Strategic technology development program, 10041772, (The Development of an Adaptive Mixed-Reality Space based on Interactive Architecture) funded by the

Ministry of Knowledge Economy (MKE, Korea), and (3) The MKE (The Ministry of Knowledge Economy), Korea, under IT/SW Creative research program supervised by the NIPA (National IT Industry Promotion Agency)” (NIPA-2012- H0502-12-1013).

References

1. Ejaz N, Tariq TB, Baik SW (2012) Adaptive key frame extraction for video summarization using an aggregation mechanism. *J Vis Commun Image Represent* 23(7):1031–1040
2. Ejaz N, Manzoor U, Nefti S, Baik SW (2012) A collaborative multi-agent framework for abnormal activity detection in crowded areas. *Int J Innov Comput Inform Control* 8(6):4219–4234
3. Shen H (2010) An efficient and adaptive decentralized file replication algorithm in P2P file sharing systems. *IEEE Trans Parallel Distrib Syst* 21(6):827–840
4. Yang S, Jin H, Li B, Liao X, Yao H, Huang Q, Tu X (2009) Measuring web feature impacts in Peer-to-Peer file sharing systems. *Comput Commun* 32(12):1418–1425
5. Khuzadiv M (2010) The collaboration management culture, *IEEE magazine on aerospace and electronic systems*
6. Mcheick H, Qi Y (2011) Dependency of components in MVC distributed architecture, *Proceeding of Canadian conference on electrical and computer engineering*
7. Online. <http://www.infragistics.com/default.aspx>
8. Patton RM, McNair W, Symons CT, Treadwell JN, Potok TE (2012) A text analysis approach to motivate knowledge sharing via microsoft sharepoint, *Proceeding of international conference on system science*
9. Buschmann F (2011) Unusable software is useless, part 1, software, *IEEE*
10. Yao M, Jinjuan F (2011) Evaluating usability of three authentication methods in web-based application, *software engineering research, proceeding of international conference on management and applications*

Feature Reduction and Noise Removal in SURF Framework for Efficient Object Recognition in Images

Naveed Ejaz, Ran Baik and Sung Wook Baik

Abstract Speeded up Robust Features (SURF) is an interest point detector and descriptor which has been popularly used for object recognition. However, in real time object recognition applications, SURF framework can not be used because of its expensive nature. In this paper, a feature reduction process is proposed by using only the most repeatable features for matching. The feature reduction step results in a remarkable computational speed up with little loss of accuracy. A noise-reduction process allows a further increase in matching speed and also reduces the false positive rates. A modified definition of the second-neighbor in the nearest neighbor ratio matching strategy allows matching with increased reliability. The comparative analysis with SURF framework shows that the proposed framework can be useful in applications where the accuracy can be sacrificed to save computational cost.

Keywords Object recognition · SURF features · Pose estimation · Noise reduction · Feature matching

N. Ejaz · S. W. Baik (✉)
College of Electronics and Information Engineering, Sejong University, Seoul,
Republic of Korea
e-mail: sbaik@sejong.ac.kr

N. Ejaz
e-mail: naveed@sju.ac.kr

R. Baik
College of Business, Honam University, Gwangju, Republic of Korea
e-mail: baik@honam.ac.kr

1 Introduction

Object recognition is a very important component of many computer vision applications including content based multimedia search [1], automated monitoring of surveillance videos [2], and action recognition [3]. Owing to this importance, a lot of work has been done in this field. However, effective object recognition still remains a challenge because of the significant variation exhibited by real world objects and inter-class similarity of objects.

An important set of object recognition techniques are those which are based on local representation of image using invariant feature points [4–6]. The basic theme of such methods is to extract interesting points of an object in an image, and then provide feature description of the object. The feature description is usually generated from a set of training images and then can be used to find the similar object in a test image having many objects. An important requirement for such methods is that the features computed from the training image must also be detected under varying lightning and noise conditions. The object recognition methods based on interest points are popular primarily because of their robustness in providing better invariance in terms of size, rotation and illumination as compared to the more traditional template based methods [7]. There are three major steps to find the point correspondences between the images: detector, descriptor and matching. The aim of detector step is to locate ‘interest points’ from the images. The descriptor develops feature vectors based on the neighborhood of each interest point. The descriptor must be distinctive in nature and must exhibit robustness against noise and geometric transformations. Finally, the matching of the descriptor vectors is performed using some distance measure. The bottleneck of such object recognition techniques is the huge matching time required to match the corresponding points using descriptor vectors. It is desirable to have low dimension descriptor vectors for faster interest point matching. However, the low dimension feature vectors tend to be less distinctive. Speeded up Robust Features (SURF) is a popular interest point detector and descriptor originally suggested by Bay et al. [4]. SURF gives comparable or better results than that of another popular technique called SIFT [6] with reduced computational cost. The standard SURF results in a 64 dimensional descriptor vector of floating point values. SURF finds a large number of highly discriminative features from an object. However, in theory, only the more repeatable features for each object must be stored in the feature database. The identification of these most repeatable features is a daunting task.

This paper presents a framework for recognition of multiple objects using SURF features in an efficient and reliable manner. This work presents a technique for the identification of the most repeatable features under affine transformations. During the training phase, two additional steps: feature reduction and feature noise-reduction, are employed to reduce the number of features in the database. This reduction of features provides a remarkable speed-up during the matching phase. During the matching phase, a modified definition of second nearest neighbor is used which increases the reliability of matching process, as a larger

number of features are correctly matched. Our proposed modifications exhibit promising results and thus can be useful in many applications (e.g. web applications) where the accuracy can be sacrificed to achieve performance benefits. Another challenge in the usage of SURF and related techniques is the identification of false positives in feature matching. In original SURF, the feature matching is performed using the nearest neighbor ratio matching strategy. This paper presents a definition of the second neighbor based on k-means clustering of SURF features, as opposed to a definition based on object classes. This allows larger number of features to be matched.

2 Problem Formulation

This section presents the design of the framework for reliable and efficient object recognition using modified SURF. The main steps of the framework are shown in Fig. 1. The rest of this section describes the two phases and feature database in detail.

2.1 Learning Phase

The three steps of learning phase include feature extraction, reduction of SURF features, and removing noisy features using k-Means clustering. Lastly, these features are stored in the feature database along with object tags. Each of these steps is explained in detail below.

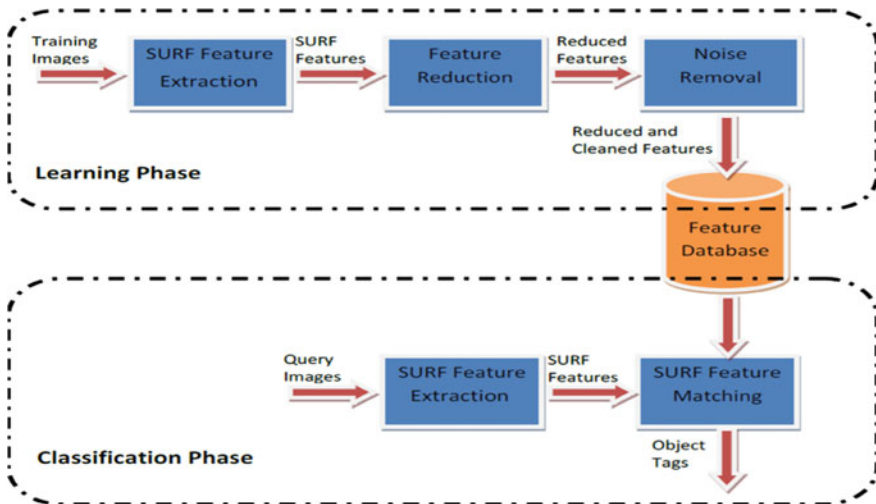


Fig. 1 Framework for object recognition using SURF

2.1.1 Surf Feature Extraction

Initially, the interest points in an image are localized using the SURF interest point detection mechanism. For details of SURF feature extraction, we refer readers to [4]. After detection of interest points, each interest point is then described using SURF feature descriptors. SURF feature descriptor is based on the distribution of first order Haar wavelet responses in x and y direction. The first step in the extraction of the descriptor is to build a square region of size 20 centered on a particular interest point. In the next step, the region is split into 4×4 sub-regions. This means that there is a 4 dimensional descriptor vector \mathbf{v} for each sub-region which is given as:

$$\mathbf{v}^T = \left[\sum d_x \sum d_y \sum |dx| \sum |dy| \right] \quad (1)$$

There are 16 sub-regions, so combining feature vectors for all sub-regions results in a descriptor vector of length 64.

2.1.2 Feature Reduction

The basic principle of the feature reduction step is that an interest point is repeatable if it is detected by the descriptor under varying view conditions. The assumption is that only the most repeatable features can be used for the purpose of object recognition. In our work, the focus is only limited towards the features that are most repeatable under affine transformations. In the first step, SURF feature vectors are extracted for an image and its affine transformed variants. The used affine transformations include $2 \times$ scale-up, $2 \times 25^\circ$ clockwise rotation and XY-skew. The most repeatable features are then found by comparing the features extracted from the original training image with the feature extracted from the affine variants of the images. The features that do not emerge in all transformed images are filtered out. On an average the number of features is reduced by about 7 times during this step.

2.1.3 Noise Removal

After reduction of features, the noise removal step aims to remove the features that are likely to be generated by background or noise in the object images. The proposed noise reduction step is based on clustering. The feature vectors are clustered and only the dominant object is selected from each cluster. The k-means algorithm [8] is used for clustering the feature vectors. An important step in k-means clustering is choosing the number of clusters 'k' a priori. For estimating the value of 'k', the following formula is used:

$$k = \alpha * |O| * |I| \quad (2)$$

In the above equation, $|O|$ is the number of objects in the database; $|I|$ is the average number of instances per object in the database and α is a factor which is used to control the size of the clusters. In most cases 3 is found to be the ideal value of α . Once the features vectors are clustered, a dominant object is selected for each cluster using a majority vote algorithm. All the features not belonging to the dominant object in the clusters are removed. This step is expected to reduce the noise to a significant extent. However, in applications/works for high dimensional feature vectors, some noise may still be left.

2.2 Classification Phase

The second step is matching of the extracted features from the image to the feature in the database. The original SURF feature matching is based on the nearest neighbor ratio matching strategy as used in SIFT. Bay et al. [4] compare an interest point in the test image with features in the database by calculating their Euclidean distance. If this distance is <0.8 times the distance with the nearest neighbor, a matching pair is detected. Lowe [6] defined the second closest neighbor as being the closest neighbor as it is known to come from a different object rather than the first. The rationale behind such an approach is that the correct matches need to have the closest neighbor significantly closer than the closest incorrect match to achieve reliable matching. For false matches, there will likely be a number of other false matches within similar distances due to the high dimensionality of the feature space [5, 6]. The second nearest neighbor can be thought to be providing a measure of the density of the false matches.

Using this principle, we modify the definition of second neighbor as being the closest neighbor known to have come from a different ‘cluster of features’ than the first. These clusters are created using the k-Means algorithm as described earlier. The rationale behind this modification is that a number of features are not matched using the definition provided by Lowe [6] as other object may have similar features. However, k-Means algorithm clusters the visually similar features in the same cluster and thus the features are matched. It is observed that the number of features matched are increased substantially using this definition and the number of false positives is also reduced. An object is said to be found in an image if three or more features from that object are matched. The pose of the object is estimated using the pose information stored in the features database.

3 Results

This section presents the results of the experimental setup used and the results of the proposed framework. We compared our proposed framework with standard SURF based on the UK Benchmark Object Recognition Dataset [9]. The dataset contains 10,200 images of about 2,500 objects with 4 images for each of the

Table 1 Comparison of measured roughness data, machining center

	Without feature reduction	With feature reduction	After noise removal
Speed-up (%) compared to SURF		634.8 %	939.6 %
Average features per object	379.3	50.5	46
Matching accuracy	81.33 %	80.00 %	79.67 %
False positives	1.33 %	6.67 %	3.67 %
Average matching time	4914 ms	774 ms	523 ms

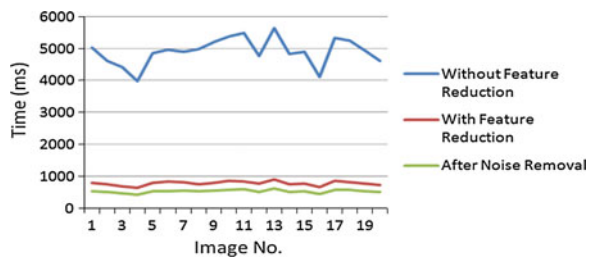
object. One image per object is used for training and three images per object for testing purposes. The same sets of images are used as training and test sets for both the methods under consideration to avoid any influence of partitioning.

The framework has been evaluated based on the speed up, matching accuracy, number of false positives and average number of features per object. The two most important evaluation parameters for object recognition algorithms are matching accuracy and matching time [10]. Table 1 summarizes the comparative results of our scheme and original SURF. By reducing the features, a speed-up of 634.8 % was achieved with <2 % reduction in matching accuracy. The speed up is achieved because the average number of features extracted for each object has been reduced significantly (more than 6 times). However, the number of false positives increases because of feature reduction step. The additional step of noise reduction eliminates the irrelevant features which pushes the false positive rate by about 50 %. Moreover, the noise reduction step results in an overall speed up of 939.6 % as compared to the original SURF. Again, the matching accuracy is not being significantly affected.

Figure 2 shows the comparison of matching time between original SURF, SURF with feature reduction and SURF with feature and noise reduction for 20 images. The graph clearly validates the substantial decrease in matching time by use of our framework.

The results show the efficacy of our scheme in speeding up the object recognition process using SURF features. The experimental results show that our technique provides excellent speed up at the loss of little accuracy. An additional overhead in our framework is the additional time taken in training because of feature and noise reduction. However, most of the times, the training is done offline and thus an increase in training time does not significantly affect the performance of overall system.

Fig. 2 Matching time comparison for 20 images



4 Conclusions

In this paper we present a framework for efficient and reliable object recognition and pose estimation using SURF features. The process of feature reduction is presented which selects only the most repeatable features from an image, allowing a substantial matching speed-up. The presented noise removal process significantly reduces the false positive rates and provides further speed up. The tree-like hierarchal database organization method presented in this paper allows for pose estimation of objects. Work is underway on two fronts (1) to use multi resolution analysis techniques during the learning phase to improve recognition of objects in low resolution images, (2) to incorporate hierarchal generative process [11] in recognition of objects and actions.

Acknowledgments This research was supported by: (1)The Industrial Strategic technology development program, 10041772, (The Development of an Adaptive Mixed-Reality Space based on Interactive Architecture) funded by the Ministry of Knowledge Economy(MKE, Korea), (2) The MKE(The Ministry of Knowledge Economy), Korea, under IT/SW Creative research program supervised by the “NIPA(National IT Industry Promotion Agency)” (NIPA-2012-H0502-12-1013).

References

1. Ejaz N, Tariq TB, Baik SW (2012) Adaptive key frame extraction for video summarization using an aggregation mechanism. *J Vis Commun Image Represent* 23(7):1031–1040
2. Ejaz N, Manzoor U, Nefti S, Baik SW (2012) A collaborative multi-agent framework for abnormal activity detection in crowded areas. *Int J Innov Comput Inform Control* 8(6): 4219–4234
3. Ejaz N, Lee JW, Kim W, Lim C, Joo S, Baik SW (2012) Automated selection of appropriate advertisements for digital signage by analyzing crowd demographics, Special issue on computer convergence technologies. *Inform Int Interdiscip J* 15(5):2019–2030
4. Bay H, Tuytelaars T, Gool LV (2008) Speeded-up robust features (SURF). *Comp Vis Image Underst* 110(3):346–359
5. Lowe DG (1999) Object recognition from local scale-invariant features, Proceedings of the international conference on computer vision, pp 1150–1157
6. Lowe DG (2004) Distinctive image features from scale-invariant key points. *Int J Comp Vis* 60(2):91–110
7. Lee S, Kim K, Kim JY, Kim M, Yoo HJ (2010) Familiarity based unified visual attention model for fast and robust object recognition. *Pattern Recognit* 43(3):1116–1128
8. MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. In: Cam LML, Neyman J (eds) In proceeding of the Berkeley symposium on mathematical statistics and probability, vol 1. University of California Press, California, pp 281–297
9. Nistér D, Stewénius H (2006) Scalable recognition with a vocabulary tree. In: Proceeding of IEEE conference on computer vision and pattern recognition, vol 2. pp 2161–2168
10. Bashir F, Porikli F (2006) Performance evaluation of object detection and tracking systems. In: Proceeding of international workshop on performance evaluation of tracking and surveillance
11. Lin L, Wu T, Porway J, Xu Z (2009) A stochastic graph grammar for compositional object representation and recognition. *Pattern Recognit* 42(7):1297–1307

Automatic Segmentation of Region of Interests in MR Images Using Saliency Information and Active Contours

Irfan Mehmood, Ran Baik and Sung Wook Baik

Abstract Magnetic resonance imaging (MRI) is the most clinically used and gifted modality to identify brain abnormalities in individuals who might be at risk for brain cancer. To date, automated brain tumor segmentation from MRI modalities remains a sensitive, computationally expensive, and a demanding task. This paper presents an automated and robust segmentation method to enable investigators to make successful diagnosis and planning of radiosurgery by reducing the risk factor and study duration. The proposed system consists of following steps: (1) remove the non-brain part from MRI, (2) estimate saliency map of MRI, (3) use the salient region (tumor) as an identification marker and segment the salient object by finding the “optimal” closed contour around the tumor. The system has been tested on real patient images with excellent results. The qualitative and quantitative evaluations by comparing with ground truths and with other existing approaches demonstrate the effectiveness of the proposed method.

Keywords Visual saliency · MRI · Tumor detection · Active contours

I. Mehmood · S. W. Baik (✉)
College of Electronics and Information Engineering, Sejong University,
Seoul, Republic of Korea
e-mail: sbaik@sejong.ac.kr

I. Mehmood
e-mail: irfanmehmood@sju.ac.kr

R. Baik
College of Business, Honam University, Gwangju, Republic of Korea
e-mail: baik@honam.ac.kr

1 Introduction

The brain tumor is an abnormal growth of cells in the brain which results in the destruction of healthy cells with the passage of time. The incident rate of brain tumors is increasing in human beings [1]. The life expectancy of brain tumor depends on many factors which includes early diagnosis. The usage of computer aided systems has obtained considerable popularity due to increasing amount of digital data and ease of accessibility [2]. In order to assist early diagnosis of brain tumors, the usage of computer aided diagnosis (CAD) systems has obtained considerable popularity. An important component of such CAD systems for brain tumor detection is the segmentation of brain tissues from the medical images. The automated segmentation of brain tumor from MRIs is a daunting task owing to the involvement of various research disciplines, varying shapes and sizes of tumors, tendency of tumors to appear at various locations, and appearance in different image intensities [3, 4].

The active contours have been frequently used in estimating the boundaries of a segmented object [5]. The active contours estimate the boundary by using the principal of minimizing internal and external energies. However, a limiting factor in the usage of active contours is that they are sensitive to initialization information. This initialization information is provided by the users, usually in the form of a line drawn on maximum diameter of the tumor. In order to resolve these limitations, researchers have attempted to automatically determine the initialization seed. The choice of initialization seed heavily affects the quality of segmentation and the computational time.

A brief introduction of segmentation in medical imaging is provided by Angelini et al. [6]. Most of the techniques for brain tissue segmentation presented in the literature are ineffective for accurate segmentation. Instead of interactive segmentation methods presented in literature [7–9], we present a fully automatic mechanism to estimate the effective seed point for active contours for the tumor segmentation.

2 Methodology

In this section a detailed description of the brain tumor segmentation framework is presented. There are three sections: pre-processing, calculating saliency map, and finding the outlines of tumor regions as shown in Fig. 1. The details of each module are discussed in the following sections.

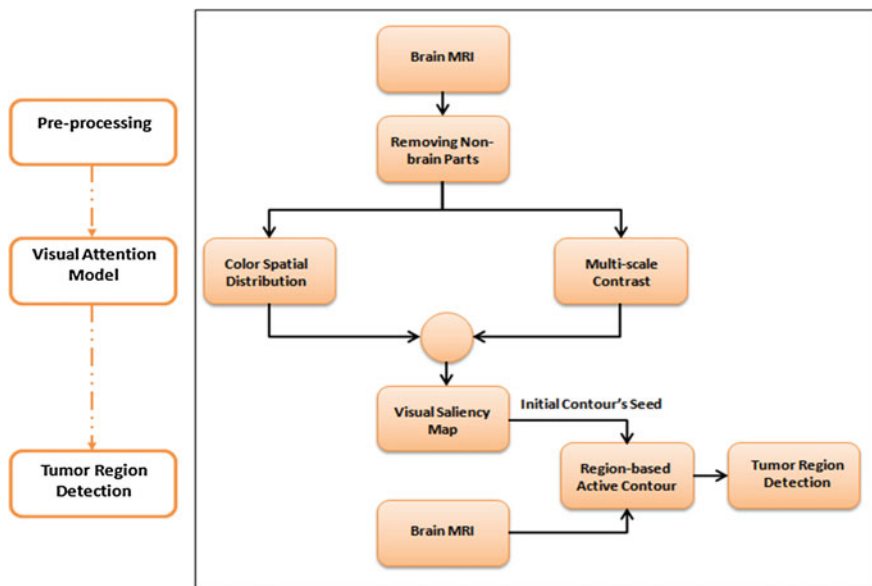


Fig. 1 Framework of proposed system

2.1 Pre-Processing

It is a very tedious job to extract brain parts manually. This task becomes even more difficult while dealing with high resolution images. In pre-processing step, we use algorithm of Boesen et al. [10] (also called McStrip) for removing non-brain parts from MRI images. This step is essential for affectively locating the tumor region using visual attention model.

2.2 Visual Attention Model

While analyzing medical images, clinicians' focus of interest is always on salient objects such as tumor and abnormal spots. Blackwell [11] gave a detailed description that Rod sensors in human eye are 100 times more sensitive to luminance than color sensitive cones. To use this ability of human visual process, contrast is calculated at multiple scales to accurately detect varying tumor size in the patient's brain. In the proposed work, we calculate multi-scale contrast using Gaussian image pyramid as explained below:

$$M_c^l = \sum_{p \in N} \left\| M^l(p) - M^l(p') \right\|^2 \quad (1)$$

Where ‘M’ is the MRI image, ‘N’ is the 5 x 5 neighbourhood around pixel ‘p’, and ‘l’ is the Gaussian pyramid decomposition level which varies from 1 to 3. The resultant multi-scale contrast is calculated by linearly combining contrast values at image decomposition level and normalized in the range [0 1].

$$M_c = \sum_{l=1}^3 M_c^l \quad (2)$$

$$M_{C(Norm)} = \left(\frac{M_c}{\max(M_c)} \right) \quad (3)$$

The contrast is a local feature which is used to measure saliency efficiently. We incorporate color spatial-distribution as a global feature. It has been observed that for salient regions color distribution is small compared to non-salient regions like background. Thus, spatial distribution of color is an important candidate for estimating salient areas in the image. For this purpose we first quantize image ‘M’ into five bands because brain MRIs consist of white matter, gray matter, cerebrospinal fluid, background and tumor tissues. The variance of spatial distribution of each pixel in horizontal direction is calculated as follows:

$$V_{h(c)} = \frac{\sum_{i=1}^m \sum_{j=1}^n (M_{(i,j)}(c) - \mu_h(c))^2}{m * n} \quad (4)$$

where ‘m’ and ‘n’ are indices of pixels having color ‘c’ and $\mu_h(c)$ is the mean of color ‘c’ in horizontal direction. Similarly, variance of color spatial distribution $V_v(c)$ is computed along vertical direction. The spatial variance of each color bin is defined as

$$SV(c) = V_{h(c)} - V_{v(c)} \quad (5)$$

The color spatial-distribution is estimated as

$$SD = \sum_c P(C/I_{(x,y)}) * (1 - SV(C)) \quad (6)$$

Finally, visual saliency map is calculated by linearly combining the multi-scale contrast and color spatial distribution:

$$V_A = M_{C(Norm)} + SD \quad (7)$$

2.3 Segmenting Salient Regions

While dealing with the detection of brain tumor using saliency information, we need to segment a specific target object “optimally”. Region-based models cannot detect a single object accurately from a complex background, because they classify

pixels into two categories (clusters) with one cluster having minimum and other maximum of the image. Although there exist some region-based techniques which efficiently segments object from complex background but they are computationally costly.

The choice of initialization seed for active contour models heavily affects the quality of segmentation and the computational time. The problem with the existing active contours is that they are sensitive to initialization information (seed). To deal with this problem, interactive methods are used to determine the seed point for segmentation. This is done by providing some label points or identification marker to manually assign the seed. In order to resolve these limitations, researchers have attempted to automatically determine the initialization seed. For this purpose saliency map computed in above section is used for initialization of active contours. We used a region-based active contours model proposed by Wang and Wu [12] by initializing its seed using visual saliency model. This algorithm uses two-phase approach having only a single cluster in its evaluation function. Thus, it is more robust and faster compared to other existing region-based contours.

3 Experiments and Result

The proposed method is tested on two different datasets as described below:

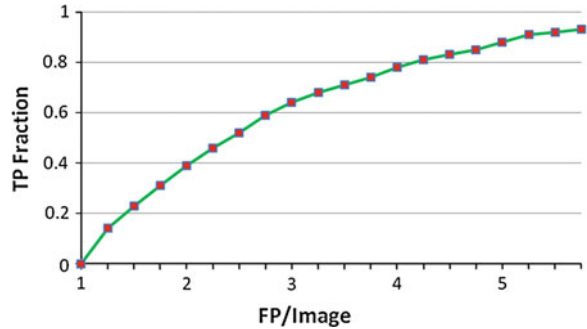
- Dataset for brain MRI analysis available at Harvard medical school [13]. It is a manually segmented dataset which helps in performance comparison.
- Real dataset obtained from a hospital [14] during radiosurgery treatment. This is a labelled MRI dataset of ten patients that was used as ground truth for validation.

In order to analyze the relative performance of the proposed technique and the state of the art brain tumor segmentation algorithms, we have compared our algorithm with Tumor-cut [7] and Grow-cut [8].

3.1 Segmentation Accuracy Analysis

In this set of experiments, the detection of tumor was our primary interest. In order to thoroughly test our technique, we apply our technique on real and synthetic data of brain MRIs with varying tumor sizes and shapes. Free receiving operating characteristic (FROC) methodology is used to compare proposed method with ground truth. We conduct a simple case study where a group of radiologists were asked to score each patient's tumor segmentation from [1–8, 10, 12]. After testing 200 MR brain images with tumors, we have obtain lesion localization fraction (overlap on gold standard) greater than 0.92/image and detection rate of 98 % as

Fig. 2 FROC curve operated for detecting tumor regions in brain MRI. FP is the false positive and TP is true positive on x and y-axis, respectively



shown in Fig. 2. These results are compatible with the segmentation method proposed by Hamamci et al. [7] which needs an input from user to initialize the contour.

Dice overlap with 10 different seed initializations was also measured to compare the accuracy of proposed method with Grow-Cut and Tumor-Cut methods as shown in Table 1. The affect of different initializations are indicated by the extent of standard deviations. The average Dice Overlap for clinical dataset is 84.24 ± 1.5 % for the proposed algorithm which is higher compared to Dice Overlap obtained by Grow-Cut and Tumor-Cut algorithms.

3.2 Stability Analysis

The experiment is conducted on 20 brain MRI having tumors in order to evaluate the impact of initialization on the stability of segmentation using active contours. In this experiment, for each MRI scan, we calculate contour curve around the tumor boundary using proposed, Tumor-cut [7] and Grow-cut [8] methods by

Table 1 Dice overlap (%) computed with ten different seeds for each tumor case

	Grow-cut Dice overlap (%)	Tumor-cut Dice overlap (%)	Proposed method Dice overlap (%)
Case 1	60 ± 5.5	76.8 ± 4.5	84.6 ± 2.8
Case 2	78.4 ± 3.1	88 ± 5.5	90 ± 6
Case 3	74.9 ± 0.8	85.3 ± 2.5	88 ± 1.8
Case 4	78 ± 2.9	78.5 ± 0.8	80 ± 2.3
Case 5	69 ± 4.5	86 ± 2	82 ± 1
Case 6	73.9 ± 1.9	83.7 ± 1.9	83 ± 3.4
Case 7	80.5 ± 1.9	80.8 ± 2.6	85.7 ± 0.6
Case 8	66 ± 2.6	85 ± 1.2	82.8 ± 5.8
Case 9	71.2 ± 2.2	80.4 ± 2.4	79 ± 2.6
Case 10	66.4 ± 3.7	73 ± 4.1	87 ± 2.1
Average	72.00 ± 2.5	81.93 ± 2.3	84.24 ± 1.5

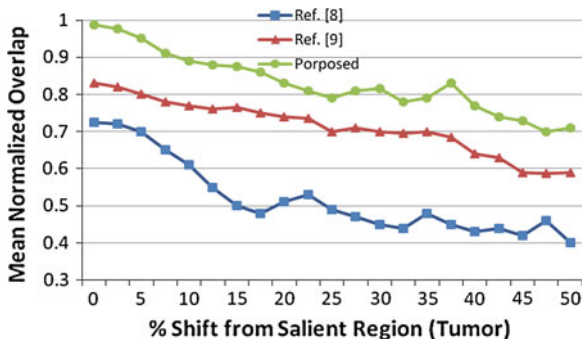


Fig. 3 Accuracy of segmentation region varies with change in distance of seed initialization from salient region

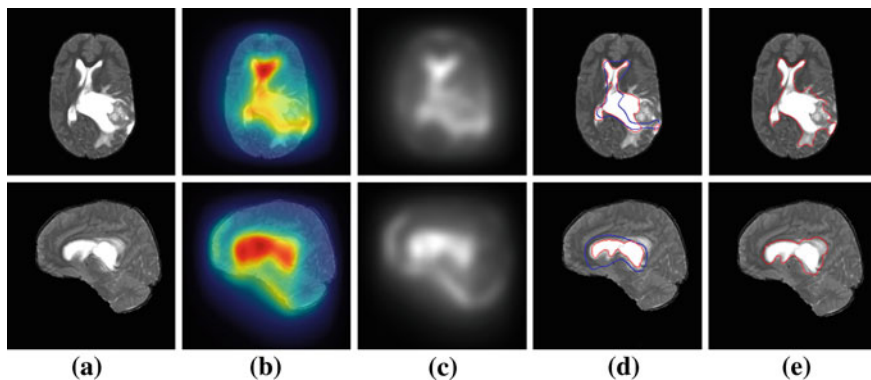


Fig. 4 Tumor detection from brain MRI scan **a** Brain Part after removing non-brain parts using McStrip [10]. **b, c** Heat and Intensity saliency map obtained using proposed model. **d** Detection of tumor region, where *blue curve* is the initialization of active contours using saliency map and *red curve* is the final contour boundary obtained from proposed method. **e** *Red curve* is the final contour boundary obtained from Tumor-Cut [7] algorithm

varying the initialization distance from 0-50 %. The segmentation results obtained from varying initialization are shown in Fig. 3 below. Some visual results (shown in Fig. 4) indicate that the accuracy of tumor region segmentation of the proposed methods is much better than other existing methods.

4 Conclusions

We have presented an automatic and fast segmentation technique for brain tumor detection that uses a visual attention model. This visual attention map works as an identification marker and helps in correctly initializing the active contour to

segment region of interest in minimal span of time with high accuracy. The experimental results show that proposed method is invariant to shape and size of tumor and that it provides robust and efficient detection of salient regions in brain MR scans. The accuracy of segmentation rate indicates that the proposed method is helpful as an effective tool to assist radiologists in diagnosis of brain. This is completely unsupervised method as compared to other existing methods like Tumor-cut method, in which user interaction is necessary to draw one line for each tumor. We hope that the false positives can be reduced by using more sophisticated features in measuring saliency map.

Acknowledgments This research is supported by: (1) The Industrial Strategic technology development program, 10041772, (The Development of an Adaptive Mixed-Reality Space based on Interactive Architecture) funded by the MKE (Ministry of Knowledge Economy, Korea) and, (2) The Ministry of Knowledge Economy (MKE), Korea, under IT/SW Creative research program supervised by the National IT Industry Promotion Agency (NIPA)” (NIPA-2012-H0502-12-1013).

References

1. <http://www.brainumor.org/>
2. Ejaz N, Tariq TB, Baik SW (2012) Adaptive key frame extraction for video summarization using an aggregation mechanism. *J Vis Commun Image Represent* 23(7):1031–1040
3. Cabezas M, Oliver A, Lladó X, Freixenet J, Cuadra MB (2011) A review of atlas-based segmentation for magnetic resonance brain images. *J Comput Methods Programs Biomed* 104(3):e158–e177
4. Boldrey E (1949) A survey of brain tumors for the general practitioner of surgery. *Am J Surg* 78(3):340–346
5. Wang X, Pang Q (2011) The research on segmentation of complex object. *Int Congr Image Signal Process (CISP)* 3:1177–1281
6. Angelini ED, Clatz O, Emmanuel M, Konukoglu E, Capelle L, Duffau H (2007) Glioma dynamics and computational models: a review of segmentation, registration, and in silico growth algorithms and their clinical applications. *J Curr Med Imaging Rev* 3(4):262–276(15)
7. Hamamci A, Kucuk N, Karaman K, Engin K, Unal G (2012) Tumor-cut: segmentation of brain tumors on contrast enhanced MR images for radiosurgery applications. *IEEE Trans Med Imaging* 31(3):798–804
8. Vezhnevets V, Konouchine V (2005) “GrowCut”—interactive multi-label N-D image segmentation by cellular automata. Presented at the Graph-icon, Novosibirsk Akademgorodok
9. Jaffer A, Zia S, Latif G, Mirza AM, Mehmood I, Ejaz N, Baik SW (2012) Anisotropic diffusion based brain MRI segmentation and 3D reconstruction. *Int J Comput Intell Syst* 5(3):494–504
10. Boesen K, Rehm K, Schaper K, Stoltzner S, Woods R, Lüders E, Rottenberg D (2004) Quantitative comparison of four brain extraction algorithms. *J Neuroimage* 22(3):1255–1261
11. Blackwell HR (1946) Contrast thresholds of the human eye. *J Opt Soc Am* (1917–1983) 36(11):624–632
12. Wang XT, Wu JT (2012) Active contours for specific target detection. *Electron Lett* 48(2):83–84
13. Harvard Medical School <http://med.harvard.edu/AANLIB/>
14. Pakistan Institute of Medical Sciences <http://www.pims.gov.pk/radiology.htm>

Digital Image Magnification Using Gaussian-Edge Directed Interpolation

Muhammad Sajjad, Ran Baik and Sung Wook Baik

Abstract This paper presents a simple and cost effective approach for digital image magnification (DIM). DIM is used in various applications and is an enthusiastic area of research at present. The proposed technique uses Gaussian edge directed interpolation to determine the precise weights of the neighboring pixels. The standard deviation of the interpolation window determines the value of ' σ ' for generating Gaussian kernels. Gaussian kernels preserve the original detail of the low-resolution image to produce high-resolution image of high visual quality. The experimental results show that the proposed technique is superior to other techniques qualitatively as well as quantitatively.

Keywords Digital image magnification · Gaussian kernel · Gaussian sigma · Weighted interpolation

M. Sajjad · S. W. Baik (✉)
College of Electronics and Information Engineering, Sejong University,
Seoul, Republic of Korea
e-mail: sbaik@sejong.ac.kr

M. Sajjad
e-mail: sajjad@sju.ac.kr

R. Baik
College of Business, Honam University, Gwangju, Republic of Korea
e-mail: baik@honam.ac.kr

1 Introduction

Digital image magnification (DIM) means to take low resolution image as an input and generate a corresponding high resolution image. The terms zooming, super resolution (SR) and scaling etc. are also used interchangeably for the same process. Image magnification aims to compute high resolution image containing accurate and precise details of the original low resolution image. Magnification technique must preserve the sharp luminance information, texture, geometrical invariance, and smoothness of the original image while producing high resolution image from the source image. DIM has various applications in many areas e.g. high-definition television (HDTV), medical Imaging, surveillance system, satellite-imaging, and entertainment etc. A variety of techniques can be used to produce SR image. The reconstruction based DIM techniques [12–16] need prior information to model the zooming scheme. This information can be acquired by down-sampling the HR images to low-resolution (LR) images. In order to construct the high resolution image, reconstruction DIM is considered as the inverse problem by adding one or more LR images. There are various other image zooming schemes based on machine learning approaches [15–20]. Therefore a set of LR and their corresponding patches of HR image are stored in database. To construct HR image from given LR image, a patch from LR input image and their corresponding HR is used for the magnification. Machine learning approaches have high time complexity and impossible to implement in real time systems. Interpolation based DIM techniques [7–10] are used most often. Such techniques take a single image as an input and produce HR image. The common interpolation techniques are Nearest Neighbor (NN), Bilinear (BL) and Bicubic (BC). These techniques are non-adaptive. The time complexity of these non-adaptive techniques is low but these techniques introduce some unwanted artifacts across the edges in the image. There are some edge-directed interpolation schemes [11] which are adaptive and preserve original information of the source image. Moreover, these schemes perform the interpolation in selective direction using the geometrical and structural information of the image. Still BC interpolation scheme is used most frequently due its low time complexity and good visual results. [1] proposed a zooming technique using block-expanding method based on intervals which associate each pixel with an interval acquired by a weighted combination of the pixels in the vicinity of the neighborhood.

In this paper, our objective is to develop an efficient magnification technique which, construct HR image from given source LR image. The idea of the proposed technique is based on the calculation of the value of the unknown pixels from given known pixels information in the neighborhood by weighted approximation. For assigning correct weight to each pixels in the neighborhood [1, 21, 22], resultant image of HR will be of high quality. To use this concept, the proposed technique uses two type Gaussian kernels, generated by two different values of standard deviation of Gaussian Function. The standard deviation of the neighborhood pixels determines which Gaussian kernels to be used.

The proposed technique is computationally simple and produces good results quantitatively as well as qualitatively. For this purpose it has been compared with contemporary techniques as discussed in detail in [Sect. 3](#)

2 Proposed Method

Suppose S_{in} is the source image of size $r \times c$ and λ is a magnification factor. The source image is first expanded by factor λ where $\lambda = \{2, 4, 8\}$. S_{in} is mapped from a smaller pixel grid to larger pixel grid i.e. $S_{out}(R,C) = S_{in}(r \times \lambda, c \times \lambda)$ where $R \times C$ is the size of expanded image. After expansion, S_{out} has undefined pixels which are interpolated from already known pixels. The magnified image of high quality is obtained by convolving a kernel $\Phi_{Gk}(s,t)$ on S_{in} .

$$S_{out}(x, y) = \sum_{s=i} \sum_{t=j} S_{in}(x + s, y + t) \Phi_{Gk}(s, t) \tag{1}$$

$\Phi_{Gk}(s,t)$ is the Gaussian kernel of size $s \times t$. k denotes type of Gaussian’s kernel. In order to get the complete zoomed image of high quality Eq. (1) must be convolved for $x = 1, 2 \dots R$ and $y = 1, 2 \dots C$. The proposed technique uses two type of Gaussian kernels depending on the luminance information of the interpolation window. The size of the interpolation window is also of size $s \times t$. Initially the standard deviation σ_w of the interpolated window is calculated according to:

$$\sigma_w = \sqrt{\frac{\sum_{s=i} \sum_{t=j} (\bar{X} - S_{out}(x + s, y + t))^2}{S_{xt}}} \tag{2}$$

where \bar{X} is the mean of interpolated window. The Gaussian Kernel is generated with $\sigma_{Gk} = 0.3$ of size $s \times t$, if σ_w is greater than threshold τ otherwise it will be generated with $\sigma_{Gk} = 0.8$ of the same size. The Gaussian kernel [4–6] is defined as:

$$\Phi_{Gk}(s, t) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{d^2}{2\sigma_{Gk}}} \tag{3}$$

where d^2 is the distance of neighbor pixels (s_i, t_j) from the center pixel (s_c, t_c) . The σ_{Gk} defines the width i.e. the degree of smoothness, of a Gaussian Kernel $\Phi_{Gk}(s,t)$. In case of sharp luminance, we use the kernel with $\sigma_{Gk} = 0.3$ and for smoothness where the interpolation region is constant, we use the kernel with $\sigma_{Gk} = 0.8$. The values for $\sigma_{Gk} = \{0.3,0.8\}$ have been chosen after a number of tests on smoothed, texture and sharpened luminance interpolated regions of the images. σ_{Gk} with value 0.3 gives good result on sharpened luminance interpolated area while maintaining the texture of the image and σ_{Gk} with value 0.8 has better-quality for smooth

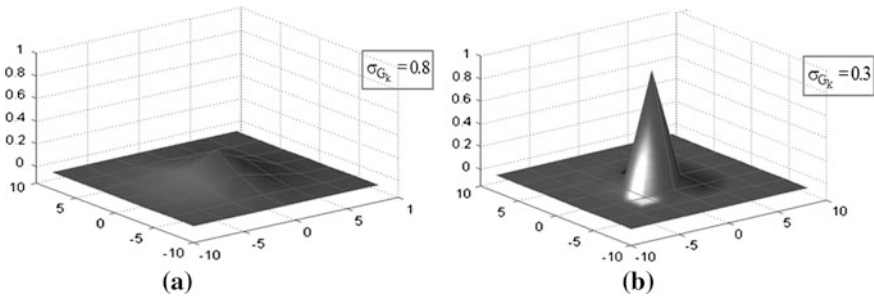


Fig. 1 The two-dimensional Gaussian function with $\sigma_{Gk} = 0.3$ and $\sigma_{Gk} = 0.8$

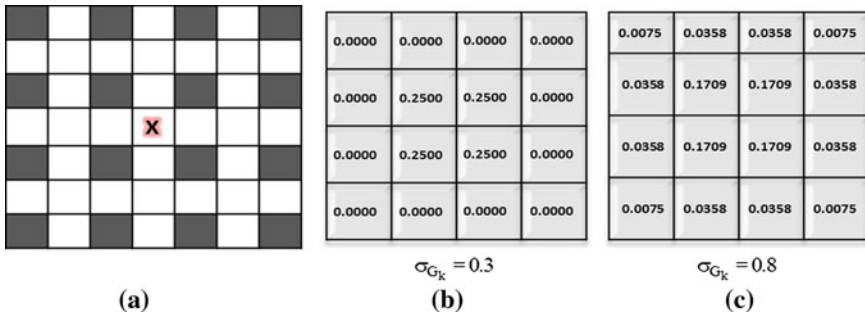


Fig. 2 **a** Example of interpolation window in expanded image **b** is the Gaussian kernel with $\sigma_{Gk} = 0.3$ **(c)** is the Gaussian kernel with $\sigma_{Gk} = 0.8$

interpolated region. It preserves not only the smoothness of the image but also preserves the texture property of the image inside smooth region. The Gaussian function has been shown with both sigma's values in Fig. 1. The Gaussian kernels $\Phi_{Gk}(s,t)$ of size $s \times t$ generated with $\sigma_{Gk} = 0.3$ and $\sigma_{Gk} = 0.8$ are shown in Fig. 2. Figure 2a shows the interpolation window in expanded image in the form of grid of size $S \times T$. The dark gray blocks in grid represent the known pixels and white blocks in the grid stand for unknown pixels. The block contain 'X' is the interpolated pixel. These both kernels convolve to the known pixels inside the interpolation region and sum of the convolved pixels is assigned to the block contain 'X'.

Figure 3 shows the system model of the proposed technique. The proposed technique take gray scale image as an input and produce the zoomed image by factor $\lambda = 2$. It also work same for the color image. Let \mathcal{F}_{rgb} be the color source input image of size $r \times c \times z$ where z stands for RGB color channels. We split this color image \mathcal{F}_{rgb} into RGB channels e.g. \mathcal{F}_r , \mathcal{F}_g and \mathcal{F}_b . On each separate color channel (gray-scaled image) the proposed technique is applied independently with magnification $\lambda = 2$ and produced the magnified images of each color channel \mathcal{F}_R , \mathcal{F}_G and \mathcal{F}_B of size $R \times C$. For combining these three \mathcal{F}_R , \mathcal{F}_G and \mathcal{F}_B images we get the zoomed color image \mathcal{F}_{RGB} of size $R \times C \times Z$.

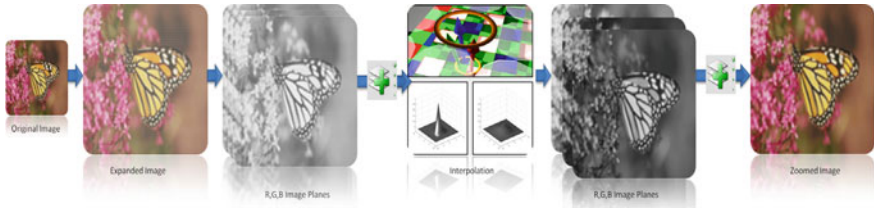


Fig. 3 System model of the proposed magnification technique

3 Experimental Results

The proposed technique has been evaluated qualitatively as well as quantitatively. For this purpose we use a set of images from public database provided by [3]. All the images were of size $M \times N$ where $M = N = 512$. The images were reduce by factor $\lambda = 1/4$ and magnified to its original size using different zooming techniques. Let O_{MN} be the original image obtained from public database and T_{MN} be the magnified image after reducing it by $\lambda = 1/4$. The proposed technique has been compared with other three techniques, one recent technique ‘InInfo’ [1] and other two are well-known classic interpolation techniques i.e. Bilinear (BL) and Bicubic (BC) interpolation techniques [2]. The program for InInfo was provided by the author and for bilinear and bicubic we used already existing implementations in MATLAB R2011b. In order to evaluate that how much the magnified image T_{MN} is similar to original image O_{MN} , we assess the similarity using Peak Signal to Noise ratio (PSNR) [1] and Mean Squared Error (MSE) as:

$$PSNR = 20 * \log_{10} \frac{255}{MSE} \tag{4}$$

$$MSE = \frac{\sum_{x=v,y=t} (O_{MN}(x_v, y_t) - T_{MN}(x_v, y_t))^2}{MN} \tag{5}$$

PSNR is the ratio between the strength of the maximum achievable signal and corrupting noise. The noise is the error introduced during zooming process and the signal is the original pixel’s intensity information. The greater value of the PSNR notifies good quality of zooming technique and vice versa. The results have been shown in Table 1. The proposed technique has the highest PSNR values which show the superiority of the proposed technique over other techniques.

Equation (5) shows how to calculate the MSE. The lower values of MSE describe high quality of zooming technique and vice versa. Figure 4 shows the MSE results calculated over ten standard images. The overall results of the proposed technique are better than other zooming technique. These both quantitative result shows that the proposed technique calculates the unknown value close to actual value. This is because of usage two separate Gaussian kernels adaptively.

Table 1 Comparison of proposed technique with other zooming techniques by calculating PSNR over 10 standard images. The zooming factor is $\lambda = 4$

No.	Image Name	BL	BC	InInfo	Proposed
1	Harbor	25.10231	26.44346	24.45197	28.35325
2	Cameraman	23.45222	22.42631	20.34227	22.67422
3	House	27.86343	28.34538	26.45323	30.45236
4	Avion	33.7638	33.23482	31.34201	35.3435
5	Baboon	22.4345	22.27906	20.54329	23.34295
6	Athens	27.86542	28.52971	26.43592	29.67568
7	Barbara	24.6433	23.2433	21.56237	25.8127
8	Boat	25.36196	26.29832	22.36509	27.47344
9	Peppers	29.54394	28.85262	27.13269	31.43224
10	Lena	27.21796	29.4329	27.43299	31.32439
	Average	26.72488	26.90859	24.80618	28.58847

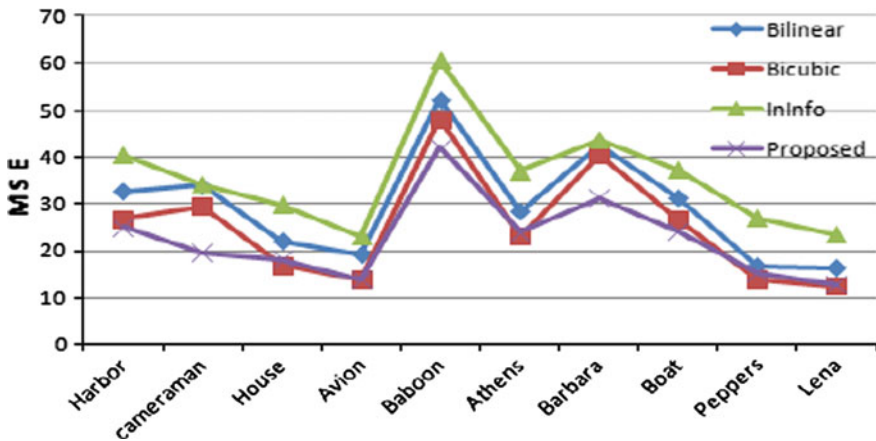


Fig. 4 MSE computed over ten images zoomed by factor $\lambda = 4$ using proposed technique, Bilinear, Bicubic and InInf zooming techniques

In Fig. 5a cropped portion of size 146×54 from the color image Bobcat has been magnified by factor $\lambda = 2$ (Fig. 5b) and by factor $\lambda = 4$ (Fig. 5c). The proposed technique preserved the texture, edges, smoothness, and geometrical variation of the zoomed image to the greater extent which proved the high quality of it. The time complexity of the proposed technique is $O(R,C)$ where R is the number of rows and C is the number of columns. Both qualitative and quantitative results of proposed technique are better than InInfo, Bilinear and Bicubic techniques.

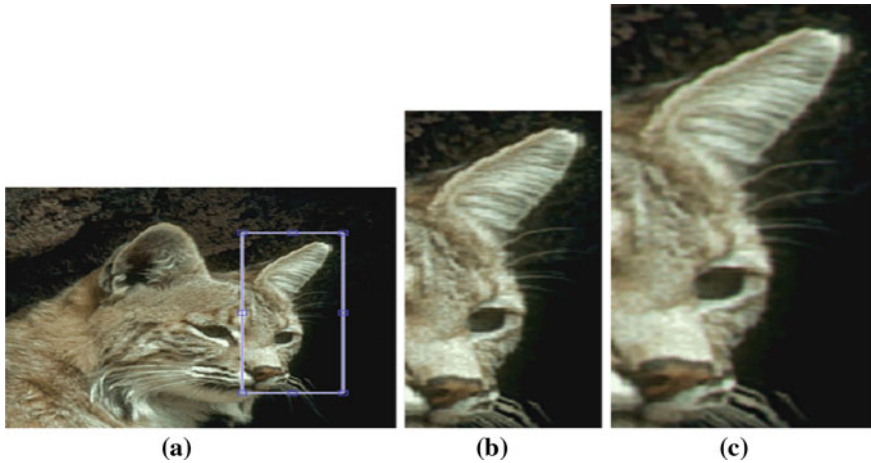


Fig. 5 a Cropped portion of size 194 x 71 of Bobcat, **b** 2 x zoomed, **c** 4 x zoomed

4 Conclusion

In this paper we have presented a DIM scheme via Gaussian edge-directed interpolation method. The technique uses two Gaussian kernels which intelligently calculate the unknown pixels. The standard deviation of the interpolation window decides the type of Gaussian kernel to be used. The Gaussian kernel with $\sigma_{Gk} = 0.3$ works well in edges and preserves the sharp luminance variation across the boundaries. The Gaussian kernel with $\sigma_{Gk} = 0.8$ works well in smooth area of the image and preserve the texture of the image. In this paper we use only two Gaussian kernels. In future, we intend to test the interpolation region for more than two Gaussian kernels with different values of σ_{Gk} to preserve the details of the original image up to greater extent.

Acknowledgments This research is supported by, (1) The Industrial Strategic technology development program, 10041772, (The Development of an Adaptive Mixed-Reality Space based on Interactive Architecture) funded by the Ministry of Knowledge Economy (MKE, Korea), and (2) The MKE (The Ministry of Knowledge Economy), Korea, under IT/SW Creative research program supervised by the NIPA (National IT Industry Promotion Agency)” (NIPA-2012-H0502-12-1013).

References

1. Jurio A, Pagola M, Mesiar R, Beliakov G, Bustince H (2011) Image magnification using interval information. *IEEE Trans Image Process*, 20(11):3112–3123
2. Amanatiadis A, Andreadis I (2009) A survey on evaluation methods for image interpolation. *Meas Sci Technol* 20(10):104015–104021
3. <http://decsai.ugr.es/cvg/dbimagenes/g512.php>

4. Yeon JL, Jungho Y (2010) Nonlinear image upsampling method based on radial basis function interpolation. *IEEE Trans Image Process* 19(10):2682–2692
5. Gonzalez RC, Woods RE (2007) *Digital image processing*, 3rd edn. Amazon
6. Shapiro LG, Stockman GC (2001) *Computer vision*. Amazon
7. Hou HS, Andrews HC (1978) Cubic splines for image interpolation and digital filtering. *IEEE Trans Acoust Speech Signal Proc* 26:508–517
8. Li X, Orchard MT (2001) New edge-directed interpolation. *IEEE Trans Image Process* 10:1521–1527
9. Tam WS, Kok CW, Siu WC (2010) A modified edge directed interpolation for images. *J Electron Imaging* 19(1):1–20
10. Wittman T (2005) *Mathematical techniques for image interpolation*. Department of Mathematics, University of Minnesota
11. Lee YJ, Yoon J (2010) Nonlinear image upsampling method based on radial basis function interpolation. *IEEE Trans Image Process* 19(10):2682–2692
12. Shan Q, Li Z, Jia J, Tang CK (2008) Fast image/video upsampling. *ACM Transactions on Graphics (SIGGRAPH ASIA)* 27:153–160
13. Hung KW, Siu WC (2009) New motion compensation model via frequency classification for fast video super-resolution. *IEEE Int Conf Image Process*
14. Baker S, Kanade T (2002) Limits on super-resolution and how to break them. *IEEE Trans on Pattern Anal Mach Intell* 24:1167–1183
15. Irani M, Peleg S (1993) Motion analysis for image enhancement: resolution, occlusion and transparency. *J Vis Commun Image Represent* 4(4):324–335
16. Mallat S, Yu G (2010) Super-resolution with sparse mixing estimators. *IEEE Trans Image Process* 19(11):2889–2900
17. Gajjar PP, Joshi MV (2010) New learning based super-resolution: use of DWT and IGMRF prior. *IEEE Trans Image Process* 19(5):1201–1213
18. Ni KS, Nguyen TQ (2007) Image super resolution using support vector regression. *IEEE Trans Image Process* 16(6):1596–1610
19. Yang J, Wright J, Huang TS, Ma Y (2010) Image super-resolution via sparse representation. *IEEE Trans Image Process* 19(11):2861–2873
20. Kim KI, Kwon Y (2008) Example-based learning for single image super-resolution and JPEG artifact removal. Technical Report 173, Max Planck Institute
21. Ejaz N, Tariq TB, Baik SW (2012) Adaptive key frame extraction for video summarization using an aggregation mechanism. *J Vis Commun Image Represent* 23(7):1031–1040
22. He H, Siu WC (2011) Single image super resolution using Gaussian process regression. *IEEE Conf Comput Vis Pattern Recognit* 449–456

Part VI
Mobile Computing and Future Networks

An Examination of Psychological Factors Affecting Drivers' Perceptions and Attitudes Toward Car Navigation Systems

Eunil Park, Ki Joon Kim and Angel P. del Pobil

Abstract To examine drivers' perceptions of and attitudes toward car navigation systems, the present study integrated perceived satisfaction and perceived locational accuracy into a modified technology acceptance model, and investigated hypothesized causal paths proposed by the research model with data collected from an online survey (N = 1,204). Results from the structural equation modeling indicated that perceived satisfaction and locational accuracy played crucial roles in determining perceived ease of use and perceived usefulness of the navigation systems. Implications and limitations are discussed.

Keywords Car navigation systems · Technology acceptance model · Perceived satisfaction · Perceived locational accuracy

E. Park (✉)

Interaction Science Research Center, Sungkyunkwan University, Seoul, South Korea
e-mail: pa1324@skku.edu

K. J. Kim · A. P. del Pobil

Department of Interaction Science, Sungkyunkwan University, Seoul, South Korea
e-mail: veritate@skku.edu

A. P. del Pobil

e-mail: pobil@icc.uji.es

A. P. del Pobil

Department of Computer Science and Engineering, University Jaume-I, Castellon, Spain

1 Introduction

Car navigation system is one of the most widely and commercially used global positioning systems. By using the system's map database and satellite information, drivers can receive geographic information such as maps of surrounding areas and directions to designated locations. With this convenience, many drivers have installed car navigation systems in their cars. In accordance with this demand, manufacturers and researchers have studied the systems for improvement. To be selected by potential drivers, the systems were focused on the two factors, geographic information accuracy and user interface of the systems, which are engineering and user satisfaction aspects of the systems [1–3].

Although these factors should be considered to be chosen by drivers, few studies have aimed to investigate driver intentions of car navigation systems [4]. Therefore, the current study examines drivers' intentions of the car navigation systems and their attitudes toward the systems by using the technology acceptance concept.

2 Theoretical Background

2.1 *History of Car Navigation Systems*

The first commercial car navigation system was introduced by Steven Lobbezoo [5] and presented at a technology fair in 1985. The first system was composed of three parts: a personal computer, a disc containing map data, and a display screen. Mitsubishi later introduced a more advanced GPS navigation system in 1990, offering real-time maps and a graphic display screen [6].

Since Magellan produced the first commercially sold car navigation system in 1995 in the U.S [7], the market has grown exponentially; iSuppli estimated that the domestic navigation market will increase from 1.6 billion dollars in 2006 to 6.5 billion dollars by 2012 [8].

2.2 *Technological Aspects of Car Navigation Systems*

Three technologies or devices are required for providing drivers with proper navigation services. First, a visualization technology is required. Recent systems offer a bird-eye view and a synchronizing function of the map. A distance measuring function and voice examination function are also equipped. Second, road database and stable storage to save the database are required. Accurate geographic data such as street names and building numbers are saved in the storage. Third, system integration and real-time synchronization are required for saving and retrieving data at the right time [9–12].

3 Hypotheses

3.1 Perceived Ease of Use, Perceived Usefulness, and Attitude

A large number of technology acceptance model (TAM) studies have consistently demonstrated that perceived ease of use and perceived usefulness are crucial determinants of users' attitudes toward an information system [13, 14]. In addition, attitude and perceived usefulness are key determinants of users' intentions to use the system. Furthermore, previous studies have found that perceived ease of use has positive effects on perceived usefulness of the system [15, 16]. Consistent with these findings, we posit the following hypotheses.

H1: Perceived ease of use has positive effects on perceived usefulness.

H2: Perceived ease of use has positive effects on attitude.

H3: Perceived usefulness has positive effects on attitude.

H4: Perceived usefulness has positive effects on intention to use.

H5: Attitude has positive effects on intention to use.

3.2 Perceived Locational Accuracy

Perceived locational accuracy is defined as the degree to which drivers are being aware of their accurate locations. Most navigation systems use an arrow-shaped figure to display cars' locations and directions [17, 18], and they interactively work in real-time to provide drivers with accurate, efficient ways to get to their destinations. Therefore, drivers are likely to perceive that navigation systems with greater locational accuracy is more useful, thereby providing them with greater satisfaction. Based on this logic, we posit the following hypothesis.

H6: Perceived locational accuracy has positive effects on perceived usefulness.

3.3 Perceived Satisfaction

Previous studies have indicated that perceived satisfaction is one of the most crucial factors that determines users' acceptance of a technology. For example, Chiu et al. [19] found that perceived usability of and satisfaction with online systems were positively related to users' acceptance and intention to continue use the systems. In addition, LeBarbera and Mazursky showed that consumer satisfaction was key to improving online shopping systems [20]. By extension, perceived satisfaction with car navigation systems is also likely to have similar effects on drivers' perceptions of the systems. In particular, we hypothesize that

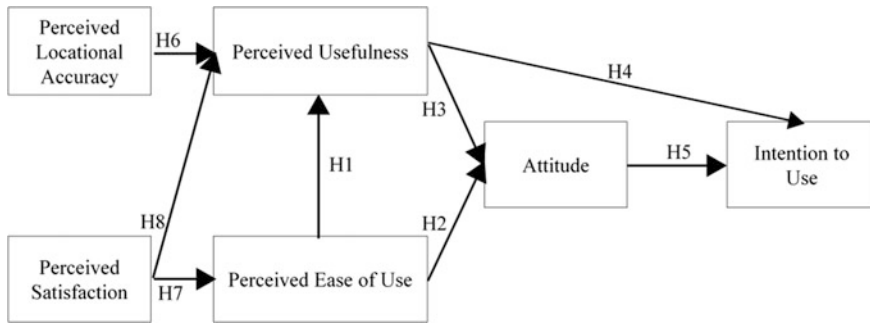


Fig. 1 Proposed research model

H7: Perceived satisfaction has positive effects on perceived ease of use.

H8: Perceived satisfaction has positive effects on perceived usefulness.

3.4 Proposed Research Model

Based on the hypotheses, the present study proposes the following research model (Fig. 1).

4 Method

4.1 Participants

1,204 drivers (71 % male, 29 % female) participated in the survey. The mean age of participants was 26.6 years old (SD = 3.91). All participants had their own cars and driver's license.

4.2 Procedure and Measures

Questionnaire items were developed and modified via three rounds of reviews by an expert group. Results of the pre-test indicated that all items had acceptable reliability.

The main survey was administered via online; it was posted on two online discussion forums for car drivers in South Korea. All participants had at least 24 weeks of experience in driving cars and using car navigation systems. The first page of the survey showed the purpose of the survey. The second page of the survey was used to gather sample demographic information such as gender and age.

All questionnaire items in the survey were adopted from validated previous studies. Perceived locational accuracy was measured with an index consists of three items adopted from Park et al. and Loomis et al.’s studies [17, 21]. Perceived satisfaction was measured with an index consists of three items used by Park and del Pobil [22]. Items measuring the four constructs of the TAM were adapted from the original TAM studies (Perceived ease of use: three items, Perceived usefulness: three items, attitude towards can navigation system: three items, and intention to use: three items) [13–16].

5 Results

Descriptive statistics are reported in Table 1.

5.1 Analysis Method

Structural equation modeling (SEM) was employed to test the proposed model. We used confirmatory factor analysis (CFA) and LISREL 8.70 to evaluate the measurement model. In addition, the maximum likelihood method was used for testing reliability and validity. Our sample size (i.e., N = 1,204) met the recommendation that at least 200 participants are needed for a valid SEM analysis. All constructs also had a strong discriminant validity because correlation between two constructs did not exceed square root of the average variance extracted (AVE) of each construct [23].

5.2 Fit Indices

As reported in Table 2, the measurement model and proposed model showed acceptable fit-indices between them.

Table 1 Descriptive analysis and the reliability of the constructs

Constructs	Mean	Standard deviation	Cronbach’s α
Perceived locational accuracy	5.54	1.13	0.91
Perceived satisfaction	5.82	1.01	0.94
Perceived ease of use	5.81	1.04	0.81
Perceived usefulness	5.58	1.00	0.92
Attitude	5.64	1.10	0.83
Intention to use	5.58	1.17	0.85

Table 2 Fit indices

Fit-indexes	The measurement model	The proposed model	Fit-indexes	The measurement model	The proposed model
$\chi^2/d.f.$	4.74	4.95	CFI	0.92	0.92
GFI	0.93	0.92	SRMR	0.048	0.048
NNFI	0.91	0.91	RMSEA	0.045	0.047

Table 3 Results of hypothesis testing

Hypotheses	Paths	Standardized coefficient	SE	CR	Results
H1	PE → PU	0.005	0.046	0.112	Not supported
H2	PE → ATT	0.487*	0.022	23.066	Supported
H3	PU → ATT	0.396*	0.022	18.732	Supported
H4	PU → IU	0.422*	0.025	18.955	Supported
H5	ATT → IU	0.452*	0.024	20.301	Supported
H6	PLA → PU	0.542*	0.019	26.046	Supported
H7	PS → PE	0.890*	0.013	67.667	Supported
H8	PS → PU	0.430*	0.047	19.428	Supported

* $p < 0.001$, PE: Perceived Ease of use, PU: Perceived Usefulness, ATT: Attitude, IU: Intention to Use, PLA: Perceived Locational Accuracy, PS: Perceived Satisfaction

5.3 Hypothesis Testing

As summarized in Table 3 and Fig. 2, all paths except H1 were supported. Perceived ease of use did not have significant effects on perceived usefulness (H1, $p > 0.1$), while perceived ease of use had effects on attitude (H2, $\beta = 0.49$, $p < 0.001$). Perceived usefulness was positively related to attitude (H3, $\beta = 0.40$, $p < 0.001$) and intention to use (H4, $\beta = 0.42$, $p < 0.001$). In addition, attitude had positive effects on intention to use (H5, $\beta = 0.45$, $p < 0.001$). 61 % of the variance of intention to use was explained by attitude and perceived usefulness.

Confirmation of H6, H7 and H8 demonstrated the effects of two external variables on perceived ease of use and perceived usefulness. Perceived locational accuracy had positive effects on perceived usefulness (H6, $\beta = 0.54$, $p < 0.001$). Perceived satisfaction had significant effects on perceived usefulness (H8, $\beta = 0.43$, $p < 0.001$) and ease of use (H7, $\beta = 0.89$, $p < 0.001$). In addition, perceived locational accuracy and perceived satisfaction explained 48 % of the variance of perceived usefulness. 79 % of the variance of perceived ease of use was explained by perceived satisfaction.

6 Discussion

With two factors and technology acceptance concept, this study examines that perceived locational accuracy and satisfaction were certainly important factors for improving car navigation systems. The results from the SEM found that perceived

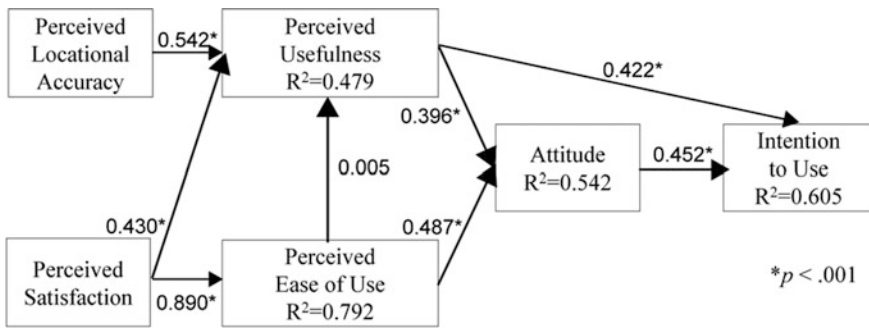


Fig. 2 Summary of hypothesis testing

locational accuracy and the satisfaction significantly affected users’ attitude and acceptance towards the systems.

In addition, the present study revealed similar effects of perceived usefulness and attitude on the intention to use, while the effect of perceived ease of use on attitude was stronger than that of perceived usefulness.

Inconsistency with H1 may be explained by findings from previous studies that focused on user acceptance toward information systems. They already found that the increase in perceived ease of use negatively affected the degree of perceived usefulness [22].

There are some issues which should be considered in future research. First, participants’ characteristics were not considered in this study. As shown in some previous studies, the characteristics of participants can affect results of user acceptance studies [16]. Second, there can be other crucial factors which are able to be related to users’ attitude. Thus, future studies are required to consider these limitations.

Acknowledgments This research was partly supported by WCU program via the NRF of Korea funded by the KMEST (R31-2008-000-10062-0), by Ministerio de Ciencia e Innovacion (DPI2011-27846), by Generalitat Valenciana (PROMETEO/2009/052) and by Fundacio Caixa Castello-Bancaixa (P1-1B2011-54).

References

1. Adrados C, Girard I, Gendner JP, Janeau G (2002) Global positioning system (GPS) location accuracy improvement due to selective availability removal. *CR Biol* 325(2):165–170
2. Kupper A (2005) *Location-based services*. Willey, Chichester
3. Kawasaki H, Murao M, Ikeuchi K, Sakauchi M (2001) Enhanced navigation system with real images and real-time information. In: *Proceedings of the 8th world congress on intelligent transport systems*, pp 1–11
4. Nobuyuki K, Tomohiro Y, Osamu S, Andrew L (2000) A driver behavior recognition method based on a driver model framework. *SAE Trans* 109(6):469–476

5. Autos.ca, Product Review: five in-car navigation systems <http://www.autos.ca/auto-articles/product-review-five-in-car-navigation-systems>
6. Nowakowski C, Utsui Y, Green P (2000) Navigation system destination entry: the effects of driver workload and input devices, and implications for SAE recommended practice (Technical Report UMTRI-2000-20). University of Michigan, Transportation Research Institute, Ann Arbor
7. Magellan, MiTAC International Corporation <http://www.magellangps.com/>
8. iSuppli, Automotive Infotainment & Telematics <http://www.isuppli.com/automotive-infotainment-and-telematics/pages/navigation-telematics-and-digital-entertainment.aspx>
9. Ariyoshi H, Iwaskaki A, Sugihara T, Ohe H, Sakamoto M (1988) Car navigation system. *NEC Tech J* 41:149–159
10. Narzt W, Pomberger G, Ferscha A, Kolb D, Müller R, Wieghardt J, Hörtner H, Lindinger C (2004) A new visualization concept for navigation systems. *Lect Notes Comput Sci* 3196:440–451
11. Robertson DP, Cipolla R (2004) An image-based system for urban navigation. In: *Proceedings of the 2004 BMVC*
12. Gilliéron PY, Merminod B (2003) Personal navigation system for indoor application. In: *Proceedings of the 11th IAIN world congress*, pp 1–15
13. Davis F (1989) Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q* 13(3):319–340
14. Davis F (1993) User acceptance of information technology: system characteristics, user perceptions and behavioral impacts. *Int J Man Mach Stud* 38:475–487
15. Koufaris M (2003) Applying the technology acceptance model and flow theory to online consumer behavior. *Inf Syst Res* 13(2):205–223
16. Venkatesh V, Davis FD (2000) A theoretical extension of the technology acceptance model: four longitudinal field studies. *Manage Sci* 46:186–204
17. Park E, Kim KJ, Jin D, del Pobil AP (2012) Towards a successful mobile map service: an empirical examination of technology acceptance model. *Commun Comput Inf Sci* 293:420–428
18. Park E, Kim S, del Pobil AP (2011) Can I go there? The effects of digital signage on psychology of wayfinding users. *J Next Gener Inf Technol* 2(4):47–58
19. Chiu CM, Hsu M, Sun S, Lin T, Sun P (2005) Usability, quality, value and e-learning continuance decisions. *Comput Educ* 45:399–416
20. LaBarbera P, Mazursky D (1983) A Longitudinal assessment of consumer satisfaction/dissatisfaction: the dynamic aspect of the cognitive process. *J Mark Res* 24:393–404
21. Loomis JM, Silva JA, Philbeck JW, Fukusima SS (1996) Visual perception of location and distance. *Curr Dir Psychol Sci* 5:72–77
22. Park E, del Pobil AP (2012) Modeling the user acceptance of long-term evolution (LTE) services. *Ann Telecommun* 1–9
23. Hair JF, Black WC, Babin BJ, Anderson RE (2006) *Multivariate data analysis*. Prentice Hall, Upper Saddle River

Handover Performance Evaluation of 4G Mobile Networks

Baik Kim and Ronny Yongho Kim

Abstract Since the core functionality of cellular network mobility management is handover and handover latency directly impacts service continuity and the end-user experience, handover has the most stringent latency requirement on service interruption time compared to other mobility related operations such as reentry from Idle mode. This paper presents the state-of-the-art handover schemes designed for IEEE 802.16m based 4G mobile networks (next generation WiMAX), approved by ITU as an IMT-Advanced technology and handover performance of 4G mobile networks is evaluated in terms of latency. A system-architectural view on handover is provided to analyze the causes of handover latency and methods to optimize them for latency reduction. Through handover latency analysis an insight on how well 4G handover scheme is designed can be provided.

Keywords Mobility management · Performance evaluation · 4G mobile networks · IMT-advanced

1 Introduction

As smartphones emerge, nowadays diverse services via wireless communication have become essential part of people's daily life. With expectation of the fast growth of mobile computing users/devices as well as higher demands for capacity and features of wireless technology, International Telecommunication Union

B. Kim · R. Y. Kim (✉)

Department of Railroad Electrical and Electronics Engineering, National University of Transportation, 157 Cheoldo Parkmulgwan-ro, Uiwang, Gyeonggi, South Korea
e-mail: ronnykim@ut.ac.kr

Radio Section (ITU-R) has commenced the process of developing ITU-R Recommendations for the terrestrial components of the International Mobile Telecommunications-Advanced (IMT-Advanced) radio interface [1]. IMT-Advanced systems are the 4th-Generation (4G) mobile systems that offer significant enhancements over IMT-2000 3G systems. IMT-Advanced systems are anticipated to provide significant improvement in performance and Quality of Service (QoS). Since the seamless operation when roaming within the network is a key requirement of IMT-Advanced, a very tight interruption time of 27 ms during (intra-radio access technology (RAT), excluding radio link synchronization) handover is required to be met in order to provide uncompromised QoS for current and future mobile Internet applications. Emerging broadband wireless air interface specification such as IEEE 802.16m [2, 3], which provides enhanced link layer amendment to the legacy IEEE 802.16 system [4, 5], is designed to meet and in many cases exceed IMT-Advanced requirements. In particular, handover latency is one of the key improvement areas during the development of the IEEE 802.16m.

In this paper, a detailed handover latency analysis of IEEE 802.16m handover scheme is provided in order to understand how IMT-Advanced handover latency requirements can be met. The remaining part of the paper is organized as follows. In Sect. 2, system architecture for handover support is described. A detailed handover latency analysis based on the IEEE 802.16m air interface specification is provided in Sect. 3. Finally, Sect. 4 concludes the paper.

2 System Architecture for Handover Support

In any cellular system, handover operation is closely related to the network architecture, where many handover optimizations are done jointly by the network as well as air interface. A simplified logical network architecture of IEEE 802.16e/m based WiMAX system [6] is shown in Fig. 1 where the core network is fully IP based. The IEEE 802.16m Access Service Network (ASN) is a single entity that manages both the data plane and control plane functions for each Mobile Station (MS) during mobility. For each MS, a designated ASN Gateway (ASN-GW) inside the ASN-GW pool is its anchor ASN, so that in the most common mobility case, the anchor ASN-GW remains the same even if the MS is subsequently served by different Base Stations (BSs) due to mobility, which is called as ASN-anchored handover. For example, in Fig. 1 MS1 is served by BS2 and its anchor ASN-GW is ASN-GW1. If it performs ASN-anchored handover to BS3, its serving ASN-GW remains at ASN-GW1 and ASN-GW2 will only be the relay point between ASN-GW1 and BS3 in the network to facilitate this anchoring function. Note that the network may also relocate MS's anchor ASN-GW to another, due to ASN-GW load balancing, data path optimization or an ASN boundary-crossing mobility, which will introduce extra latency if this operation happens during the mobility. Such non-collocated ASN mobility becomes more likely when network becomes heterogeneous, e.g., a mixed deployment of macro cells with home femto cells.

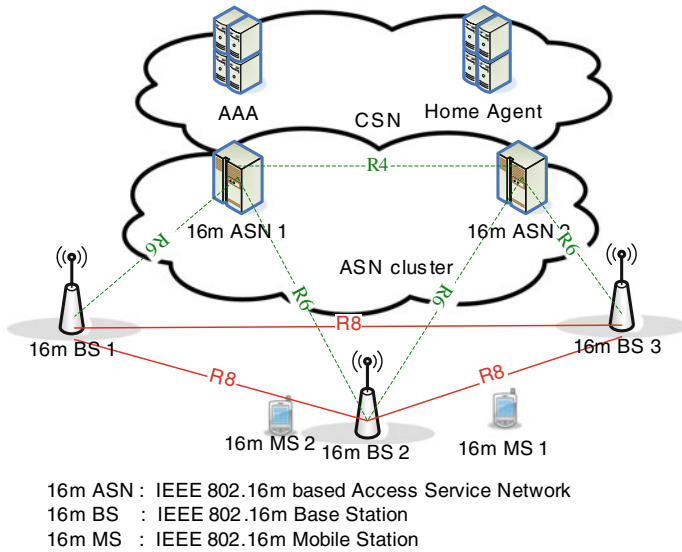


Fig. 1 The network architecture of IEEE 802.16m based systems

3 Handover Latency Analysis

In this section, we provide quantitative analysis on the measurable handover performance metric: handover latency. The minimum latency requirement for IMT-Advanced system such as IEEE 802.16m is 27.5 ms [1], which assumes a normal handover instance without any erroneous matters. Because the latency requirement of IMT-Advanced does not consider the effect of wireless environment and user mobility, handover latency caused by handover control packet errors due to wireless channel impairment or abnormal user mobility is not considered in the handover latency analysis. The 27.5 ms requirement latency includes steps for performing uplink synchronization (CDMA code ranging) and completing network re-entry at the target BS. Hence, the transmission of control messages for network re-entry (ranging MAC control message) is also considered since data communication may not be performed before this when seamless handover option is not used. In addition, service interruption may also occur before the actual handover execution. Table 1 shows the latency of each step during the handover execution, including physical-layer operation for CDMA ranging and transmission of ranging MAC control messages at the target BS. The numbers are derived based on analytical models, which account for the frame structure design of IEEE 802.16m [3]. In particular, a transmission happens within a subframe, with length of 0.617 ms. Eight subframes make up one frame, in which the device may receive or transmit a signal. The frame Downlink-to-Uplink ratio is configurable, e.g. 5 DL subframes and 3 UL subframes in one frame. Overall fully

Table 1 Overall handover latency analysis based on HO procedures

HO latency	Synchronization	10.617 ms (=5 + 0.617 + 5) RF switch + preamble acquisition + Fast_RNG_IE reception
	UL allocation Request and Grant for MS identification	7.234 ms (=0.617 + 2+0.617 + 2 + 2) RNG-REQ signaling + Resource Reservation + RNG-RSP reception + RNG-RSP decoding + ID update
	HO Confirmation	6.234 ms (=0.617 + 0.617 + 5) CQICH allocation reception + CQI code signaling
	Total/requirement (ms)	24.085/27.5

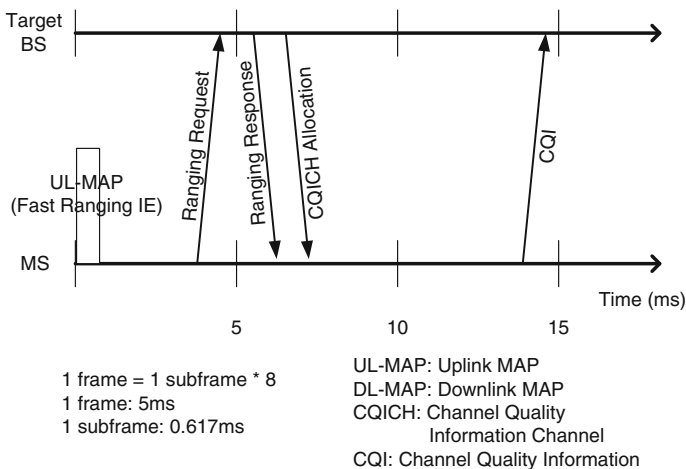


Fig. 2 Overall handover procedure (Fully optimized)

optimized handover procedure with timing is shown in Fig. 2. With the proper handover preparation as discussed in Sect. 3, the target BS transmits a UL resource allocation Information Element (IE) at the arranged Action Time, which takes 0.617 ms. Assuming the device has succeeded in synchronizing to the UL channel of the target BS, the device subsequently performs MAC layer network re-entry procedure by transmitting the ranging MAC control message to the target BS, which requires one UL subframe (i.e. 0.617 ms). Upon reception of this message, the BS will authenticate the device and if successful, it will reserve resources for proceeding the network re-entry for the device which will incur an operation latency of approximately 2 ms. The information on reserved resources will be signaled to the device with the ranging response MAC control message consuming one DL subframe. The device needs to decode and update its information based on

the conveyed information within the ranging response MAC control message which takes approximately 4 ms. Hence, the total latency incurred for the MAC layer network re-entry sums up to 7.851 ms. After successful network re-entry, to continue data communication, the device needs to report its Channel Quality Information (CQI) for efficient DL data, which requires first to be allocated a Channel Quality Information Channel (CQICH) and its following CQI report. This operation takes 6.234 ms in total. In summary, the whole handover procedure as described incurs a total of 14.085 ms, which satisfies latency requirement of 27.5 ms by a good margin.

4 Conclusion

In this paper, a detailed handover performance analysis of IEEE 802.16m based next-generation WiMAX have been provided. Through the performance analysis, we can understand how IMT-Advanced handover latency requirements can be met. The whole handover procedure incurs a total of 14.085 ms, which satisfies latency requirement of 27.5 ms by a good margin.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A1014610).

References

1. IEEE 802.16m-09/0034r2, IEEE 802.16m System Description Document, Sept 2009
2. IEEE Standard for Local and metropolitan area networks Part 16: Air Interface for Broadband Wireless Access Systems Amendment 3: Advanced Air Interface, IEEE Std 802.16m-2011, May 12 2011
3. IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Broadband Wireless Access Systems, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1, IEEE Std 802.16e-2005, Feb 28 2006
4. IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Broadband Wireless Access Systems, IEEE Std 802.16-2009 (Revision of IEEE Std 802.16-2004), May 29 2009
5. ITU-R M.2134, Requirements Related to Technical System Performance for IMT-Advanced Radio Interface(s) [IMT.TECH], draft new report, Nov 2008
6. WiMAX End-To-End Network System Architecture, Stage 3: detailed protocols and procedures, WiMAX Forum, Aug 2006

An Emergency Message Broadcasting Using Traffic Hazard Prediction for Vehicle Safety Communications on a Highway

Sang Yeob Oh

Abstract On a highway, it has high potential of ‘chain-reaction collision’ to occur, since vehicle has high velocity. As a preventive method for chain-reaction collision, native broadcasting and intelligent broadcasting that sends emergency messages to rear vehicles were proposed. However, these methods were ineffective when vehicles were concentrated in one area. This paper offers you a broadcasting method that considers stopping distance and improved back-off algorithm based on the number of neighbor nodes, which solves previous problems. As a result, the frame reception success rate comparing two different methods, broadcasting methods considering dangerous factors, was 7 % improved than native and intelligent broadcasting.

Keywords VSC · Back-off · Broadcasting · Emergency message

1 Introduction

Vehicular Network is one of the critical technologies in ITS (Intelligent Transport System). In IEEE, they standardized IEEE 802.11p WAVE to support Vehicular Network technology. Usually, Vehicular Network is divided into two specific categories; IVC (Inter-Vehicular Communication) and RVC (Roadside-Vehicular Communication). However, when Vehicular Network seems to have both characteristics of IVC and RVC, we call them HVC (Hybrid-Vehicular Communication) [1].

S. Y. Oh (✉)

School of Interactive Media, Gachon University, 1342 Seong-Nam-Dae-Ro,
Su-Jeong-Gu, Seong-Nam-Si, Kyung-Gi-Do 461-702, South Korea
e-mail: syoh1234@gmail.com

On a highway, it has high potential of ‘chain-reaction collision’ to occur, since vehicle has high velocity. In order to prevent these incidents, studies about VSC (Vehicle Safety Communication) has been done. One of the method from the studies was called NB (Native Broadcasting) [2]. However, possible occurrence of “broadcast storm is relatively high when cars are concentrated. To solve this problem, such as IB (Intelligent Broadcasting) which is random based selective rebroadcasting system, came out [3]. In this situation, IB acts differently with NB by backing out rebroadcasting, if one node identifies another node is already broadcasted. However IB has limitation and drawbacks when vehicle’s concentration is high, it is more likely to have frame collision in randomly given delayed time.

This paper suggests technique applied number of neighbor node with back-off algorithm in order to shorten the delaying time. Unlike basic system of sending emergency message after collision happens, this paper also offers alternative that sends emergency message when it senses abrupt deceleration of vehicle.

2 Proposed Broadcasting Protocol

2.1 Neighbor Node Table

Every node broadcast its ID and GPS information while sending hello message periodically. After sending hello message, it saves information (Node ID and Distance field) in neighbor node table. Distance field is calculated with reception node’ GPS location and location of GPS that received. Hello message is key element to calculate the distance between other vehicles so it changes transmission period based on vehicle’s velocity. The equation for solving the calculation is as below.

$$I_{hello\ message} = I_{min} + \left(1 - \frac{v_n}{v_{max}}\right) \times I_{max} \quad (1)$$

In this equation, I_{min} and I_{max} each represents interval maximum value and minimum value of hello message. v_{max} represents maximum value of velocity and v_n represents current velocity of node in Eq. (4). Following the equation, faster the velocity is, shorter the interval of hello message. If a node receives emergency message, transmission period is suspended for 5 s.

2.2 Improved Back-Off Algorithm

As far as emergency message’s source node and receiving node is the signal range of it increases. On that account, this paper offers system that makes node to compete in each distanced section in order to give nodes that are more distant a highest priority to rebroadcast. Firstly, we can solve out the SI from the equation below.

$$SI = C_s - \left\lfloor \frac{D_{s-d}}{D_{mp}} \times C_s \right\rfloor \tag{2}$$

C_s represents distanced section phase number and D_{s-d} represents simultaneous transmission and reception (or two-way messenger) nodes' distance. D_{mp} represents maximum propagation distance. In this equation SI represents section index of signal-receiving node. By using this equation below, we can solve out the back-off section.

$$[(SI - 1) \times n, (SI \times n - 1)] \tag{3}$$

In this equation, n represents neighbor node count by using neighbor node table.

2.3 Traffic Hazard Prediction Algorithm

On a highway, when driver senses risk element, he tries to a quickly brake its car. This paper presents a system that's aware of possibility of collision when certain threshold value is below than average while monitoring node's acceleration. In order to apply this system, calculate current velocity (V_n) for every I_v second.

$$V_n = (1 - \alpha) \times V_{n-1} + \alpha \times V_{current}, (n = 1, 2, \dots), \text{ if } n = 1 \text{ then } V_{n-1} = 0 \tag{4}$$

In this equation, V_{n-1} represents velocity calculated previously and a controls effect of current velocity over accumulative velocity. After solving equation calculate acceleration (V_a) for every I_v second by using equation below.

$$V_a = \frac{V_{current} - V_{n-1}}{I_{dv}} \tag{5}$$

For every I_v period, Eqs. (4) and (5) are calculated. Apply accumulative velocity V_{n-1} in Eq. (5) not V_n , in order to calculate the acceleration (Table 1).

3 Simulation

3.1 Environment and Scenario

Figure 1 shows scenario of vehicle movement of 4 lanes and measure nodes' frame reception success rate at maximum 1000 m distant distance. Distance between vehicles as safe headway. We used [4]'s stopping distance calculator to calculate stopping distance. For example, in dry asphalt, the stopping distance of vehicle with 100 and 120 km/h of velocity, stopping distances of each vehicle is 56.23 and 80.98 m. Then if node senses black spots inside range of 2500 m, the

Table 1 Simulation parameters

Road size	2500 m × (3.5 × Lane count) m
Lane count	4
Safe headway	2 s
Velocity	40–160 km/h
Transmission range	250 m
MAC	IEEE 802.11p
TTL	5
Frame size	200 bytes
C_s	4
I_{min}	0.1 s
I_{max}	1.5 s
V_{max}	200 km/h
I_v	0.5 s
Threshold value	-20 m/s
α	0.2

distance where vehicle actually decelerates its speed is 2443.77 and 2419.02 m each. In this experiment we assume that when braking activates, it will uniformly decelerate until the velocity is 0 km/h.

3.2 Simulation Result

Figure 2 shows a successfully received rate of sending emergency message to vehicles, when node senses black spots inside range of 1 km, based on changes of velocity in 0.5–0.7 s. As velocity increases, as safe distance is receded, rate of nodes that competes other nodes to broadcast decreases, thus, possibility of frame collision also decreases. Therefore as vehicle’s velocity increases reception rate also increases.

According to result of this experiment, method that this paper offered averagely increased the frame reception success rate by 7 %. This is because of sending emergency message faster by using traffic hazard prediction algorithm comparing with other methods.

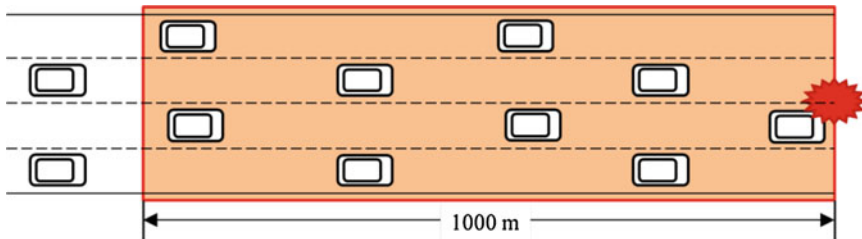


Fig. 1 Vehicle movement scenario (4-lanes)

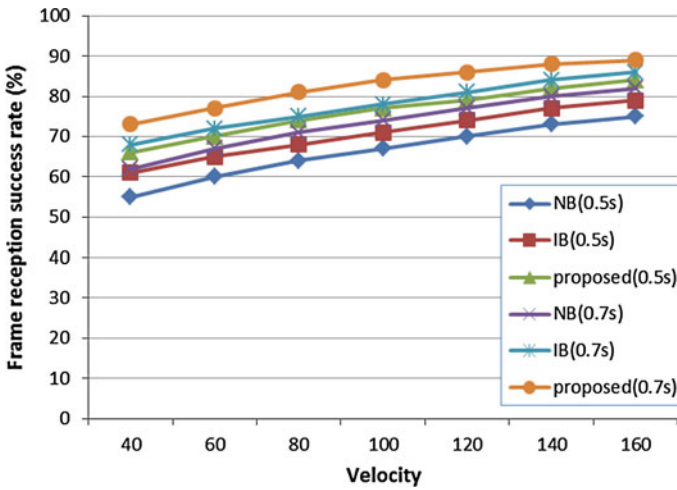


Fig. 2 Frame reception success rate per velocity in given time

4 Conclusions

This paper offered broadcasting method for VSC in highway. Our method, broadcasts emergency message coadaptationally, using number of neighbor nodes based on how vehicles are concentrated in one area. Also, rather than sending emergency message after the collision, predicting traffic hazard made faster communication possible. And experimenting with NB and IB, frame reception success rate has been improved about 7 %.

Acknowledgments This work was supported by the Gachon University research fund of 2012 (GCU-2012-R158).

References

1. Sichertiu M, Kihl M (2008) Inter-vehicle communication systems: a survey. *IEEE Commun Surv Tutor* 10(2):88–105
2. Biswas S et al (2006) Vehicle-to-vehicle wireless communication protocols for enhancing highway traffic safety. *IEEE Commun Mag* 44(1):74–82
3. Fukuhara T, Warabino T (2005) Broadcast methods for inter-vehicle communications system. In: *Proceedings of IEEE wireless communications and networking conference, New Orleans*, pp 2252–2257
4. <http://forensicsdynamics.com/stopping-distance-calculator>, Stopping (braking) distance calculator

Detecting SIM Box Fraud Using Neural Network

Abdikarim Hussein Elmi, Subariah Ibrahim
and Roselina Sallehuddin

Abstract One of the most severe threats to revenue and quality of service in telecom providers is fraud. The advent of new technologies has provided fraudsters new techniques to commit fraud. SIM box fraud is one of such fraud that has emerged with the use of VOIP technologies. In this work, a total of nine features found to be useful in identifying SIM box fraud subscriber are derived from the attributes of the Customer Database Record (CDR). Artificial Neural Networks (ANN) has shown promising solutions in classification problems due to their generalization capabilities. Therefore, supervised learning method was applied using Multi layer perceptron (MLP) as a classifier. Dataset obtained from real mobile communication company was used for the experiments. ANN had shown classification accuracy of 98.71 %.

Keywords Multi layer perceptron · SIM box fraud · Classification · Telecom fraud

A. H. Elmi (✉) · S. Ibrahim · R. Sallehuddin
Faculty of Computer Science and Information System, Universiti
Teknologi Malaysia, 81300 Skudai, Johor, Malaysia
e-mail: sirabdikarim@yahoo.com

S. Ibrahim
e-mail: subariah@utm.my

R. Sallehuddin
e-mail: roseline@utm.my

1 Introduction

The theft of service and misuse of voice as well as data networks of telecom providers is considered as fraud. The perpetrator's intention could be to avoid the service charges completely or reduce the charges that would have been charged for the service used. The intention could also be deeper than that and the fraudster's aim might be to gain profit by misusing the network of the provider [1]. Losses due to fraud in telecom industry are highly significant.

Even though telecommunication industry suffers major losses due to fraud there is no comprehensive published research on this area mainly due to lack of publicly available data to perform experiments on. The data to be used for the experiments contains confidential information of customers and in most cases law and enforcement authorities prohibit exposing the confidential information of customers [2]. On the other hand, any broad research published publicly about fraud detection methods will be utilized by fraudsters to evade from detection [1, 3]. Existing research work is mainly focusing on subscription and superimposed types of fraud which are the dominant types of fraud in telecom industries worldwide. However, another type of fraud called SIM box bypass fraud has become a challenging threat to telecom companies in some parts of Africa and Asia. The success of this fraud depends on obtaining SIM cards. Therefore the effects of SIM box bypass fraud vary across countries. In countries where unregistered SIM cards are not allowed and the government laws recognize the SIM box devices as illegal equipment, the effect is less compared to countries where obtaining of SIM cards by customers is very cheap or even free and government laws do not prohibit unregistered subscribers. The fact that this type of fraud is not a problem for all telecom companies worldwide might justify the reason why the publicly available research on this type of fraud is very limited.

SIM box' fraud takes place when individuals or organizations buy thousands of SIM cards offering free or low cost calls to mobile numbers. The SIM cards are used to channel national or international calls away from mobile network operators and deliver them as local calls, costing operators' millions in revenue loss [4, 5]. A SIM box is VoIP gateway device that maps the call from VoIP to a SIM card (in the SIM box) of the same mobile operator of the destination mobile [5].

In this paper we present a study on which set of descriptors that can be used to detect SIM cards originating from SIM box devices have been identified. Neural Networks are promising solutions to this type of problem as they can learn complex patterns and trends within a noisy data. Neural networks have particularly shown better performance results than other techniques in the domain of telecom fraud. Therefore, supervised learning method was applied using Multi layer perceptron (MLP) as a classifier. The dataset that was used for this study is obtained from a real mobile communication network and contains subscribers/SIM cards that have been tested and approved by the operator to be SIM box fraud SIM cards and normal SIM cards. The next section describes the dataset and descriptors used.

Section 2 discusses the method applied and Sect. 3 presents the results obtained from the experiments. Last section concludes the work.

1.1 Dataset and Descriptors

The huge volumes of data stored by telecommunication companies include Customer Data Record (CDR) which is the database that stores the call information of each subscriber. Whenever a subscriber makes a call over the operator’s network a toll ticket is prepared which contains complete information of the call made including the subscriber id, the called number, duration of the call, time, destination location etc. In this fraud scenario the CDR database serves a suitable source of information where useful knowledge about callers can be extracted to identify fraudulent calls made by subscribers.

This study is based on Global Systems for mobile communications (GSM) network and specifically the Customer Data Record (CDR) database of prepaid subscribers. The dataset used for the experiments contained 234,324 calls made by 6415 subscribers from one Cell-ID. The dataset consisted of 2126 fraud subscribers and 4289 normal subscribers which is equivalent to 66.86 % of legitimate subscribers and 33.14 % of SIM box fraud subscribers. The total duration of these call transactions was two months.

A Total of 9 features have been identified to be useful in detecting SIM box fraud. Table 1 shows the list of these features and their description.

Table 1 Selected descriptors

Field Name	Description
Call sub	This is the subscriber identity module (SIM) number which was used as the identity field
Total calls	This feature is derived from counting the total calls made by each subscriber on a single day
Total numbers called	This feature is the total different unique subscribers called by the customer (subscriber) on a single day
Total minutes	Total duration of all calls made by the subscriber in minutes on a single day
Total night calls	The total calls made by the subscriber during the midnight (12:00 to 5:00 am) on a single day
Total numbers called at night	The total different unique subscribers called during the midnight (12:00 to 5:00 am) on a single day
Total minutes at night	The total duration of all calls made by the subscriber in minutes at midnight (12:00 to 5:00 am)
Total incoming	Total number of calls received by the subscriber on a single day
Called numbers to total calls ratio	This is the ratio of the total numbers called/total calls
Average minutes	The is the average call duration of each subscriber

2 Materials and Methods

Neural Network is a group of simulated neurons interconnected to represent a computation mathematical model that can take one or more inputs to produce an output by learning the complex relationships between the inputs and outputs [6]. Supervised learning requires an input pattern along with the associated output values which is given by an external supervisor [7, 8].

2.1 Multi Layer Perceptron

Feed Forward Neural Network contains neurons and edges that form a network. The neurons are set of nodes and are of three types: input, hidden and output. Each node is a unit of processing. The edges are the links between two nodes and they have associated weights [8]. In Multi layer perceptron the network consists of multiple layers of computational units, usually connected in a feed-forward way. Each neuron in one layer has direct connections to the neurons of the subsequent layer although not to other nodes in the same layer. There might be more than one hidden layer [9, 10].

A neuron has a number of inputs and one output. It combines all the input values (Combination), does certain calculations, and then triggers an output value (activation) [8, 11]. There are different ways to combine inputs. One of the most popular methods is the weighted sum, meaning that the sum of each input value is multiplied by its associated weight. Therefore, for a given node g we have:

$$Net_g = \sum w_{ij}x_{ij_1} = w_{0j}x_{0j} + w_{1j}x_{1j} + \dots + w_{ij}x_{ij} \quad (1)$$

where x_{ij} represents the i 'th input to node j , w_{ij} represents the weight associated with the i 'th input to node j and there are $I + 1$ inputs to node j .

The value obtained from the combination function is passed to non-linear activation function as input. One of the most common activation functions used by Neural Network is the sigmoid function. This is a nonlinear functions and result in nonlinear behaviour. Sigmoid function is used in this study. Following is definitions of sigmoid function:

$$\text{Sigmoid} = \frac{1}{1 + e^{-x}} \quad (2)$$

where x is the input value and e is base of natural logarithms, equal to about 2.718281828. The output value from this activation function is then passed along the connection to the connected nodes in the next layer.

Back-propagation algorithm is a commonly used supervised algorithm to train feed-forward networks. The whole purpose of neural network training is to minimize the training errors.

Equation 3 gives one of the common methods for calculating the error for neurons at the output layer using the derivative of the logistic function:

$$Err = O_1(1 - O_1)(T_i - O_1) \tag{3}$$

In this case, O_i is the output of the output neuron unit i , and T_i is the actual value for this output neuron based on the training sample. The error calculation of the hidden neurons is based on the errors of the neurons in the subsequent layers and the associated weights as shown in Eq. 4.

$$Err_i = O_i(1 - O_i) \sum_j Err_j W_{ij} \tag{4}$$

O_i is the output of the hidden neuron unit I , which has j outputs to the subsequent layer. Err_j is the error of neuron unit j , and W_{ij} is the weight between these two neurons. After the error of each neuron is calculated, the next step is to adjust the weights in the network accordingly using Eq. 5.

$$W_{ij,new} = W_{ij} + l * Err_j * O_i \tag{5}$$

Here l , is value ranging from 0 to 1. The variable l is called learning rate. If the value of l is smaller, the changes on the weights get smaller after each iteration, signifying slower learning rates.

To obtain the best Neural Network architecture for this research, four parameters settings were considered. The number of hidden layers in the network architecture as well as the number of neurons in each hidden layer is considered. The learning rate and momentum parameters which have significant effect on the performance of any neural network architecture are also considered.

Three architectures of neural network were considered in this research; one, two and three hidden layers and 5, 9 and 18 hidden nodes in each hidden layer. The learning rate is a constant chosen to help the network weights move toward a global minimum of Sum Square Error (SSE). Therefore, in this research four values of learning rate are considered: 0.1, 0.3, 0.6 and 0.9. The back-propagation algorithm is made more powerful through the addition of a momentum term. Momentum helps in the early stages of the algorithms, by increasing the rate at which the weights approach the neighbourhood of optimality. Therefore, four values of momentum term are used in this study: 0.1, 0.3, 0.6 and 0.9.

3 Results and Discussions

This section discusses the results obtained in comparing the ANN models created to find the neural network architecture which provides the most reliable and accurate predictions. All possible combination of the parameter settings was experimented and as a result, 240 neural network models were created. The models were evaluated based on their prediction accuracy, generalization error, time taken

to build the model, precision and recall. 10—Fold cross-validation results of the models were compared.

Classification accuracy ranged from 56.1 to 98.71 %. It has been observed that the models show the worst performance results when both momentum and learning rate are increased to range of 0.6–0.9. The highest classification accuracy that could be achieved in these values was 86.16 % and the root mean square error was as high as 0.66. In all network layers; 3, 4, and 5, the overall accuracy degraded significantly when this range was used. This could be explained by the fact that higher values of learning rate and momentum could lead the algorithm to overshoot the optimal configuration.

Figure 1 compares the best classification accuracy of the three hidden layers experimented with respect to the learning rate parameter. When one and two hidden layers are used, the accuracy degrades as the learning rate is increased from 0.1 to 0.3. But accuracy again increases until the learning rate is 0.6 where it starts to decline dramatically if further increased. However, when three hidden layers are used, the accuracy increases as the learning rate is increased from 0.1 to 0.6 and then the accuracy declines if the learning rate is increased from this point.

The highest accuracy was achieved when two hidden layers were used at a learning rate of 0.6. Another observation shown by the graph is that the classification accuracy for all hidden layers decreases as the learning rate is increased from 0.6 to 0.9.

Figure 1 also compares the classification accuracy of all hidden layers with respect to momentum term. From the figure, it can be clearly seen that hidden layer 2 at a momentum of 0.3 shows the best performance. The performance also degrades after the momentum of 0.3 for two and three hidden layers. The classification accuracy degrades significantly for all hidden layers at a momentum of 0.9.

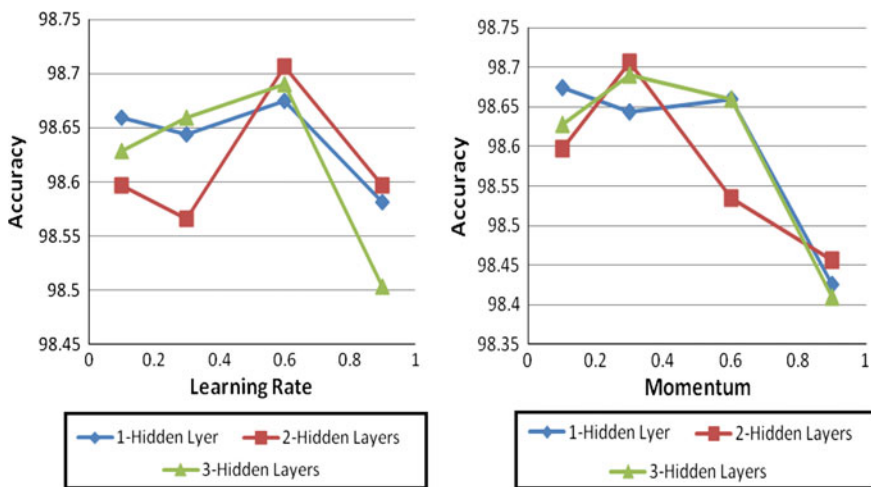


Fig. 1 Accuracy against learning rate and momentum for all hidden layers

Table 2 Confusion matrix of Selected ANN model

	Normal	Fraud
Normal	4269	20
Fraud	63	2063

Table 3 Results of the best ANN

	Best model
RMSE	0.10380
Accuracy	98.7061 %
Time	17.17
ROC area	0.997
Precision	0.987
Recall	0.987

Table 4 Parameter values of best ANN model

	Best model
Input layer nodes	9
Hidden layer 1 nodes	5
Hidden layer 2 nodes	5
Hidden layer 3 nodes	N
Output nodes	2
Learning rate	0.6
Momentum	0.3

The conclusion that can be made from these figures is that, very high learning and momentum rates significantly degrade the classification accuracy of the models. The best results are obtained when lower value of momentum is used with relatively higher value of learning rate.

From the analysis discussed in this section, the best results were obtained when two hidden layers each having five hidden neurons was used with learning rate of 0.6 and a momentum term of 0.3.

In the confusion matrix shown in Table 2, the columns represent the predicted values and rows represent the actual cases. The model was able to correctly classify 2063 out of the 2125 fraud subscribers and 4269 out of the 4289 normal subscribers. Fraud is the negative target value, false negative count is 63 and false positive count is only 20. Table 3 shows performance results of the best model and Table 4 shows the parameter values used in this model.

4 Conclusions

The focus of this work was to come up with a set of features that can be used to effectively identify SIM cards originating from SIM box devices and an algorithm that can classify subscribers with high accuracy. The learning potentials of neural

network for the detection of SIM box fraud subscribers were investigated. The experimental results revealed that ANN has high classification accuracy. SVM has recently found considerable attention in classification problems due to its generalization capabilities and less computational power. In future work SVM will also be investigated and compared with ANN.

Acknowledgments The authors first thank the anonymous reviewers for their valuable comments and to Universiti Teknologi Malaysia (UTM) for the FRGS Grant Vote number 4F086 that is sponsored by Ministry of Higher Education (MOHE) and Research Management Centre, Universiti Teknologi Malaysia, Skudai, Johor.

References

1. Taniguchi M, Haft M, Hollmen J, Tresp V (1998) Fraud detection in communications networks using neural and probabilistic methods. In: Proceedings of the 1998 IEEE international conference on acoustics speech and signal processing, vol 2. IEEE, Los Alamitos, pp 1241–1244
2. Hilas C, Mastorocostas P (2008) An application of supervised and unsupervised learning approaches to telecommunications fraud detection. *Knowl Based Syst* 21(7):721–726
3. Azgomi NL (2009) A taxonomy of frauds and fraud detection techniques. In: Proceedings of CISTM 2009, Ghaziabad, India, pp 256–267
4. Telenor GS (2010) Global SIM box detection
5. Nokia Siemens Networks Corporation (2008). Battling illegal call operations with fraud management systems
6. Larose DT (2005) *Discovering knowledge in data*. John Wiley and Sons, Inc., Hoboken
7. Ghosh M (2010) Telecoms fraud. *Comput Fraud Secur* 2010(7):14–17
8. MacLennan J (2009) *Data mining with Microsoft SQL Server 2008*. Wiley Publishing Inc, Indianapolis
9. Mark EM, Venkayala S (2007) *Java data mining strategy, standard, and practice*. Diane Cerra, San Francisco
10. Cortesao L, Martins F, Rosa A, Carvalho P (2005) Fraud management systems in telecommunications: a practical approach. In: *Proceeding of ICT, 2005*
11. Pablo A, Este'vez CM, Claudio AP (2005) Subscription fraud prevention in telecommunications using fuzzy rules and neural networks. In: *Proceedings of the expert systems with applications*. Santiago, Chile, 2005
12. Hilas C, Mastorocostas P (2008) An application of supervised and unsupervised learning approaches to telecommunications fraud detection. *Knowl Based Syst* 21(7):721–726

Overlapping-Node Removal Algorithm Considering the Sensing Coverage of the Node in WSN

Doo-Wan Lee, Chang-Joon Kim and Kyung-Sik Jang

Abstract In this paper, we propose Overlapping-node Removal Algorithm that utilizes the backup node to the detection of the nodes in the WSN sensing coverage overlapping node. Overlapping sensing coverage of the node that has a data network, energy efficiency is very low. In the proposed algorithm, the initial network connection each node to know the location information of neighbor nodes within their communication coverage, because according to the radius of the sensing coverage overlapping node may be able to identify. Overlapping node that occurs during the network tasks is changed to the standby-node until a specific event occurs, replace the fault node can improve the network lifetime.

Keywords Overlapping node · Sensing coverage · WSN · Self-organization

1 Introduction

In Wireless Sensor Network (WSN) under the limited energy source, one of the main design goals is self-organization. WSNs are a large number of small devices capable of executing sensing, data processing and communication tasks. As sensor nodes may be placed everywhere, this type of network can be applied to healthcare,

D.-W. Lee · C.-J. Kim · K.-S. Jang (✉)

Korea University of Technology and Education Byeongcheon-myeon, Cheonan-si,
Chungnam, South Korea
e-mail: ksjang@kut.ac.kr

D.-W. Lee
e-mail: neomenie@kut.ac.kr

C.-J. Kim
e-mail: chjkim@kut.ac.kr

environment monitoring system, a military, disaster surveillance, and so on. WSNs developments and proposals in existence have been designed to build a network for just one type of node, where all nodes can communicate with any other nodes in their communication coverage area.

WSN arranged to Area of Interest (AOI) perform the work that communication coverage and performing the network event generated in sensing coverage and delivers data to Base Station. Sensing coverage is senses the event generated by the area smaller than communication coverage in the sensor node around. Detected event applies the hierarchy structure routing algorithm of delivering with the parents-node. Therefore, the network energy efficiency falls down by the same event sensing from the area in which the high density of the node. In this paper, Overlapping Node Removal Algorithm (ODRA) which converts the node to the atmosphere node and prevents unnecessary redundancy data and improves the network lifetime of the total network that the sensing area is overlapped through the network assessment in order to remove this redundancy data are proposed.

2 Related Works

2.1 Backup Node Fault Tolerant Routing Algorithm

In order to extend WSN lifetime, it should manage energy efficiently, it has to save all data and communicate them to users in temporal failure. Therefore, it needs routing algorithm so as to take action in that kind of case. This research suggests Backup-Node Fault-Tolerant Routing Algorithm (BN-FTRA) that is able to maintain the condition of network communication by backup-node when the node is died or occurred sudden fail of network communication in WSN. BN-FTRA saves their information in backup-node that based on neighbor-node list table at the beginning of network configuration. Therefore BN-FTRA is expert to take an action to failure-node (Fig. 1).

BN-FTRA Neighbor List Table saves information of neighbor-node, it is able to check self-organization and present condition of the node by verifying its information and neighbor-node information. The Energy information of the neighbor-node is stored in NLT. Therefore if the temporally-failed node is occurred, the node in which the energy is higher is replaced with the backup-node in NLT.

2.2 Received Signal Strength Indication

The Received Signal Strength Indicator (RSSI) method is based on the fact that the radio signal strength decreases with the distance. In this context, the path loss is the attenuation that a signal undergoes in travelling over a path between two points.

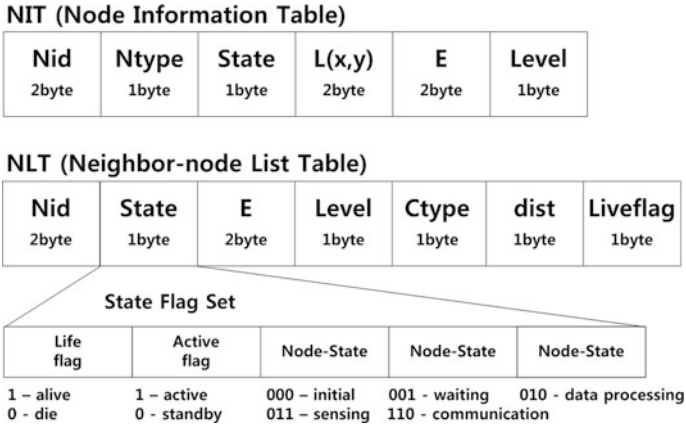


Fig. 1 NIT and NLT packet format

Using the signal strength to determine the distances usually yields to a number of errors because the actual path loss depends on many factors related to the environment, such as reflections, diffraction, scattering and antenna orientation.

3 Overlapping Node Removal Algorithm

3.1 Level Decision Command

At the beginning of network configuration, every node including Base Station is placed in AOI and afterward, it starts network configuration. Base Station broadcast level decision command (LDC) it updates NLT that receive LDC. LDC minimize loss of communication by broadcasting each node 3 times.

Base Station sets its own level to 0 and commands to determine level to all nodes positioned in communication coverage. All nodes compare level to NIT and if you do not receive a message than the level to increase their own +1 level setting. The transmitted node is set to the parent node and broadcasting does LDC to the communication coverage. The information of the transmitted node is stored in NLT if the level setting is completed, and ACK message send to the parent node. If there is no response message for a certain period of times, it determines that oneself is the lowest level in the network. The node LDC_Done message is set to the parent node. The Base Station terminates LDC command when it receives LDC_Done message.

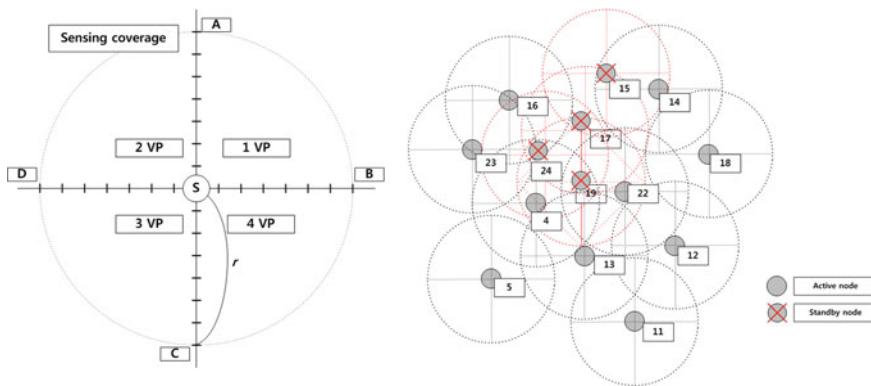


Fig. 2 Overlapping node removal algorithm

3.2 Overlapping Node Removal Algorithm

LDC command setting is completed, the Base Station broadcasting ONRA command. The node in which NodeID is the highest among is referred. The oneself sets to the center of Virtual Place (VP). If more than 1 node is located in 2 VPs in each VP after dividing into the VP1($0 < x < r, 0 < y < r$), VP2($-r < x < 0, 0 < y < r$), VP3($-r < x < 0, -r < y < 0$), and VP4($0 < x < r, -r < y < 0$), oneself is set to the overlapping node ($r = \text{sensing coverage}$). Overlapping node converts to the standby-node, active-node joins in network connection. If the fail node occurs in network, it is replaced by standby-node which is selected on basic of the priority determined by its power capacity in NLT. Base Station is able to include standby-node, and set parent node and neighbor-node if there is no standby-node (Fig. 2).

4 Simulation and Performance Evaluation

4.1 Simulation Environments

Table 1 shows simulator conditions for analyze the performance of ONRA.

In our simulation, it is assumed that all the sensor nodes have the same communication coverage of 20 m and sensing coverage of 5 m. Also, 1 thousand sensor nodes are randomly distributed in a 50×50 m square AOI, and each node has a unique NodeID, NIT, and NLT. Once a sensor node is deployed, it will not change its location. The network life-time is the case that 15 % or less of the number of the total node survived. The fault node of 0.5 % generated for the Simulation-Time unit. The energy efficiency of the applies ONRA network and which it doesn't apply network was measured and the network life-time was confirmed.

Table 1 Simulation conditions

Descriptions	Value
Area of interest (AOI)	50 × 50 m
Base station location	Random deployment
Sensor node number	1000 ea
Communication coverage	20 m
Sensing coverage	5 m
Initial energy	100,000 mJ
Fault node rate	0.5 %/Sim.Time
Network life-time threshold	Total node #/15 %

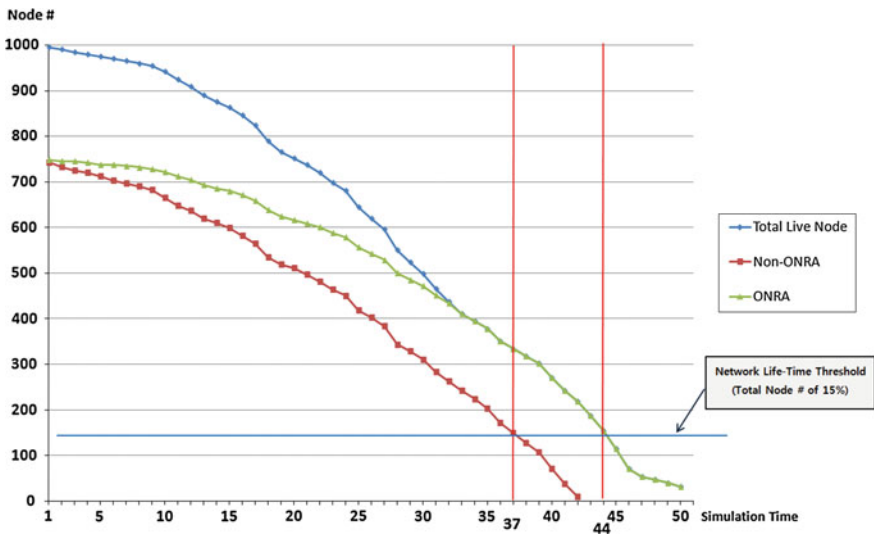


Fig. 3 Effect of ONRA

4.2 Effects of ONRA

It could confirm that network life-time of the experimental result ONRA was improved with 14 %. When connecting of the network, the node which was selected as the overlapping node and is converted to the standby-node operated with the backup-node in the fault node of network fail and it could confirm that network life-time was improved. In addition, the duplex of data which consequently is concentrated on Base Station and parent-node can be minimized because the selection process of standby-node is decided overlapped sensing coverage. The unnecessary energy efficiency which is used up in the overlapping data transmission can be enhanced.

5 Conclusions

In WSN under the limited energy source, one of the main design goals is self-organization. WSNs are a large number of small devices capable of executing sensing, data processing and communication tasks (Fig. 3).

In this paper, ODRA which converts the node to the atmosphere node and prevents unnecessary redundancy data and improves the network lifetime of the total network that the sensing area is overlapped through the network assessment in order to remove this redundancy data are proposed. It could confirm that network life-time of the experimental result ONRA was improved with 14 %. Standby-node was selected as the sensing coverage overlapping node. In addition, the duplex of data which consequently is concentrated on Base Station and parent-node can be minimized because the selection process of standby-node is decided overlapped sensing coverage. The unnecessary energy efficiency which is used up in the overlapping data transmission can be enhanced.

References

1. Lloret J, Farcia M, Bri D, Diaz JR (2009) A cluster-based architecture to structure the topology of parallel wireless sensor networks. *Open Access Sens* 9:10513–10544
2. Chen Y-S, Lo T-T, Ma W-C (2010) Efficient localization scheme based on coverage overlapping in wireless sensor networks. In: 5th international ICST conference on communications and networking in China (CHINACOM), 2010
3. Maraiya K, Kant K, Gupta N (2011) Application based study on wireless sensor network. *Int J Comput Appl* (0975–8887) 21(8):9–15
4. Iqbal M, Gondal I, Dooley LS (2007) HUSEC: a heuristic self configuration model for wireless sensor networks. *Comput Commun Science Direct* 30:1624–1640
5. Shirmohammadi MM, Chhardoli M, Faez K, (2009) CHEFC: cluster head election with full coverage in wireless sensor networks. In: *Proceedings of the 2009 IEEE 9th Malaysia international conference*
6. Kim C-J, Lee D-W, Jang K-S (2011) Fault tolerant routing algorithm using the backup-node in WSN. In: *IIEK conference*
7. Hussain S, Rahman MS (2009) Using received signal strength indicator to detect node replacement and replication attacks in wireless sensor networks. In: *SPIE Proceedings on data mining, intrusion detection, information assurance, and data networks security, Orlando, vol 7344*
8. Smolau S (2009) Evaluation of the received signal strength indicator for node localization in wireless sensor networks
9. Ye Y, Hilaire V, Koukam A, Wandong C (2008) A cluster based hybrid architecture for wireless sensor networks. In: *2008 international symposium on information science and engineering, pp 297–302*

Implementation of the Personal Healthcare Services on Automotive Environments

Kabsu Han and Jeonghun Cho

Abstract Personal healthcare devices are made for elderly people and chronic illness that needs a health monitoring continuously. There are several standard specifications for personal healthcare devices to interact each other safely and compatibly. Depending on evolution of technology, the elderly and the chronic disease can have their own personal healthcare devices and they can monitor themselves easily. But some kind of limitation is still remaining in their life even if they have powerful smart devices. They need to drive and go to medical office and drug store but automotive system does not support standard for personal healthcare device. We discuss about several standards of the personal healthcare device and propose several implementations of personal healthcare services to support standard of personal healthcare devices on automotive system.

Keywords Personal healthcare · ISO/IEEE11073 · u-Health · Smart vehicle

1 Introduction

Personal healthcare devices are developed and used for elderly people and the patients who have chronic disease. The people of the world are getting older and a large number of elderly people may use healthcare device for therapies and monitoring. According to World Health Organization (WHO), the numbers of the people who have chronic degenerative diseases are 600 million worldwide and

K. Han · J. Cho (✉)

School of E/E, Kyungpook National University, Daehak-ro 80, Buk-gu, Daegu, South Korea
e-mail: jcho@ee.knu.ac.kr

K. Han

e-mail: kshan@ee.knu.ac.kr

who have metabolic diseases are 90 million. They need personal healthcare devices for monitoring of their disease.

Today, because of the evolution of electronic device, personal healthcare device, which are based on microcontroller technologies and easy to use, is widely spread. To interact with other devices compatibly, standard specifications have developed by several work groups.

The elderly and chronic illness, that need health monitoring and therapies, need to drive a vehicle. Terrible accident can happen when elderly driver gets bad situations of health status. To avoid bad situations, personal healthcare services on automotive environments need to be considered.

Section 2 describes backgrounds for standard specification of personal healthcare services and Section 3 describes personal healthcare services with smart devices. Section 4 considers personal healthcare services on automotive environments. Section 5 describes experimental environment and finally, the last section presents conclusion and future works.

2 Background

2.1 *Continua Health Alliance*

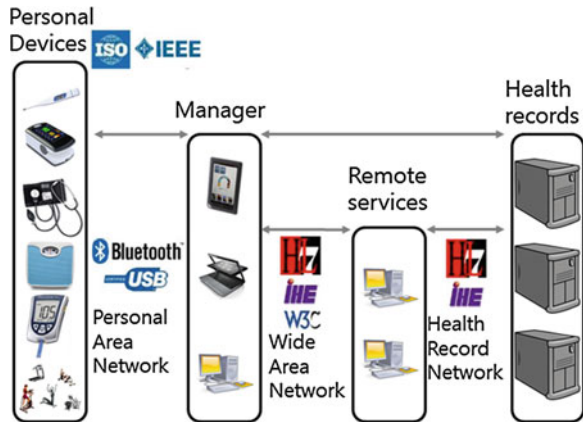
Continua Health Alliance, founded in 2006, is a non-profit open industry organization of nearly 240 healthcare providers, communications, medical and fitness device companies around the world joining together in collaboration to improve the quality of personal healthcare [6, 8]. Continua Health Alliance is dedicated to establishing a system of interoperable personal healthcare solutions and services in three major categories, chronic disease management, aging independently and health-physical fitness.

Continua Health Alliance provides several design guidelines for healthcare devices. Design guideline, version 1 released in 2009, is based on standards of connectivity, which include Bluetooth and USB. Continua Health Alliance products make use of ISO/IEEE11073 for exchange of data. The group provides product certification program to assure the interoperability of applications and transport with other continua certified products.

Figure 1 shows typical service flow and devices of personal healthcare service. Personal healthcare sensor devices which called agent, e.g., pulse oximeter, pulse/heart rate, collect information of people and send the information to a personal health processing device which called manager for processing, display and backup, e.g., PC, cell phone. Manager can send the information which processed within manager to remote system for more services and backup. This is including disease management, health and fitness or aging independently applications.

The communication path between agent and manager is logical point to point connection whatever physical connection is wireless or wired. Generally, an agent

Fig. 1 Overview of continuous healthcare devices and services



connected with a manager at once, called 1:1 communication. A manager can communicate with several agents simultaneously using each point to point connection, called 1:n communication.

2.2 ISO/IEEE11073

IEEE 11073 work group is established to develop new medical device standard specification for typical personal healthcare devices. Personal healthcare device is medical embedded system which has limited functionality and resource typically, and has network interface to connect each other. New medical device standard specifications have to consider about restrictions like weight, cost, convenience, durability and duration [1–3]. The work group develops a common base protocol for initial set of six device specializations (pulse oximeter, pulse/heart rate, blood pressure, thermometer, glucose, weighing scale).

The main scope of the IEEE 11073 standard is the data exchange and interface between the agent and the manager. The data on the connection between the agent and the manager, wireless or wired, may be converted and transmitted across different interfaces. The data exchange through the interfaces is the scope of the IEEE 11073 in application layer but transport interface under the application layer is out of the specification, shown as Fig. 2.

2.3 Transport Interfaces

Various transport interfaces, e.g., Serial, IrDA, Bluetooth, US, ZigBee, which are wired or wireless are used for personal healthcare device [4, 5]. Among them, only USB, which is wired transport interface, and Bluetooth, which is wireless transport

Fig. 2 Area of IEEE 11073 standard

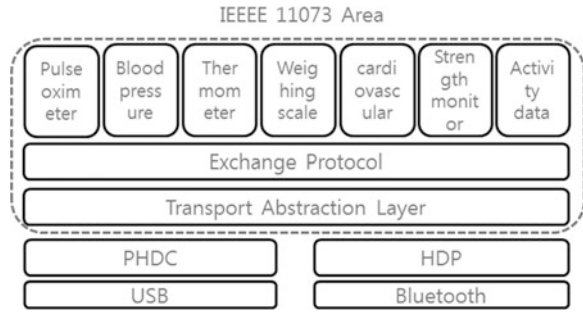


Fig. 3 Transport interfaces

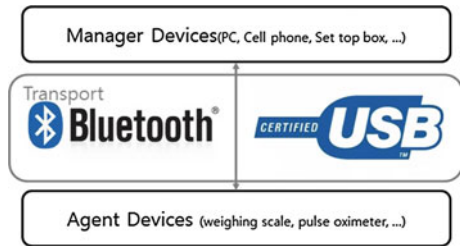
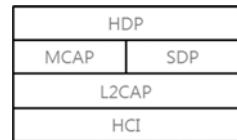


Fig. 4 Health device profile for bluetooth



interface, are certified communication standard in design guideline of Continua Health Alliance [7]. But the ZigBee is certified for Sensor LAN only. Figure 3 shows certified wired/wireless transport interface, called Personal Area Network (PAN).

The benefits of Bluetooth interface are all of the benefits of wireless technology and compliance with global standard in wireless communication. Also, various access points include PDA, cell phone, etc.

The benefits of USB interface are high signaling rate and compliance with global standard in wired interface. Also, USB is dynamically attachable interface and supports backward compatibility.

In 2007 the wireless technology Bluetooth became the first communication standard certified by Continua Health Alliance, the profile layers of Bluetooth is shown as Fig. 4. Looking for a higher data rate option, up to 400 Mbps, Continua Health Alliance keep their attention to USB since the USB work group produces the Personal Health Device Class (PHDC) specification.

In 2008 the Continua Health Alliance approved the PHDC of USB which is specific class for use in home portable medical devices. Communication stack, based on USB’s PHDC and IEEE-11073, provides the standard communication interface for the next-generation medical devices, the class layers of USB is shown as Fig. 5.

Fig. 5 Personal health device class for USB

Mouse	Medical	USB-serial	Storage
HID	PHDC	CDC	MSD
Device Layer			
HCI			

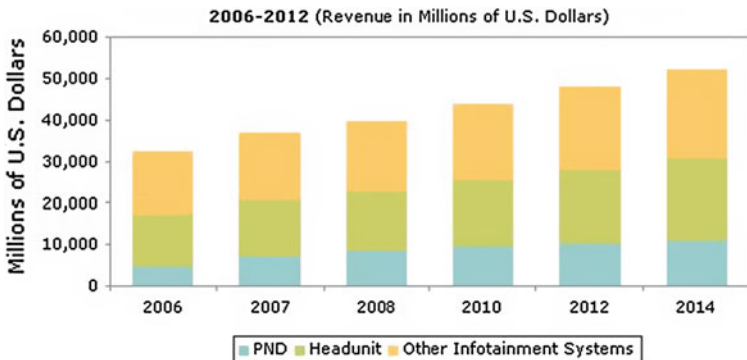
Wireless transport protocol, Bluetooth low energy and USB a fast data rate will continue to be used in personal healthcare devices for years.

3 Automotive Environments

With the development of automotive electronic technology, automotive system has more comfort devices too. Depending on development and propagation of vehicle to vehicle (V2V) communication technology and vehicle to infrastructure (V2I) communication technology, shortly V2X, infotainment system that collects information and provides value-added information to driver has emerged and spread widely, shown as Fig. 6. Also, infotainment system can send various internal information include status of driver to outside via V2X communication. Infotainment system can make emergency call and emergency alarm automatically, if some kind of trouble is happened to vehicle and driver. E.g. OnStar of GM, SYNC of Ford.

Nowadays, the most of infotainment system have various interfaces like USB for flash memory and i-pod, and Bluetooth for headset and hands-free for external expansion.

For safety reason, the research has continued to warn driver or control the vehicle autonomously by detecting status of driver. But in this case, OEM and supplier of automotive electronic system do not consider the personal healthcare device standard specification which mentioned in Section 2 for development at all.



Source: iSuppli Corp., April 2008

Fig. 6 Revenue forecast of infotainment system

They develop their own personal healthcare device exclusively and do not open their healthcare device technology, even do not support standard interface like USB class and Bluetooth profile.

To solve these problems and support standard specification for personal healthcare device interface, propose several ways to extend the personal health device functionality on automotive environment.

4 Experimental

This section describes experimental environment of personal healthcare services and experimental scenarios. Also, describes test scenarios of PAN and services.

4.1 Experimental Environment

For implementation of personal healthcare service on automotive environment, ARM 6410 development kits are used for managers, which have Bluetooth and USB interface for PAN and Wi-Fi interface for WAN. Software of development kit is based on Linux kernel 2.6.36, which has Bluetooth stack, and android 2.0. TI C5505 development kits are used for agents as Bluetooth oximeters. We use a driver's seat environment, which is development environment of automotive E/E system and remote services based on PC, which receive health records and reply requests, shown as Fig. 7. Also, for verification and validation of the personal healthcare service, we use continua certified devices, nonin 9560 bluetooth oximeter, and continua open-source manager.

4.2 Experimental Scenarios

4.2.1 PAN Test

Testing for the PAN interface consist of compliance testing and interoperability testing. Compliance testing includes compliance test suite against ISO/IEEE



Fig. 7 Test environment and result



Fig. 8 Example of streaming data



Fig. 9 Example of dual role streaming device

11073 and applicable ISO/IEEE 11073 and applicable Continua design guidelines. Continua Health Alliance verifies qualification of Bluetooth Health Device Profile and certification of USB Personal Health Device Class also. Interoperability testing includes interoperability test against continua vendor assisted source code.

Test labs for transport of USB PHDC are available on official USB website. But full test labs for Bluetooth HDP/MCAP are not available, only 15 % of tests are available on official Bluetooth website. But many test cases for Bluetooth HDP are classified. HDP devices acting as a source device such as weight scales, which transmit application data over a reliable data channel to a sink. Other source devices such as pulse oximeter, transmit application data over a streaming data channel, shown as Fig. 8.

Multiple source devices transmit application data over reliable and streaming data channels. But we consider only multiple pulse oximeters over streaming data channel and some kind of smart device, called dual role device, can receive data and relay data to manager, shown as Fig. 9.

4.2.2 Services

To implement personal healthcare service on automotive environment, we consider several ways in Section 4. In case of manger, smart device, acting as a manager, need to store health records on local storage device to assure reliability. Because smart device does not have enough storage capacity, health records have to be deleted after transmission of health records. For data synchronization between smart devices, which acting as managers, file synchronization routine based on android is developed. Personal healthcare services on infotainment system have to support multi process environment. If other processes are running on foreground, e.g., navigation, radio and hands-free, personal healthcare service processes have to assure precise services, which include reliable channel connection and transmission of health records, on background.

5 Conclusion

Personal healthcare services on automotive environments can be helpful for elderly people and the chronic disease. It can prevent terrible accident due to health problem of the driver.

Also, because of long-range driving, personal health monitoring is needed for safety. OEM and supplier try to solve this problem for a long time. Now, they have simple answer which can do easily with standard specification for personal healthcare device.

In this paper, we considered standard specifications for the personal healthcare device and implementation on automotive environments as agents and manager of personal healthcare device. But not all of personal healthcare services, only PAN and simple services on automotive environment. Research about implementation of WAN on V2X network is needed. Implementation of HL7, IHL and W3C on WAN will be included. Also, more practical service model, which can combine already existed services on automotive, and theoretical service model, for future technology like a CPS, are needed.

Automotive industry has to accept specification as soon as possible and provide interface which compatible with standard specification. Make it safe more and more. It is impossible to overemphasize safety.

Acknowledgments This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the CITRC(Convergence Information Technology Research Center) support program (NIPA-2012-C6150-1202-0011) supervised by the NIPA(National IT Industry Promotion Agency) and the Ministry of Education, Science Technology (MEST) and National Research Foundation of Korea(NRF) through the Human Resource Training Project for Regional Innovation and Kyungpook National University Research Fund, 2012.

References

1. Lim J, Park C, Park S (2010) Home healthcare settop-box for senior chronic care using ISO/IEEE11073 PHD Standard. In: EMBC'10
2. Clarke M, Bogia D (2007) Developing a standard for personal health devices based on 11073. In: EMBC'07
3. Fioravanti A, Fico G (2010) Integration of heterogeneous biomedical sensor into an ISO/IEEE 11073 compliant application. In: EMBC'10
4. Martinez I, Escayola J (2008) Standard based middleware platform for medical sensor network and u-health. In: ICCCN'08
5. Waluyo AB, Pek I, Chen X, Yeoh W (2009) Design and evaluation of lightweight middleware for personal wireless body area network. *J Pers Ubiquit Comput* 13:509–525
6. Integration of heterogeneous biomedical sensors into an ISO/IEEE11073 compliant application, Continua design guideline, Continua Health Alliance
7. Building a linux-based continua compliance software stack, Intel
8. Recommendation for Continua USB PHDC device driver interoperability, Continua Health Alliance

A Design of WSN Model to Minimize Data-Centric Routing Cost for Many-to-Many Communication

A. S. M. Sanwar Hosen and Gi-hwan Cho

Abstract Wireless sensor networks (WSNs) differ from traditional networks in several ways: sensor nodes have severe energy constraints, redundant low-rate data, and many-to-many flows. This paper deals with a data-centric merging and by-passing routing schemes based on our own network model. This model permits to gather data at a tuple-centroid node for merging the same type of data. The data-centric merging mechanism is an efficient approach which reduces the number of hops per source and to by-pass the different types of data. It is competent to reduce the unnecessary processing cost of an intermediate node. Our network model shows offers significant performance gains across the high routing cost environment.

Keywords Policy wireless sensor network • Spanning tree • Tuple • Routing costs

1 Introduction

A wireless sensor network (WSN) is generally used to acquire information in various physical environments. The tiny MEMS based sensor nodes forming WSN are resource constraints, i.e., lack of battery supply, low computational capabilities, and insufficient memory space, even the available bandwidth. Therefore,

A. S. M. Sanwar Hosen
Division of Electronics and Information Engineering, Chonbuk
University, Jeonju, South Korea
e-mail: sanwar@jbnu.ac.kr

G. Cho (✉)
Division of Computer Science and Engineering, Chonbuk University, Jeonju, South Korea
e-mail: ghcho@jbnu.ac.kr

WSNs require passing the data cooperatively through the network to sink. Meanwhile, data from sensor nodes are correlated in terms of time and space, transmitting only the required and partially processed data is more meaningful than that of sending a large amount of raw data.

There are two network scenarios familiar in WSNs: one-to-one communication and many-to-many communication. In one-to-one communication, normally sensor nodes report its data to a single sink [1]. It is sufficient to find out the shortest route toward the sink in order to transmit data from a source node. On the other hand, in many-to-many communication, usually multiple sensor nodes need to transmit data to the multiple sinks. Both of these need to build multiple independent trees from source to sink(s). That is, the network implies to have a linear hop-by-hop data forwarding mechanism. As a result, more nodes might be involved in its data forwarding route corresponding to source-sink.

Moreover, more hops might be increased the unnecessary processing costs of the intermediate nodes. Also, in a linear hop-by-hop network, it is impractical to use the Time Division Multiple Access (TDMA) for avoiding collusion and saving energy of deployed nodes. Generally, the intermediate nodes are not aware of the merging mechanism in data routing. This situation will cause high cost routing, and then bring about early death of constituted nodes in the network, consequently make short the network lifetime as a whole.

In the WSN environment, we introduce a new network model that is efficient in terms of minimizing routing cost. To begin with, we propose that a network would be partitioned into $\delta(SP_T)$ consisting of n -tuples. Among the tuples, each tuple will be formed with optimal number of tuple-nodes based on the balance traffic load. Then, the nodes which represent a minimum routing cost within the tuple are elected as tuple-centroid nodes. These nodes act as a merging and/or an intermediate node to forward data further. Here, data merging with the same type is preferred in the tuple-centroid nodes. Otherwise, the nodes will simply act as a gathering and forwarding (by-passing) node.

2 Related Works

Existing challenges in WSN require a well-consolidate network design based on the application implementation scenario. Meanwhile, the network requires to minimize the energy consumption in every factor of its design architecture. Many routing protocols already have been introduced in the last decade. For instance, some are location-based protocols MECN [2] and TBF [3], where sensor nodes are identified by means of their location. From the location information, most of them can estimate the energy consumption in their routing mechanism. The data centric protocol is one of the efficient techniques to reduce the extra computational costs. For instance, in the protocols as DD [4] and REEP [5], the aggregated data from multiple source nodes are routed to the sink in order to save on transmission costs. Most works have explored a hierarchical clustering from different perspectives as

described in LEACH [6] and HEED [7]. Clustering is an effective method to group the communication paths.

Considering a data transmission in-between a source and a sink, there are two routing paradigms: single-path routing and multipath routing. In single-path routing, each sensor sends its data to the sink via the shortest path. In multipath routing, each source finds the first k shortest paths to the sink and divides its load evenly among these paths as described in DP [8] and N-to-1 MD [9].

Our proposed network model is very similar to the data centric, hierarchical, and multipath based protocol. To the best of our knowledge, the existing protocols are not well-suited for energy efficiency. For instance, data centric protocols emphasize the aggregation of the same type of data, whereas our model permits different types of data. Therefore, it may decrease the computational costs for processing at the intermediate nodes. In hierarchical protocols, a cluster header node is elected as to aggregate data from its member nodes for further forwarding to the sink. Whereas, our model makes use of a tuple with tuple-centroid node to permit the most probable scope of merging and/or aggregating data. It permits the constituted nodes are involved with the different tasks in a tuple. Moreover, our method could select the multiple minimum routing cost paths in the network as a whole. As a result, our approach is fully competent to accomplishing the energy efficient routing in one-to-one and many-to-many communications simultaneously.

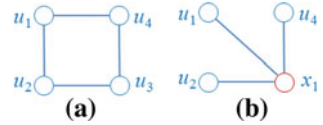
3 Formation of a New Network Model

Our work provides a comprehensive network model inspired from the Spanning Trees (STs) [10]. A spanning tree for a network is a sub-graph that contains all the sensor nodes of that network. There are many scenarios to find out a minimum routing cost spanning tree. Whenever anyone wants to find simple, cost reliable, and even well-organized way to link a set of deployed nodes for many-to-many communication, a prominent solution is normally to make use of a kind of spanning tree. Therefore, we are interested to build a sub-tree to minimize inter-communication routing cost, and then to find paths to minimize the communication cost among sub-trees. Eventually, the network aims to support one-to-one communication and many-to-many communication simultaneously. Before going into details to design our network model, we are willing to describe some important terms such are as following:

3.1 Routing Cost Estimation

A tuple is an ordered set of nodes. An n -tuple is an ordered of n nodes, where n is a positive integer. If $n = 4$ is called a 4-tuple of nodes (u_1, u_2, u_3, u_3) as shown in Fig. 1a. The tuple-centroid node is a centroid node in routing within this tuple to

Fig. 1 An example of an n -tuple and tuple-centroid node



connect all member nodes (tuple-nodes). For example, x_1 in Fig. 1b stands for the tuple-centroid node.

3.2 Routing Cost Estimation on Tuple

In general, when the link cost represents a cost for routing messages in-between two nodes (such as transmission cost), the routing cost for a pair of nodes in a given spanning tree is defined as the sum of the costs of the links in the unique tree path between them. The routing cost of the tuple itself is the sum over all links of nodes of the routing cost for the link in this tree. Therefore, we can derive the routing cost $Cost(t)$ of a tuple as follows:

$$Cost(t) = \sum_{e \in E(t)} l(t, e)w(e) \tag{1}$$

Where $e \in E(t)$ denotes to a set of links between the nodes, $l(t, e)$ denotes to the number of bits in a message, $w(e)$ denotes to the weight assigned to each link (i.e., transmission cost based on the distance of (u, v)).

3.3 Routing Cost Estimation on Tuple-Connecting Edge

The routing cost on a connecting edge in-between two tuples is the sum over all possible connecting links contained in that tuples. In general, we can define the routing load on a connecting edge e as follows:

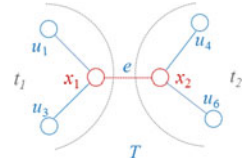
Definition 1 Let T be a tree and $e \in E(T)$ is an edge on the tree. Assume t_1 and t_2 are two tuples that result by removing e from T . The routing load on the edge e is defined by $l(T, e) = 2|V(t_1)||V(t_2)|$

From Definition 1, in Fig. 2, the routing cost on the connecting edge $e \in (x_i, x_j)$ in-between two tuples t_i and t_j can be defined as;

$$Cost(e) = l(T, e)w(e) \tag{2}$$

where $l(T, e)$ denotes to the number of bits in a message on that connecting link, $w(e)$ denotes to the weight assigned to this link (in-between two tuple-centroid nodes).

Fig. 2 An example of a tree T of two tuples



3.4 Routing Cost Estimation on Forwarding Paths

The routing cost of a forwarding path in a network is the sum over all connecting link(s) on that path in-between the source and the destination. Therefore, the routing cost of a forwarding path $Cost(P)$ can be defined as follows:

$$Cost(P) = \sum_{e \in E(P)} l(P, e)w(e) \tag{3}$$

Where $e \in E(P)$ denotes to a set of links between the nodes on a path, $l(P, e)$ denotes to the number of bits in a message needed to be forward, $w(e)$ denotes to the weight assigned to each forwarding link(s) (in-between intermediate nodes (u, v))

4 Network Design

4.1 Network Partition into $\delta(SP_T)$

To generalize the idea of $\delta(SP_T)$ is partition the network into sufficiently small n -tuples. The n -tuple is a partition of that network which contains the number of nodes based on the balance traffic load. Therefore, how many tuples and tuple-nodes need to be selected can be defined in equation (4). For this, we assume $A(t)$ is the total amount of data can be generated in a particular time in the entire network, $info$ for set of different types of data in the network contains the data set $\{a_1, a_2, \dots, a_n\}$, t_c for the maximum transmission capacity of a node, and avg_a for the average data size of a node u_i in the network. Therefore, the number of nodes that are suitable to group into a tuple can be defined as follows:

$$n - tuple = \left\{ \frac{A(t)}{t_c} \right\} / avg_a(u) \text{ Where } A(t) \in \text{info} \tag{4}$$

By using Eq. (4), we can partition the entire network in the following equation based on balance traffic load, where $\delta(SP_T)$ for the total number of tuples in the network and $V(T)$ for the set of deployed nodes.

$$\delta(SP_T) = \frac{1}{n - tuple} |V(T)| \tag{5}$$

4.2 Tuple-Centroid Node Election

A tuple centroid node is elected among the constituted tuple-nodes in a tuple based on the routing cost estimation described in Eq. (1). To evaluate the internal routing cost within a tuple, we use $l(t_i, e)$ to denote the load of transmitting bits/data in a message, $w(e)$ to denote the assigned weight of the transmission cost of bits on that distance (u, v) , and A_i for the collected total data from a particular tuple. The node $x_i = u_i$ is the tuple-centroid node among the tuple-nodes. For example, the nodes $\{x_1, x_2, \dots, x_n\} = \{u_2, u_5, \dots, u_n\}$ belong to tuples $\{t_1, t_2, \dots, t_n\}$ in the network shown in Fig. 3. Here, we can obtain the minimum internal routing cost $Cost(t_i)$ defined as follows:

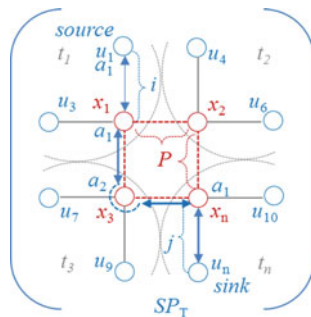
$$Cost(t_i) = \min \sum_{e \in E(t_i)} l(t_i, e)w(e) \tag{6}$$

where $\{l(t_i, e) \in A_i(t_i)\} \leq t_c$ and $w(e) \in d_T(x_i, u_i)$

4.3 Data Merging and By-passing at the Tuple-Centroid Node

In data routing, the data forwarding path from a source to a sink should be constructed based on the n -tuple concept. This means that the tuple-nodes are connected at their own tuple-centroid node, and the elected tuple-centroid node is the data merging/ by-passing point for this formed tuple. We assume the different types of data $\{a_1, a_2, \dots, a_n\} \in info$ belonged to a set of nodes $\{u_1, u_2, \dots, u_n\} \in V(T)$ can be generated within a tuple. In our network model, the tuple-centroid nodes have two types of forwarding strategies. The tuple-centroid node x_i is receiving the data from a constituted tuple-node $a_i(u_i)$, that is belonging to $a_i(x_i) \in info_i$ would be merged together for further

Fig. 3 An example of a path construction that supports one-to-one and many-to-many communications in 3-tuple networks



forwarding to an intermediate node or directly to the sink(s). On the other hand, the received data from a constituted tuple-node $a_j(u_j)$, which is not belonging to $a_j \notin info_i$ would be simply by-passed by the tuple-centroid node x_i to the intermediate node x_j or directly to the sink(s). For example, the data of $a_1(u_1)$ is merging at tuple-centroid nodes $\{x_1(a_1), x_n(a_1)\}$ and by-passing at tuple-centroid node $x_3(a_2)$ towards the sink node u_n as shown in Fig. 3.

4.4 Data Forwarding Path Selection

Institutively, a forwarding path P is the connected general path which contains a set of tuple-centroid nodes as intermediate nodes. Starting from any tuple-centroid node, there a sufficiently amount of tuple-nodes which can only be reached after passing the tuple-centroid node.

For example, Fig. 3 shows two nodes, u_1 and u_n , as the source and destination respectively in different tuples $t = \{t_1, t_2, \dots, t_n\}$ connected by the path $P = \{P_1, P_2, P_3, \dots, P_n\}$. The path between them can be divided into three sub-paths: from edge $e(u_i, x_i) \in i$, the paths in $\{x_1, x_2, \dots, x_n\} \in P$, and the edge $e(x_j, u_j) \in j$.

If the equality of hop distance is $d_T(u_i, x_i) = d_T(x_j, u_j)$ in different tuples, from Eq. (8), the tuple-centroid node can choose a minimum cost path in data routing from the initiator node to the destination node defined as follows:

$$Cost(P_i) = \min\{2n \sum_{u \in V(t_i)} d_T(u, P) + \sum_{x_i, x_j \in V(P)} d_T^P(x_i, x_j)\} \tag{7}$$

From the definition of the routing cost and by using Eqs. (2) and (3), we can derive the path cost which obtains in path P as;

$$Cost(P_i) = \sum_{e \in E(P)} l(P_i, e)w(e) \tag{8}$$

Therefore, we can calculate the minimum routing cost from Eqs. (6) and (8) for the entire network, where $Cost(t)$ contains the cost of tuples $\{t_1, t_2, \dots, t_n\}$ and the $Cost(P)$ contains the cost of paths $\{P_1, P_2, \dots, P_n\}$ as follows:

$$Cost\delta(SP_T) = \min\{\sum_{i=1}^n Cost(t) + \sum_{i=1}^n Cost(P)\} \tag{9}$$

5 Conclusions and Future Work

Our work presents a novel network model motivated from the spanning trees, where it partitions the network into the optimum number of tuples. The network

partitioning is based on the load balance in data gathering and in merging at the tuple-centroid nodes as well as the overall communication costs which rely on the minimum routing costs tuple formation. The tuple-centroid node election in each tuple and minimum cost path selection in data routing play an important role that reflects on designing an efficient routing mechanism in the aspect of the network lifetime. As a future plan, we would like to demonstrate this model in terms of performance.

References

1. Mottola L, Picco G (2011) MUSTER: adaptive energy-aware multisink routing in wireless sensor networks. *IEEE Trans Mob Comput* 10(12):1694–1709
2. Rodoplu V, Meng T (1999) Minimum energy mobile wireless networks. *IEEE J Sel Areas Commun* 17(8):1333–1344
3. Nath B, Niculescu D (2003) Routing on a curve. *ACM SIGCOMM Comput Commun Rev* 33(1):155–160
4. Intanagonwiwat C, Govindan R, Estrin D (2000) Directed diffusion: a scalable and robust communication paradigm for sensor networks. In: *Proceedings on ACM MobiCom*, pp 56–67
5. Zabin F, Misra S, Woungang I, Rashvand H, Ma N, Ali M (2008) REEP: data centric, energy-efficient and reliable routing protocol for wireless sensor networks. *IET Commun* 2(8):995–1008
6. Lindsey S, Raghavendra C (2002) PEGASIS: power-efficient gathering in sensor information systems. In: *Proceedings of the Conference on Aerospace*, pp 1125–1130
7. Younis O (2004) HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. *IEEE Trans Mob Comput* 3(4):366–379
8. Lindsey S, Raghavendra C, Sivalingam K (2001) Data gathering in sensor networks using the energy delay metric. In: *Proceedings of the 5th International conference on parallel and distributed processing symposium*, pp 2001–2008
9. Chu M, Haussecker H, Zhao F (2002) Scalable information-driven sensor querying and routing for ad hoc heterogeneous sensor networks. *J High Perform Comput Appl* 16(3):293–313
10. Wu B, Chao K (2004) *Spanning trees and optimization problems*. Chapman & Hall, Boca Raton

Improving of Cache Performance using Trust Value of Nodes

Seung-Jae Yoo and Hwan-Seok Yang

Abstract Caching scheme is important to improve data access performance and decrease bandwidth in MANET which consists of only MN. Many study like to solve cache consistency problem in this caching technique is achieved, but most of scheme is exposed to various security attacks. In particular, malicious node causes error from receiver through insertion and modification of stale data. In this paper, we proposed the technique which removes threat factor for cache consistency through authentication process using trust check about node. Cache table management scheme based cluster is applied to decrease overhead and manage efficient cache as MN discovers cache. We identified that proposed scheme in experiment result showed improved performance in average query latency and cache hit ratio.

Keywords Cache consistency · Mobile ad-hoc network · Trust value

1 Introduction

Wireless mobile communication among communication industry field is developed very rapidly and specific gravity of wire network which is using in wide field decreases remarkably because of this. This wireless network can divide to

S.-J. Yoo · H.-S. Yang (✉)

Department of Information Security Engineering, Joongbu University,
101, Majeon-ri, Chubu-myeon, Geumsan-gun, Chungnam, South Korea
e-mail: yanghs@joongbu.ac.kr

S.-J. Yoo

e-mail: sjyoo@joongbu.ac.kr

structure based infrastructure like AP and structure without support of infrastructure. Mobile Ad-hoc Network (MANET) is network which is composed to only Mobile Nodes (MNs) without support of any infrastructure and every node has to perform delivery function like router [1, 2]. MANET is not easy to maintain route and transmission delay time is long because network topology changes frequently by MNs [3–5]. Caching scheme which MN can access rapidly to requested data is very important than routing scheme for connection of nodes to use efficiently limited resource of low bandwidth and MNs [6]. Most of caching schemes can cause error by malicious node because it is exposed to various attacks.

In this paper, we proposed authentication process scheme which check trust value about nodes to prevent modification of cache data by malicious nodes and maintain cache consistency. Threat factor like insertion, modification of stale data by malicious node is removed though this. Network is composed to cluster form to manage reliability of nodes and cluster head uses member cache table and cluster cache table to trust management of member nodes and caching.

The rest of the paper is organized as follows: Sect. 2 reviews existing caching scheme. Safe caching scheme through the proposed trust evaluation in this paper is presented in Sect. 3. Section 4 evaluates efficiency of proposed scheme through comparison experience with ZC and COOP scheme. The conclusion is presented in Sect. 5.

2 Related Works

Zone Cooperative Cache [7] forms zone by nodes which are 1-hop distance and is cooperating structure. This technique composes zone to decrease message exchange and energy consumption. Each node has cache to save accessed data frequently. Cache strategy checks own local cache when nodes receive request of data. The next step is that it checks consistency if requested data exists. The data is transmitted to node which request data if it pass through consistency examination. If not, it request effectiveness examination to other node. But requested data doesn't exist to own local cache, retrieval is performed again in home zone which itself is. It broadcasts request to neighboring zone if it doesn't retrieve requested data even here. Location of node which has requested data information is detected after this process. Request node replaces own cache if response packet about requested data is received. Cache replacement policy uses VALUE. When cache is replaced, VALUE deletes data which has the lowest value from cache. It calculates approach probability from host to data.

Group caching [6] is that each node transmits hello message periodically and form group with neighborhood nodes which is 1-hop distance. Master node exists in Formed group like this and communicates with member nodes directly. It use cache control message and check caching condition of group member nodes to utilize cache space of each node in group. Redundancy of saved data in cache

space become low and data approach can be improved because of this method. Member nodes use self table and group table in order to save caching state. The first own self table is retrieved if MN receive data request. Member nodes request to group member in group table if it doesn't retrieve wanted information here. Request message is transmitted to neighboring group in order to retrieve information of wanted data by method like this. Figure 1 is shown cache retrieval process of group caching.

Energy consume of nodes is increased because hello message is sent periodically to update information about cache state of neighboring node and nodes which leave or come in this method. It has a disadvantage that performance of network decrease because quantity of control message increases.

3 Proposed System Model

In this section, we describe caching scheme which can reduce threat about cache consistency through trust value examination of nodes and overhead and bandwidth when cache is discovered.

3.1 Architecture

Cluster is used in this study in order to prevent increase of cache update by MNs and perform management of cache table efficiently. Cluster head is elected based number of connection after nodes which consist of network compose cluster based

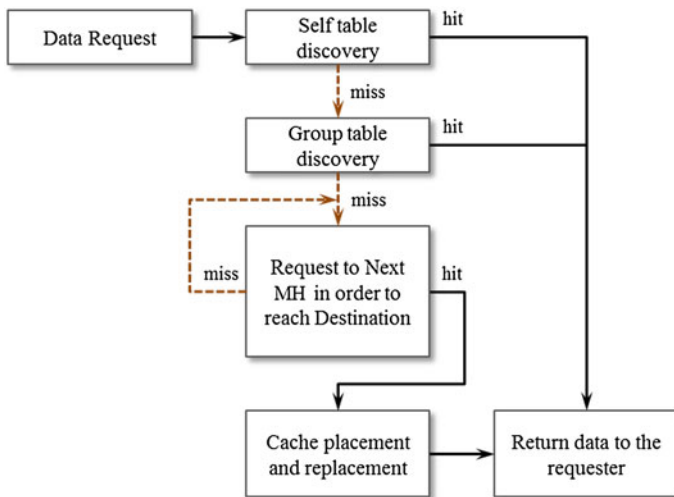


Fig. 1 Process of a request by GC

on neighboring nodes which are 1-hop distance. We can decrease overhead for cache discovery by MNs because cluster head connects with many nodes. Cache table management is easy by using cluster scheme.

3.2 Cache Table for Cache Consistency

Every node which consists of network have self cache table (SCT). SCT consist of data info and saved time field. Cluster head has member cache table (MCT) and cluster cache table (CCT) in order to offer nodes efficient information. Member nodes causes consistency problem when many nodes save same data or malicious node transmits wrong cache. Trust value is used to prevent this. Here it means packet transmission ratio of nodes and is managed by cluster head. Nodes which do malicious act can be blocked because the more participation ratio about packet transmission is high the more trust value increases. MCT has detailed information about cached data item by member node and CCT save cache data information of neighboring cluster in the cluster head. Figure 2 is shown structure of each cache table.

3.3 Cache Discovery and Update

Figure 3 shows the processing for cache discovery. Node confirms first own SCT to discover cache about data D. If Cache of data D is discovered in SCT, if the value that time which receive data request subtracts saved time of the data D is small than TTL, the data regards as valid and information of the data is transmitted. Cluster head is increased trust value of the node in MCT. If the value is large than TTL, every information of the data D is deleted from SCT and request about data D is sent to cluster head. Cluster head which receives this check MCT which information about MNs is held. Trust value of the node is confirmed if requested data D is found. If trust value is large than threshold of trust value of nodes in MCT, information of data D is sent to MN. Here threshold means average trust value in MCT. If trust value is small than threshold, information of the node

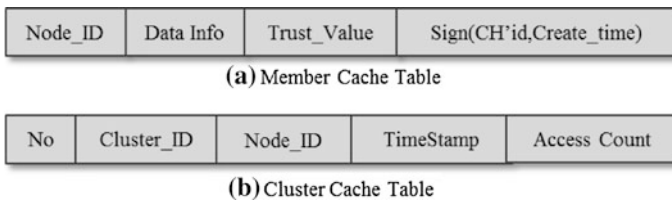


Fig. 2 Structure of cache table

is deleted and request query is sent to neighbor cluster head to using CCT. Neighboring cluster head discover cache of data D in the same way and cluster head which receives data do update MCT. Node which receives requested data D information from cluster head do update own SCT.

4 Performance Evaluation

In this section, we evaluated performance of proposed scheme through comparison experiment with existing ZC and COOP.

4.1 The Simulation Model

Ns-2 simulator is used and experiment time is 600 s in this simulation. The number of used node is 50,100 and mobility model used random way point. The simulated traffic is Constant Bit Rate (CBR) and the transmission range of MNs is 250 m. The simulation parameters are listed in Table 1.

4.2 Simulation Result

Evaluation metrics are cache hit ratio, average query latency and response reject ratio to performance evaluation of proposed scheme in this paper.

Figure 4 shows cache hit ratio according to different cache size and the number of node. The more cache size increases, the more cache hit ratio increases like

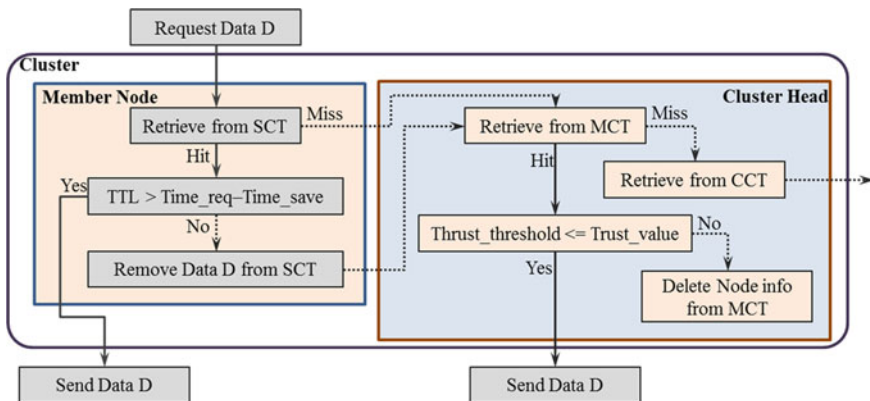


Fig. 3 Query processing of requested data D

Table 1 Simulation parameters

Parameter	Values
Network size	1500 × 1500
MN Speed	0–10 m/s
Transmission range	250 m
Bandwidth(MB)	2
Pause time(sec)	20
Cache size	200–1200
TTL(sec)	100–300
Average query rate(sec)	0.5

shown in figure. When the number of node is many, cache hit ratio also appears highly. ZC scheme is out of result because it is not cooperative structure to discover cache. COOP scheme shows a little good result because it uses cache table of group members. Proposed scheme shows the best performance because it uses trust value to remove factor which threats cache consistency and cache table management by cluster head is performed.

Figure 5 shows average query latency. ZC scheme has long latency because request signal is sent to zone which itself belong and wait response when MN receives data request. COOP scheme shows good result than ZC scheme because table which itself has is checked when data request receives. But latency to retrieve this presents a little high when requested data in table is not retrieved. The more cache size a decrease, the more latency is high because retrieval time of cache table is long.

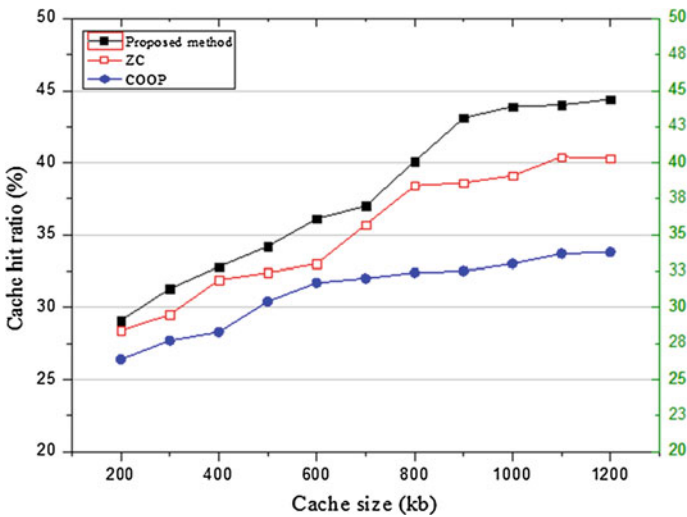


Fig. 4 Cache hit ratio as different cache size

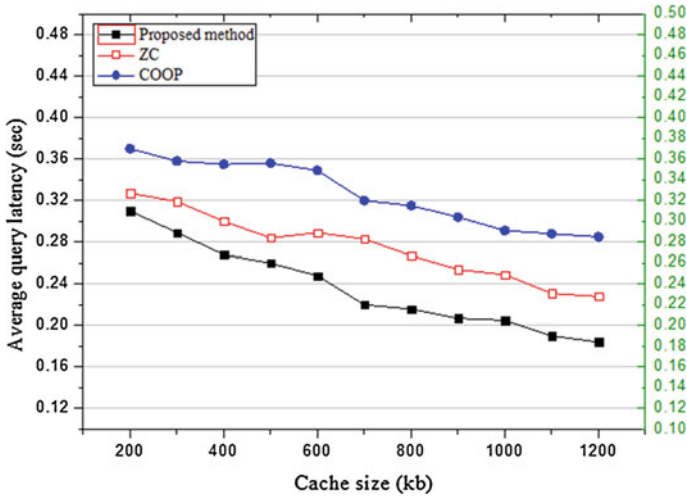


Fig. 5 Average query latency as different cache size

Figure 6 shows response reject ratio. Trust value is small than threshold in MCT which is managed by cluster head. This is ratio which response about cache request is rejected. Response rejection ratio is appeared a little highly because the more cache size is small the more cache check request is many. Cache data modification of malicious node can be blocked by excluding response of nodes which has low trust value.

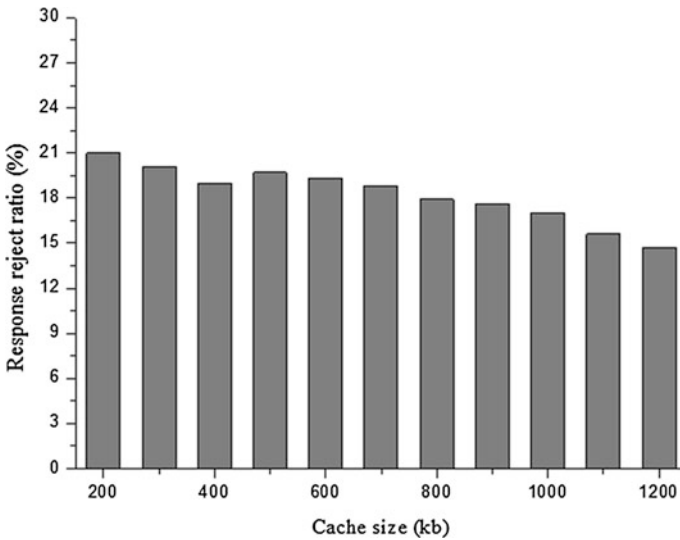


Fig. 6 Response reject ratio as different cache size

5 Conclusion

In this paper, we proposed caching scheme to maintain consistency of cache data from threat of malicious nodes, decrease overhead for cache discovery, and improve data availability. Trust value of every nodes is managed by cluster head. Cluster head has MCT which has information about member nodes in cluster and CCT which has information about neighboring cluster. Information is offered by cache request using this. Response of the node less than trust value threshold of MCT which manages by cluster head is excluded when cache request is received. Safety about cache consistency is guaranteed and performance of caching scheme can be improved.

References

1. Toh CK (2002) Ad hoc network wireless networks. Protocol and system, Prentice Hall PTR, New York
2. Hara T (2001) Effective replica allocation in ad hoc networks for improving data accessibility. In: Proceedings of INFOCOM, pp 1568–1576
3. Lim S, Lee WC, Cao G, Das CR (2006) A novel caching scheme for improving internet based mobile ad hoc networks performance. Elsevier J Ad Hoc Netw 4(2):225–239
4. Cao G (2003) On improving the performance of cache invalidation in mobile environments. IEEE Trans Knowl Data Eng 15(5):1251–1265
5. Cao G, Yin L, Das C (2004) Cooperative cache based data access framework for ad hoc networks. IEEE Comput Soc 37(2):32–39
6. Ting Y-W, Chang Y-K (2007) A novel cooperative caching scheme for wireless ad hoc networks: group caching. In: IEEE international conference on networking, architecture, and storage
7. Chand N, Joshi RC, Manoj M (2011) A cooperative caching strategy in mobile ad hoc networks based on clusters. Int J Mob Comput Multimed Commun 3(3):20–35

Design of Microstrip Patch Antenna for Mobile Communication Systems Using Single-Feed

Chanhong Park

Abstract Efforts to downsize antennas were mostly made in microstrip antenna with great success and progress in a wide variety of fields. Using dielectric substance is one way to downsize antennas but it hinders such antenna characteristics as antenna bandwidth and radiation efficiency. Because of such limitations, there is an ongoing research to modify antenna structure. Radiation pattern of small antennas becomes nearly non-directional and antenna gain becomes lower. Furthermore, bandwidth gets narrower because of weakened input resistance and extremely high reactance. Developing a small-sized antenna that is free of such shortcomings is not easy and the major job is to downsize antennas while matching impedance to protect its characteristic. In this paper, we proposed novel particle swarm optimization method based on IE3D is used to design a mobile communication Microstrip Patch Antenna. The aim of the thesis is to Design and fabricate an inset fed rectangular Microstrip Antenna and study the effect of antenna dimensions Length (L), Width (W) and substrate parameters relative Dielectric constant (ϵ_r), substrate thickness on Radiation parameters of Band width. When the antenna was designed, a dual-band, dual-polarized antenna was used to secure the bandwidth and improve performance, and a coaxial probe feeding method so that the phased array of antenna is easy.

Keywords Microstrip patch antenna • Mobile communication antenna • PSO • IE3D

C. Park (✉)

SamSun Technology Research Co. Ltd., Company-affiliated Research,
564-6, Sang-dong, Wonmi-gu, Bucheon-si, Gyeonggi-do, Korea
e-mail: iparka1028@gmail.com

1 Introduction

As the society develops information-based and in a desperate need of the wireless technology, development in the field of mobile communication is on the rise, with the smaller but powerful mobile phones in conjunction with the user's needs. As the size of antenna matters most in PCS, GSM, PHS, and any other satellite iridium services, as well as for use in airships and spacecrafts where aerodynamic disturbances are incorporated. As for the mobile phone, it has been developed in the perspective of compactness, function, lightness, and lower power use, with use of the smaller antenna likely affecting the speech quality and design scheme of the device. Having developed for the said reason and with the consistent demand in the compact antenna, varied type of compact antennas came out, including microstrip antenna [1]. Use of dielectric substance, having suggested as a solution for compact antenna, is deemed affecting the bandwidth, radiation efficiency, and other antenna performances. Another alternative involving use of a circuit board with thickness, lower dielectric constant, and parasitic elements to improve bandwidth, inevitably enlarges the size of antenna [2]. Compact antenna features omni-direction, lower antenna gain, lower input resistance, and higher reactance all of which shrinks bandwidth, challenging use of Chip Technologies or ceramic elements with higher dielectricity factoring in the loss in dielectric substance volume. Impedance matching is thereby the name of the game for compact antenna, while maintaining the originality thereof. In that regard, microstrip antenna is regarded as an optimal solution [3]. Attachable to plain surface or on curvature, microstrip antenna features cheap manufacture cost, with use of photolithograph technology. Also applicable to monolithic microwave integrated circuit (MMIC), this varies, as necessary, resonant frequencies, polarization, pattern, and impedance at the phase of manufacture by way of insertion of pin diode, varactor diode, and other active elements between patch and surface contacting. Such a microstrip antenna, however, shall bear lower efficiency, higher electricity use, higher selectivity (Q), lower polarization purity, lower directivity, unnecessary feeding radiation, and narrow bandwidth all of which are deemed setback in developing compact antenna [4, 5].

Narrow bandwidth often resolved by the thicker circuit board and lower dielectricity, often occurring surface wave affecting the radiation efficiency of antenna and unintended radiation pattern. Antenna feeding and grounded part of the circuit board could leak the electricity, likely distorting the radiation orientation toward the grounded surface. Higher mode also distorts impedance characteristics and radiation pattern, needing another solution to adjust resonant frequencies of microstrip antenna, which is so called Linear-frequency Transposition. Also discussed as an alternative is a dual-resonance, multi-band antenna to feature duality in frequency transposition. As a means to expand bandwidth, this is widely applied to make an antenna to be able to operate services falling into different bandwidths for better efficiency in operation, with use of a single antenna to cover multiple bandwidths as well as seamless electricity supply. One example of such an antenna

is ‘Stacked Antenna’ [6, 7], layering multi-layer patches resonating within the different frequencies. Stacked Antenna is operable under the broadband condition when such frequencies work distantly each other and can be compacted by way of ground wire around the antenna edge, locating slots, magnetic substances, and insertion of slots into the ground surface [8–10]. Such an insertion is done with the use of four upswept slots or a pair of widthwise/lengthwise slots along the edge [11]. Stated hereunder has incorporated IE3D Simulator to apply parameters with accuracy, in manufacturing a couple of antennas for the next-generation mobile network communication. The square-shape single-feeding microstrip patch antenna herein has been designed for LTE use. Note that LTE is aiming, for data transfer at 20 MHz bandwidth, downlink and uplink data transfer rate of 100 Mbps and 50 Mbps or 30 Mbps and 15 Mbps for in-motion (at 120 km/h) data transfer rate, respectively. Antenna for LTE use is now under development, on the basis of array antenna and MIMO antenna previously used for the existing base station. As an improvement thereof, square-shape microstrip patch antenna for mobile use has been presented herein, with the use of ‘micro wire feeding method’ to facilitate circular polarization for higher antenna gain.

The rest of this paper is organized as follows. Section 2 describes designed the microstrip patch antenna for mobile communication using sing-feed. Section 3 describes the simulation result of antenna. The conclusions are given in Sect. 4.

2 Microstrip Patch Antenna for Mobile Communication

2.1 Antenna Design and Simulation Environment

The three following parameters are absolutely necessary in designing microstrip antenna for mobile communication. Resonance frequency, which has to be appropriately selected, was set at 2.12 GHz to run the antenna within this frequency range since mobile communication system moves within the frequency range of 2.1–5.6 GHz. For dielectric substance, an oxidized aluminum with 9.8 in permittivity was selected for designing mobile communication square microchip patch antenna since permittivity in panels used for antenna design is around 2.2–10. This is maximum calculation of dielectric substance panel constant to downsize antenna (Table 1).

Table 1 Design parameter of antenna

parameter	value
Resonance frequency (f_0)	2.12 GHz
Dielectric constant (ϵ_0)	9.8
Height of dielectric substance board (h)	1.58 mm
Velocity of light (c)	3×10^8 m/s

Table 2 Simulation parameter

Parameter	Value (mm)
Width (w)	30.7
patch length (l)	23.2
Effective permittivity (ϵ_{eff})	8.89
Effective length of Antenna (L_{eff})	23.9

Set value required to design patch antenna was calculated based on design parameters in Table 2 and PSO algorithm was applied to generate the remaining variables.

Antenna proposed in this paper first located the optimal power feed point via single feed point and each constant value was set to determine optimal length of antenna patch and realized linearly polarized wave. Power feed points were located diagonally to the patch. Two orthogonal degenerated modes produced by square patch to radiate circular polarization should always be in diagonal line against perturbation segments (Fig. 1).

Antenna designed in this Chapter uses multi-angular, polyhedral mesh creation method with which to design fixed points connecting antenna feed line and the remaining antenna structure connecting microstrip feed line. Cells per wavelength (CPW) determining mesh density is set at 30 since frequency of the designed antenna is 2.12 GHz (simulation becomes more precise when the number of

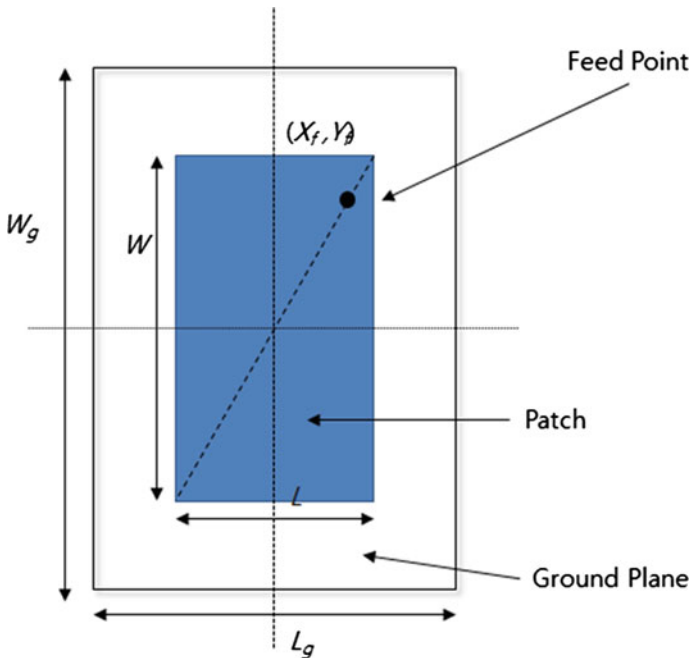


Fig. 1 Design of microstrip patch antenna using single feed

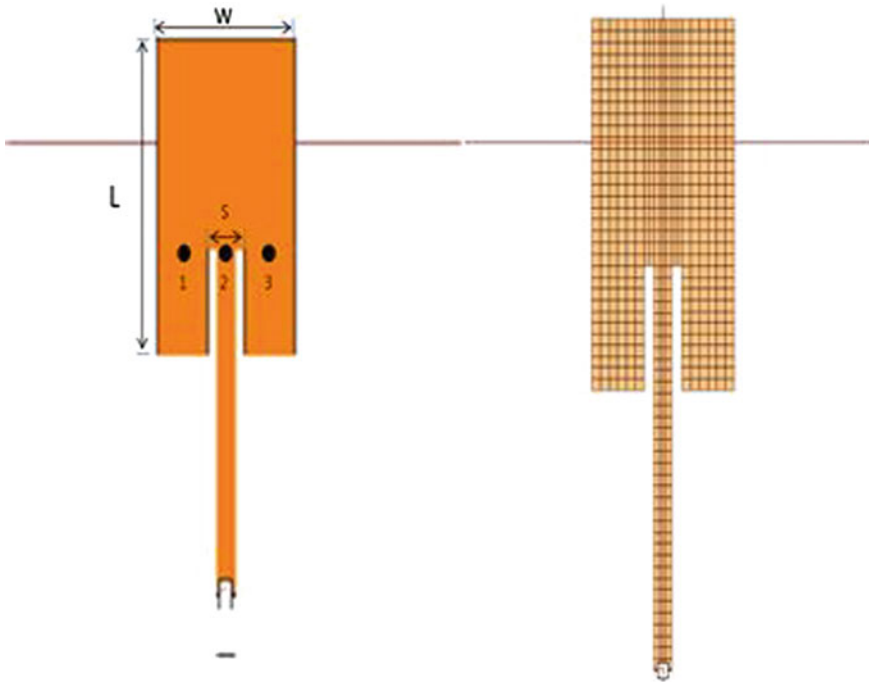


Fig. 2 IE3D mesh pattern of the patch antenna

CPW is higher). CPW of 20–30 is used in most simulations during antenna manufacturing to secure enough precision. The designed antenna uses multi-angular, polyhedral mesh creation method with which the remaining antenna structure connecting fixed points and microstrip feed lines is designed. As shown in Fig. 2 feed line to provide RF electricity to the patch during polyhedral mesh creation is designed in the antenna and port number to be used as the standard during when feed line is designed and scattering parameter is estimated has been designated.

Manipulation of three-dimensional current distribution offers relationship between co-polarization and cross-polarization. Precise antenna’s polarized light is offered based on the characteristics of polarized light in the propagation field through the patch antenna. Figure 3 shows three-dimensional structure of meshed rectangular patch antenna using microstrip feed line.

3 Simulation

Figure 4 shows results of return loss measured via circuit network analyzer. Frequency in the point where patch antenna’s VSWR is 3 ranges in between 1.614 and 3.437 GHz. 1.823 GHz in frequency range in which the antenna can be used

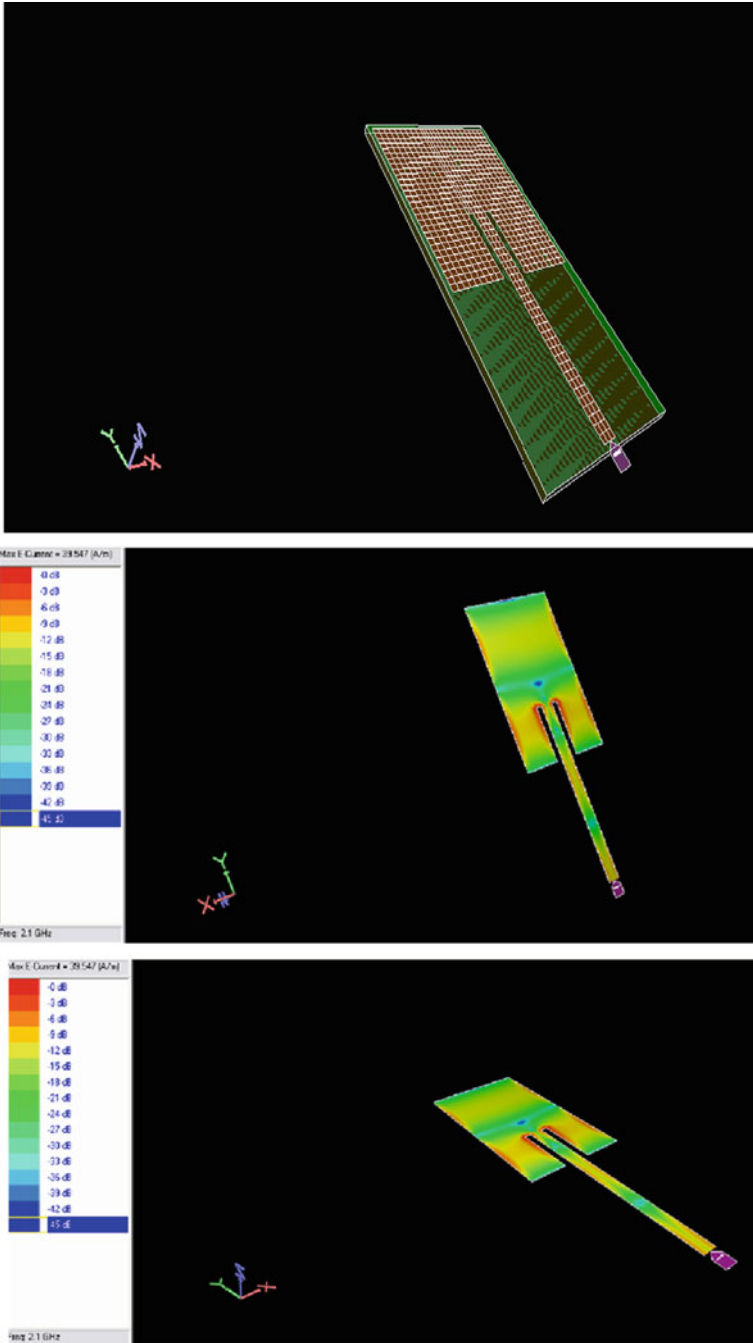


Fig. 3 3D structure of the meshed rectangular patch antenna using microstrip feed line

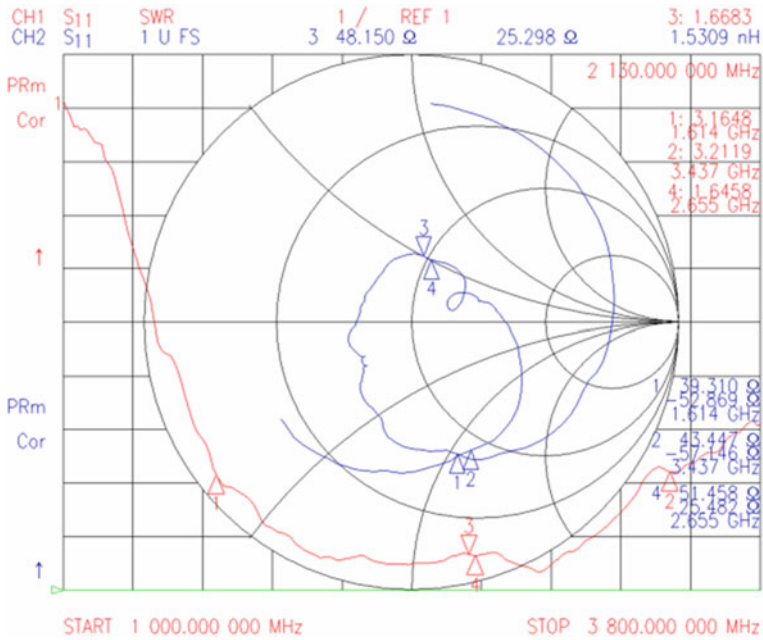


Fig. 4 S₁₁ measurement results of designed patch antenna

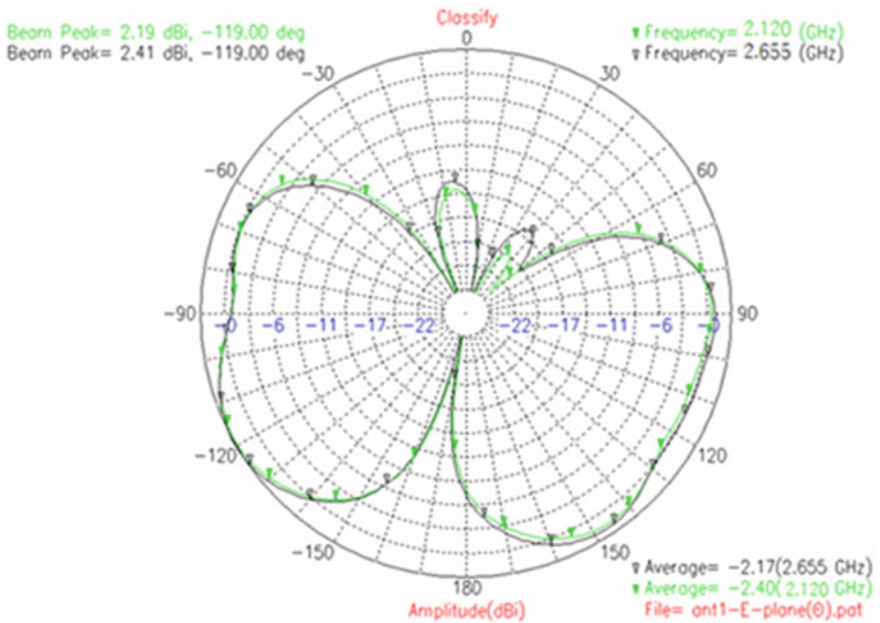


Fig. 5 E-field radiation pattern

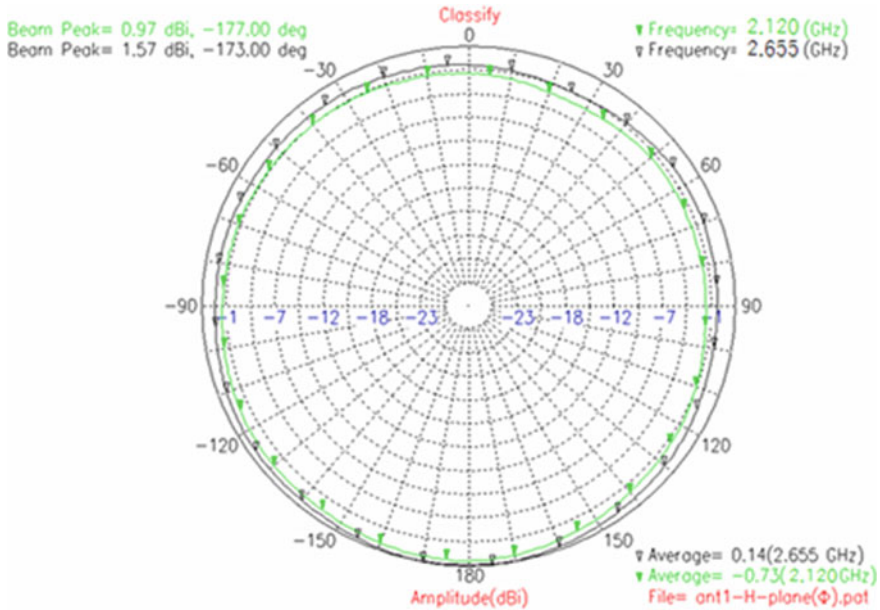


Fig. 6 H-field radiation pattern

shows broadband characteristics. Return loss in frequency band of 2.12 GHz is approximately -12.683 dB indicating healthy antenna characteristics in mobile communication band.

Figure 5 shows antenna’s E-domain radiation pattern in which antenna gain is 2.19 dBi in 2.120 GHz band and 2.41 dBi in 2.655 GHz band.

Figure 6 illustrates antenna’s H-domain radiation pattern in which antenna gain is 0.97 dBi in 2.120 GHz band and 1.57 dBi in 2.655 GHz band. This indicates that the antenna is more directional compared to isotropic antenna but reception is available in all directions, making it fit for mobile communication as confirmed in the figure. It can also be used as multi band antenna.

4 Conclusion

Resonance frequency of microstrip patch antenna for mobile communication was set at 2.4 GHz since it should be used in mobile communication system. Dielectric constant of panel used in this study was set at 2.45, height of dielectric substance panel at 1.58 mm for simulation. According to the simulation antenna’s optimal feed depth was 13.2 mm when it was -27 dB during when center frequency was from 1.9120 GHz to 23.28 MHz. VSWR was <3 in 2.2–3.5 GHz showing broadband characteristics and antenna’s maximum gain when resonance frequency was 1.87 dBi in 2.4 GHz indicating outstanding characteristics. Radiation pattern in

antenna's E-domain was 1.60 in 2.630 and 1.81 dBI in 2.655 GHz. In H-domain, radiation pattern was 1.24 in 2.630 and 1.81 dBI in 2.655 GHz, making it fit for mobile communication.

References

1. Finkenzeller K (2003) RFID handbook, 2nd edn. Wiley, England
2. Lier I, Jakobsen KR (1983) Rectangular microstrip patch antennas with infinite and finite ground-plane dimension. *IEEE Trans Antennas Propag AP-31(6)*:978–984
3. Mailoux RJ (1987) On the use of metallized cavities in printed slot arrays with dielectric substrates. *IEEE Trans Antennas Propag AP-35(5)*:477–487
4. Ikonen PMT, Rozanov KN, Osipov AV, Alitalo P, Tretyakov SA (2006) Magneto-dielectric substrate in antenna miniaturization: potential and limitations. *IEEE Trans Antennas Propag 54(7)*:3391–3399
5. Ruiming Z, Xin Z, Xi L, Qun P, Yinglong F, Dacheng Y (2009) Performance evaluation on the coexistence scenario of two 3GPP LTE system. *VTC 2009-Fall 3(2)*:213–221
6. Song M-H, Woo J-M (2003) Miniaturisation of microstrip patch antenna using perturbation of radiating slot. *IEEE Electr Lett 39(5)*:417–419
7. Seo J-S, Woo J-M (2004) Miniaturisation of microstrip antenna using irises. *IEEE electr Lett 40(12)*:718–719
8. Lier I, Jakobsen KR (1983) Rectangular microstrip patch antennas with infinite and finite ground-plane dimension. *IEEE Trans Antennas Propag AP-31(6)*:978–984
9. Fang ST A novel polarization diversity antenna for WLAN applications. In: *Proceeding of IEEE antennas and propagation society International Symposium Salt Lake City*, vol 2, no 1. pp 282–285
10. Shelokar PS (2007) Particle swarm and ant colony algorithms hybridized for improved continuous optimization. *Appl Math Comput 188(1)*:129–142
11. Kennedy J, Eberhart R (1995) Particle swarm optimization. *Proc IEEE Int Conf Neural Netw 4(2)*:1942–1948

Microstrip Patch Antenna on UHF Band Using Multiple Meander for Metal Attached in U-City

Chanhong Park

Abstract The tag antenna has been particularly influential in the performance of RFID system. Many applications require tag antennas to be of low profile mounted on electrically metallic objects. Several designs have been developed for RFID patch-type antennas or planar inverted-F antennas (PIFAs) mountable on metallic objects. Although most of these reported antennas can give the required reading-range performance, they may suffer from inconvenient mounting on metallic object because of their high profile. However, it is inherent for a patch-type antenna that lowering the antenna profile would degrade its radiation efficiency and antenna gain. Hence, an effort is being made to further improve the antenna gain of the low-profile patch antenna needed to provide the applicable reading range in a RFID application. In this paper, we designed meander-type microstrip patch antenna which displays the best performance at the frequency of 910 MHz, the RFID standard available in metal environment. Square-shaped power feeder, connected to the body for coordination with common-use tag chip attached to the antenna, is located in the body, while the patch device is designed in the form of multiple meanders to efficiently scale back the size of body of antenna. Then the characteristics of bandwidth, efficiency, and recognition distance are compared and analyzed by the size of proposed antenna and the number of being folded. It was found that the efficiency and gain characteristics changes by the size of antenna and the number of being folded in the form of meander have a significant influence over recognition distance of antenna.

Keywords RFID · UHF · Microstrip patch antenna · Multiple meander

C. Park (✉)

SamSun Technology Research Co. Ltd., Company-affiliated Research,
564-6, Sang-dong, Wonmi-gu, Bucheon-si, Gyeonggi-do, Korea
e-mail: iparka1028@gmail.com

1 Introduction

RFID system, which refers to a technology that enables wireless transmission of information on the exploitation of wireless interface based on radio frequency, is a state-of-the-art technology that brings innovative changes in human way of life and existing industrial structure instead of established barcode system. In addition, the rapid spread of applications of RFID system to the society at large, including service, sales, distribution, industry, manufacturer, and logistics system not only pushed it into the limelight as a core technology to generate massive economic implications, but also made it a critical technology for realization of ubiquitous society [1–3]. One of the biggest challenges in implementation of RFID tag antenna lies in development of tag form attached to conductive substance. As conductive substance such as aluminium can, metal box, and cigarette packet wrapped with foil increases directivity of antenna and wields great influence upon antenna performance such as resonant frequencies and radiant efficiency, it is hard to build up RFID system. Under these circumstances, implementation of a tag antenna without deterioration (chemical or physical degradation often caused by external or internal influence of insulation) has now become a supreme task [4–6]. Various studies have been made of antennas attached to metal, including spiral dipole and folded dipole with parasitic element. These antennas, however, tend to have a larger area, a half-wavelength, or a considerable height [7]. Even if PIFA antenna may have a smaller size, the antenna performance shall differ from the size of ground contact area, when it has a contact area smaller than 1λ . Parasitic capacitance between metal substance and tag antenna is likely to cause resonant frequencies, antenna impedance, and radiant efficiency, leading to deterioration of antenna performance [8]. In this respect, this paper designed and manufactured a metal tag antenna attached to metal substance that is likely to compensate the above defects. IE3D was employed to apply the precise values of parameters, while PSO [9, 10] algorithm was used to minimize the errors of parameters caused from producing antenna. Lastly, low input resistance, bandwidth, and antenna gain, which used to be shortcomings of existing antenna, are substantially improved. The rest of this paper is organized as follows. Section 2 describes designed the microstrip patch on UHF band using Multiple meander and IE3D. Section 3 describes the simulation result of antenna. The conclusions are given in Sect. 4.

2 UHF Microstrip Patch Antenna Attached to Metal

2.1 Simulation Environment

This paper designed meander-type microstrip patch antenna to acquire broadband matching characteristics, various radiation patterns, broad VSWR and ratio

bandwidth, while minimizing the size of tag in the metal environment. Square-shaped power feeder is connected to the body for coordination with common-use tag chip attached to the antenna, while the patch device is designed in the form of multiple meanders to efficiently scale back the size of body of antenna. When it comes to designing an antenna, parameters for resonant frequencies of antenna, dielectrical constant of board, and height of dielectrical board are required. Table 1 shows the parameters required for designing an antenna. Resonant frequency is set as 910 MHz, the default value of antenna, to be available between 860 and 960 MHz, the standard UHF bandwidth compatible with RFID system. Aluminum oxide, which was employed in dielectric substance used for antenna design, has a dielectric constant of 4.4. The height of dielectric board, the thickness of antenna, was set as 0.6 mm.

The tag chip used for antenna design is Alien Higgs which has capacitive input impedance of $13 - j111 \Omega$ at 915 MHz, and is connected with square-shaped power feeder at its end. As input reactance of tag chip has a capacitive value, the input impedance of tag antenna has a low resistance similar to that of tag chip for tag antenna to have broadband characteristics, whilst the input reactance has inductive characteristics for desirable impedance conjugate matching. Figure 1 shows the structure of designed antenna. The width of ground is fixed to 100 mm, while the breadth increased by 10 mm by the size of patch device. Every patch device takes the form of non-uniform linear section. Width 't' and length 'l' are adjusted for perfect impedance matching between antenna and tag chip.

The width of folded patch, which connects devices, is devised to have a wavelength of 0.5 at 915 MHz. As the path of each device is 0.5 wavelength with 180° phase shift, the current flow of all devices has the same direction with an increase in the antenna gain, where $\theta = 0$. Tag chip is located at the end of patch device, taking the end of patch as feeding point.

Figure 2 shows input impedance by the changes of width 't'. Two of the variables that can be adjusted to gain desired impedance include width 't' and length 'l'. Where $t = 4$ mm, reactance curve shall not exceed 111Ω , which implies that it's impossible to acquire desired matching state, when $t = 4$ mm. When width 't' is reduced, the input reactance increases. When width 't' is scaled back to 0.2 mm, it may acquire an input impedance of $14 + j111 \Omega$ with the frequency of 895 MHz. Thus, width 't' and t_a are set as 3 mm.

Figure 3 shows the influence of changes in length over input impedance, when the width of patch device 't' = 3 mm. When length 'l' reduces, impedance curve hardly changes but moves to higher frequency. Meanwhile, it was found that there

Table 1 Simulation parameter

Parameter	Value
Resonance frequency (f_0)	915 MHz
Dielectric constant (ϵ_0)	4.4
Height of dielectric substance board (h)	0.6 mm

Fig. 1 Structure of designed antenna

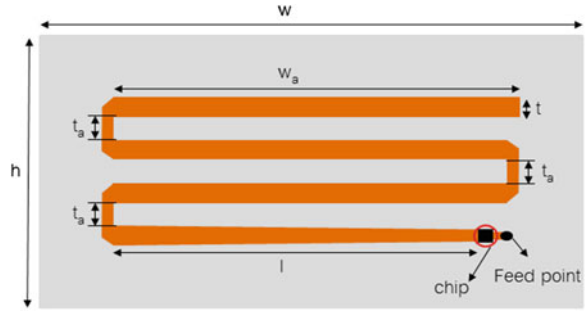


Fig. 2 Impedance of the patch element by change in width (t). **a** Input resistance. **b** Input reactance

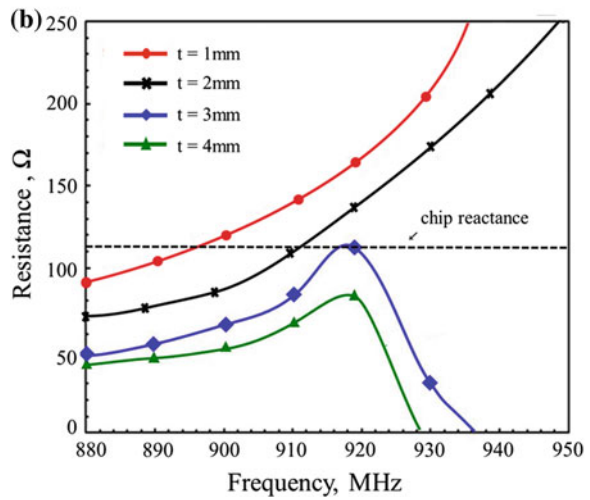
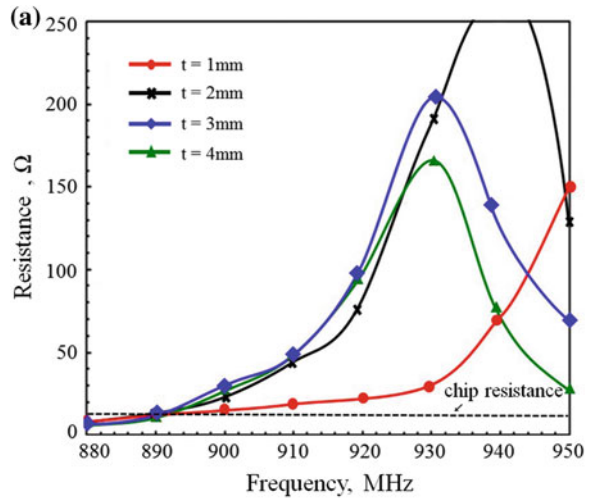
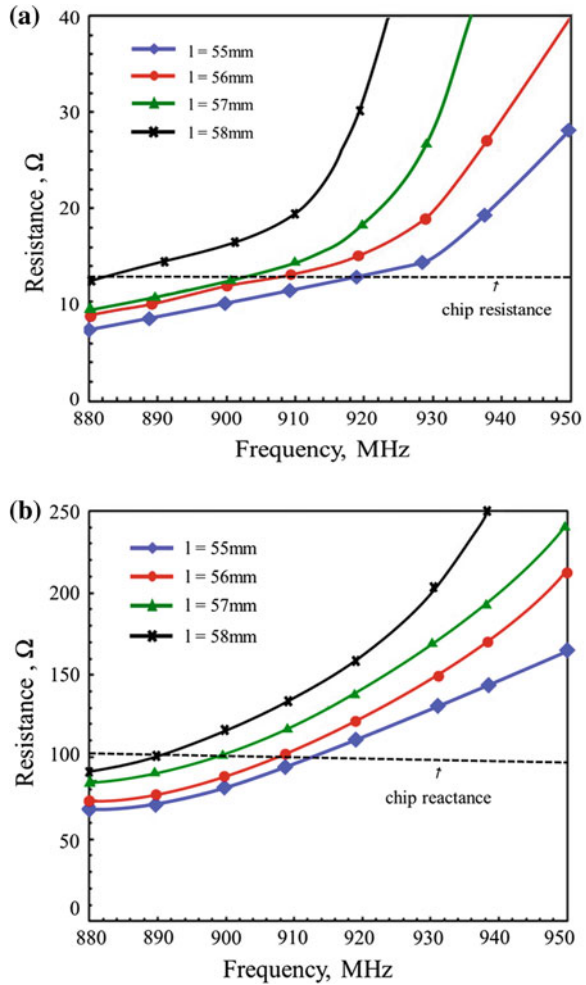


Fig. 3 Impedance of the patch element by change in length (l). **a** Input resistance. **b** Input reactance



had been conjugate impedance matching between antenna and chip, with $13 + j112 \Omega$ input impedance of antenna at 915 MHz, where length ' l ' = 56 mm.

Table 2 shows the design variables of designed antennas. The aforementioned values are applied to each antenna to compare the performance of designed antennas, with slight changes in ground height.

3 Simulation

The tag chip used for antenna design is Alien Higgs made of strap package which has an impedance value of $13 - j111 \Omega$ at 915 MHz. The input impedance of antenna has to be $13 + j111 \Omega$ at 915 MHz to convey the maximum power

Table 2 Parameter of antenna

Parameters	Value (mm)
w	80
h	40
wa	71
t	3
l	56

between chip and antenna. The bandwidth of designed antenna is found to be 909–929 MHz in free space and 908–928 MHz in metal section, which suggests that all antennas are well matched with chip impedance, satisfying the international RFID standard. Figure 4 shows the reflection loss by diagonal slot and air space, which implies that the reflection loss of microstrip patch antenna with diagonal slot and air space is the lowest at 915 MHz, and that the bandwidth is 26 MHz when VSWR is <1.2 under the reflection loss of -19.59 dB.

Figure 5 shows the 2D radiation pattern of designed antenna in free space. The result of simulation shows 2.6 dBi and 70 % of efficiency in the air with 3 dB angular width of 81.6°.

Figure 6 shows the 2D radiation pattern of designed antenna in radiation section. The result of simulation shows 4.21 dBi and 70 % of efficiency in the air with 3 dB angular width of 92°.

The designed antenna is measured by common-use RFID reader(XR440) with operating frequency of 902–928 MHz, output of 30.0 dBm, and circular polarized antenna gain of 6.0 dBi. It show the performance of antennas measured in a free

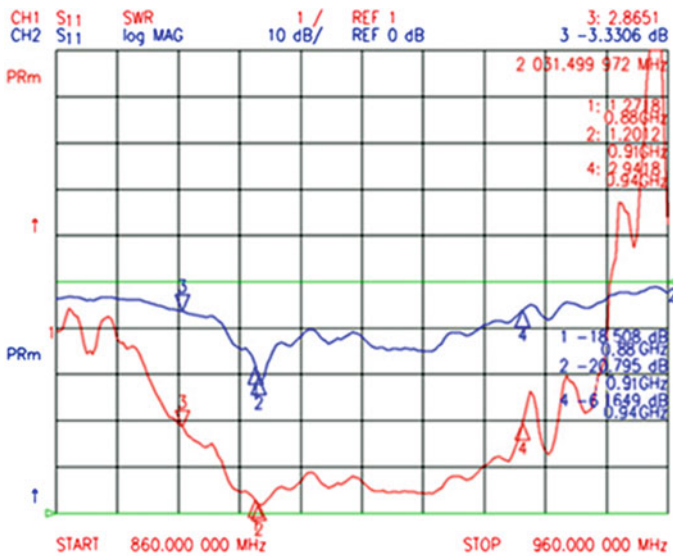


Fig. 4 Return loss measurement results of designed antenna

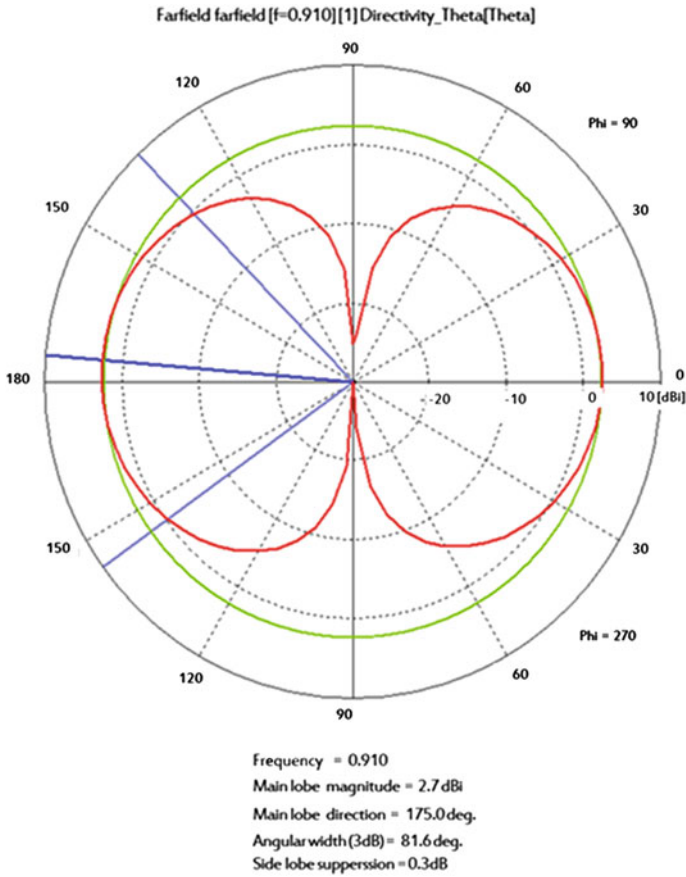


Fig. 5 Radiation pattern of free space

space and on $400 \times 400 \text{ mm}^2$ metal plate, which demonstrates maximum induced antenna gain of all antennas at around 910 MHz. It was found that the antenna gain decreased by the decrease in antenna profile by the result of induction. The recognition distance of antenna is calculated by frequency through Friis formula (2) posterior to measuring minimum radiation power that drives the tag chip by the direction of antenna [11].

$$\text{Readable range} = \frac{\lambda}{4\pi} \sqrt{\frac{P_{reader} G_{reader} G_{tag}}{P_{tag}}} \tag{1}$$

- P_{reader} = Output port power of reader antenna
- G_{reader} = Gain of reader antenna,
- P_{tag} = Minimum threshold power conveyed to tag chip,
- G_{tag} = Gain of tag antenna,

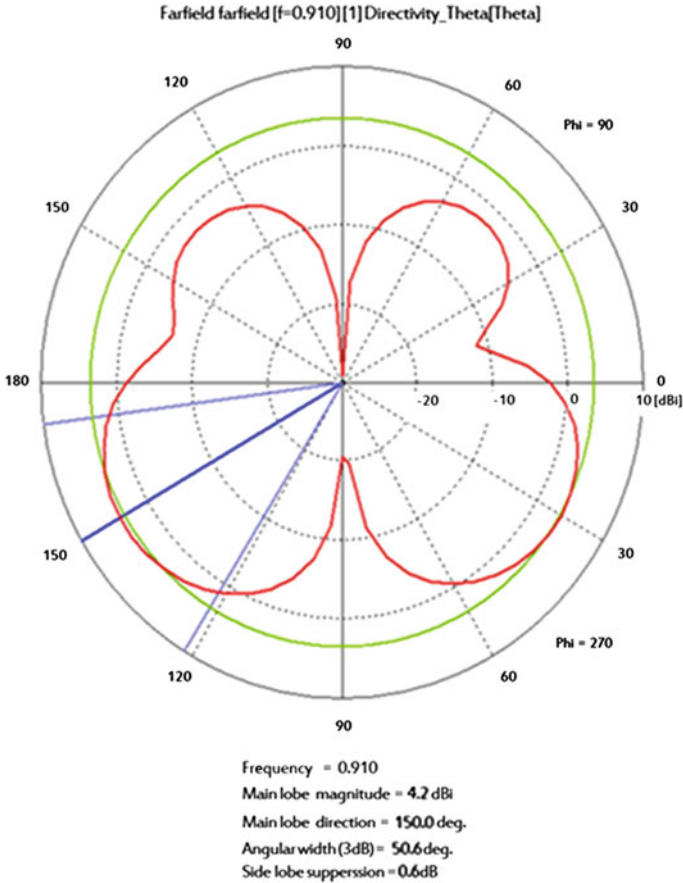


Fig. 6 Radiation pattern of metal space

$$\lambda/4\pi = \text{Loss coefficient in 1 m free space}$$

P_{tag} , the minimum threshold power of tag, is calculated by adding EIRP by P_{reader} and loss coefficient in 1 m free space. EIRP is the multiplication of P_{reader} , the input power of reader antenna, and G_{reader} , the gain of reader antenna, and generally has maximum 4 W or <36 dBm, according to ISO 18000. Accordingly, EIRP lower than 36 dBm is computed by multiplying G_{reader} , the gain of 3 dBi, to minimum output power of reader for operation of tag, and is later used to calculate tag sensitivity, that is to say P_{tag} , the minimum threshold power of tag, in consideration of loss coefficient in 1 m free space, G_{reader} , the gain of reader antenna, and cable loss. In addition, P_{tag} is divided by multiplication of EIRP and loss coefficient in 1 m free space to indicate G_{reader} , the gain of tag antenna [12]. The recognition distance measured by frequency shall be maximum 5.8 m at 908 MHz,

Table 3 Reading ranges of designed antenna

MHz	Recognition distance (m)	
	Free space	Metal plate
908	9.9	5.8
915	10.4	6.1
920	11.5	7.3

the minimum frequency range of antenna; 6.1 m at 915 MHz; and 7.3 m at 920 MHz (Table 3).

4 Conclusion

This paper designed meander-type microstrip patch antenna to acquire broadband matching characteristics, various radiation patterns, broad VSWR and ratio bandwidth, while minimizing the size of tag in the metal environment. Square-shaped power feeder is connected to the body for coordination with common-use tag chip attached to the antenna, while the patch device is designed in the form of multiple meanders to efficiently scale back the size of body of antenna. The design variables of antenna are the same, as the designed antenna is devised to find the optimal antenna by comparing the performance by the increase of patch device in the identical environment. The bandwidth of designed antenna is found to be 908–920 MHz, which suggests that all antennas are well matched with chip impedance, satisfying the international RFID standard. Antenna gain is found to be 4.2 dBi, which has no significant difference with the metal antenna in use. However, the recognition distance of existing antenna is maximum 4 m in the metal section, while that of the designed antenna in this study ranges from minimum 5.8 m to maximum 7.3 m, which shows a considerable improvement in recognition distance.

References

1. Finkenzeller K (2003) RFID handbook, 2nd edn. Wiley, England
2. Lier I, Jakobsen KR (1983) Rectangular microstrip patch antennas with infinite and finite ground-plane dimension. *IEEE Trans Antennas Propag AP-31(6):978–984*
3. Balanis CA (1997) Antenna theory analysis and design. Wiley, New York
4. Ukkonen L, Sydanheimo L, Kivikoski M (2004) Anovel tag design using inverted-F antenna for radio frequency identification of metallic objects. *IEEE AW&WC 41(3):91–94*
5. Raunonen P, Sydanheimo L, Ukkonen L, Kivikoski M (2003) Folded dipole antenna near metal plate. *IEEE AP-S 2(3):848–851*
6. Ukkonen L, Engels D (2004) Planar wire-type inverted F RFID tag antenna mountable on metallic objects. *IEEE AP-S 1(1):101–104*

7. Ukkonen L, Sydanheimo L, Kivikoski M (2005) Effects of metallic plate size on the performance of microstrip patch-type tag antennas for passive RFID. *IEEE Antennas Wirel Propag Lett* 4(1):410–413
8. Huynh M, Stutzman W (2003) Ground plane effects on planar inverted-F antenna performance. *IEEE Proceedings microwaves, antenna and propagation*, vol 150, pp 209–213
9. Shelokar PS (2007) Particle swarm and ant colony algorithms hybridized for improved continuous optimization. *Appl Math Comput* 188(1):129–142
10. Kennedy J, Eberhart R (1995) Particle swarm optimization. *Proc IEEE Int Conf Neural Netw* 4(1):1942–1948
11. Friis HT (1946) A note on a simple transmission formula. *Proceeding of IRE*, pp 254–256

Multi-Hop Relay Protocol Techniques for Performance Enhancement of LTE Downlink System

Chanhong Park

Abstract In this paper, we proposed research to boost reception performance in link-down transmission method of LTE system, which is the next-generation mobile communication technology standard underway in 3GPP. At the moment, in LTE downlink system, OFDM which is suitable for high speed data transmission and multipath has been used. However, OFDM method has a weakness displaying the relatively greater PAPR at the terminal because it basically uses multi-carrier. To this end, SC-FDMA has been used in LTE uplink system in order to compensate this big defect related to a great PAPR of OFDM in such an important terminal where power efficiency really matters. However, when signals are deteriorated by the channels in the frequency domain, SC-FDMA reveals a defect in that the impact of deteriorated parts is spreading and causes performance degradation. To this end, it proposed installing relay (RS) in between station(BS) and terminal(MS), set the distance between BS and RS at 500 and 1,000 m, each, and chose OFDMA and SC-FDMA as transmission method of RS. The paper found SC-FDMA to be better choice in RS when it is closer to BS and OFDMA to be a better choice in RS when the distance between BS and RS is farther. The system's reception performance improved when the most appropriate transmission method fitting the circumstances was used in the middle between BS and MS.

Keywords 4G · LTE · OFDMA · SC-FDMA · Multi-hop relay

C. Park (✉)

SamSun Technology Research Co. Ltd, Company-affiliated Research,
564-6, Sang-dong, Wonmi-gu, Bucheon-si, Gyeonggi-do, Korea
e-mail: iparka1028@gmail.com

1 Introduction

In the wake of the high demand on wireless multimedia services and with the necessity of high-capacity, high-speed data transmission, development of a radio-access-based 3G Mobile Communication Service is on the steady rise, improving its data transfer speed, voice-data integration, and ATM-based networking trouble inevitably taken [1]. Having improving the relevant technologies of HSDPA, HSUPA, and MBMS since WCDMA standardization, 3GPP in the long run has been providing highly competitive wireless accessing technology by way of HSDPA/HSUPA. In need of the advanced mobile technology, however, 3GPP is now discussing 4G Mobile Communication Service in depth, a much improved communication means in comparison to ‘Release 6 Technology’, in terms of latency, data transfer rate, system capacity, and coverage. Among the variation of 3GPP LTE, Wibro Evolution, and 3GPP2 UMB, 4G Mobile Communication Service is projected to be standardized into LTE-Advanced, an OFDM-based LTE communication means deemed most suitable to be called as ‘4G’ [2]. Projecting the data transfer rate of 155Mbps–1Gbps or 100Mbps while abeyant or in motion, respectively, 4G Mobile Communication Service is aiming integration of the wired and wireless communication means in a higher capacity [3]. Adopting OFDMA SC-FDMA for uplink and downlink, LTE-Advance supports bi-directional multiple antenna MIMO, as well as offering Hybrid OFDMA/SC-FDMA in compensation of PARR efficiency and performance differential between OFDMA and SC-FDMA. Multi-hop Relay is another alternative to advance data capacity and coverage [4, 5]. As for OFDMA, a multi-access means incorporating sub-channels in part, this capacitates multi-assignment by way of OFDM involving use of the multiple frequency scopes bearing orthogonality each other for better endurance to frequency fading and featuring efficiency in operation. Incorporated for downlink is OFMDA, improving frequency efficiency and cell capacity but with some setback due mostly to the higher PARR. Uplink incorporates SC-FDMA to compensate thereof, using a single carrier wave to lower PARR but likely contorting the signal from frequency scope channel as using an assimilator. Spreading the contorted data, SC-FDMA loses PARR efficiency with higher modulation level [6, 7].

Therefore, in this paper, we set up the distance between base station and terminal as 500 m, 1000 m in order to improve the performance difference between OFDMA and SC-FDMA, and to increase the performance of the system and its coverage, then a relay was installed between base station and terminal, and, on the base of the base station and the relay already installed, the we conduct a test for select proper way of the transfer mode of OFDMA and SC-FDMA [8]. The rest of this paper is organized as follows. [Section 2](#) describes the Hybrid OFDMA/SC-FDMA system. [Section 3](#) describes data transmission protocol, frame structure in relation to Multi-hop Relay and as the means enhance LTE performance. [Section 3](#) describes the relay location, transmission set-up, and Relaying performance by simulation. The conclusions are given in [Sect. 4](#).

2 Hybrid OFMDA/SC-FDMA

2.1 OFDMA

Semi-carrier waves are grouped into sub-channels, which are a larger unit, and these sub-channels are grouped into bursts to be allotted to wireless users. Each burst allotment can be changed in every frame within the order of modulation and this enables dynamic adjustment of bandwidth use according to what the current system in a station requires. Furthermore, power consumption by each user can also be adjusted according to the current system’s requirement since each user consumes only part of the whole bandwidth [9, 10] (Fig. 1).

2.2 SC-FDMA

SC-FDMA refers to the single carrier-frequency division multiple access mode. This is a multi-user modulation based on single carrier-frequency domain equalization. SC-FDE can be seen as a linearly pre-coded 7OFDM mode, whereas SC-FDMA can be seen as a linearly pre-coded OFDMA or LP-OFDMA mode. As in the case of OFDM mode, SC-FDMA is inserted between blocks of symbol for circulating protection period caused by multipath propagation to efficiently eliminate extension of time between blocks. In OFDM mode, FFT is applied to receiver of each block, while IFFT is applied to transmitter. In SC-FDE mode, FFT and IFFT alike are applied to receiver, not to transmitter. In SC-FDMA mode, FFT and IFFT are applied to both receiver and transmitter. Equalization is made by multiplying Fourier coefficients with complex number after calculating FFT not only in SC-FDE and SC-FDMA but also in OFDM. Then frequency-selective attenuation and phase distortion can be prevented. The good thing is that less computing power is needed in FFT and frequency domain equalization than in the existing time domain equalization. In SC-FDMA mode, multiple access is available by inserting silent Fourier coefficients to transmitter prior to the implementation of IFFT and eliminating them

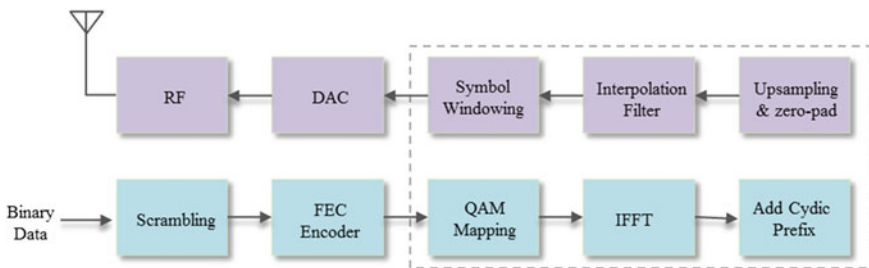


Fig. 1 Transmitters/receivers block diagram of OFDMA downlink system

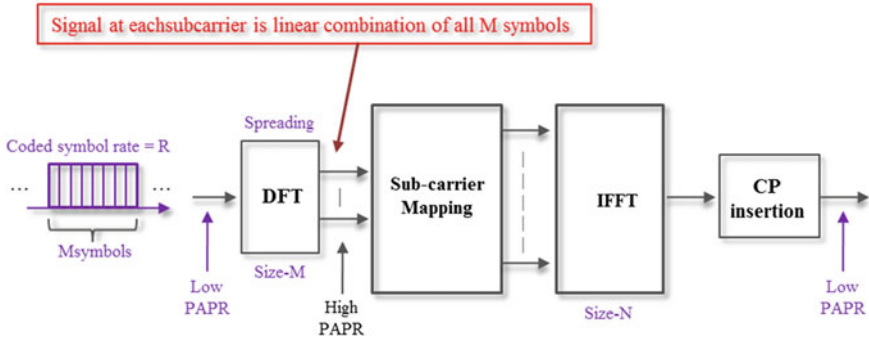


Fig. 2 Structures of SC-FDMA transmitter unit

at the receiver before implementation. Other users are assigned to other Fourier coefficients (subcarrier).

SC-FDMA, which stands for Single Carrier Frequency Division Multiple Access, can stave off frequency selection attenuation and phase distortion as both FFT and IFFT are applied to both transmitter and receiver [11, 12] (Fig. 2).

3 LTE Reception Performance Improvement Test Multi-Hop Relay Technology

3.1 Existing LTE Transmission Protocol

While existing LTE protocols have the merit of enhancing channel capacity and reducing transmission power through the transmission from base station (BS) to mobile station (MS) via relay station (RS), it is hard to expect improved reception performance from them. In this respect, diversity system, which BS sends the same signal to RS and MS and MS combines the signals from BS and RS, is applicable. Gain from transmission rate is to be taken into consideration on the application of spatial multiplexing technique different data signals are transmitted at the same time. Where is RS, one of relay source of signals, existing protocols shall be as follows in three cases [13] (Table 1).

3.2 Process of Associative Crayon Colors Pattern

This paper proposed making a selection between SC-FDMA or OFDMA depending on each terminal environment in RS in order to bring about optimal performance under LTE environment. Relay installation to optimize performance

Table 1 Relay protocol

	Protocol 1	Protocol 2	Protocol 3
Time slot 1	BS → RS, BS → MS	BS → RS, BS → MS	BS → RS
Time slot 2	BS → MS, RS → MS	RS → MS	BS → MS, RS → MS

Table 2 The proposed relay protocols

	Protocol	Existing LTE (HSDPA)	Proposal		
			Protocol 1	Protocol 2	Protocol 3
Time slot 1 (phase 1)	BS → Relay	OFDM	OFDMA	SC-FDMA	SC-FDMA
Time slot 2 (phase 2)	Relay → MS	OFDM	SC-FDMA	SC-FDMA	OFDMA

in shadow domains and the subsequent protocol needed is also proposed in the paper. In cases where transmitter unit and receiver unit is in short distance, OFDMA is used to ensure frequency efficiency and boost cell capacity. In cases of long-distance, SC-FDMA employed in LTE uplink is used to make up for OFDMA's weakness and boost system's performance (Table 2).

In the proposed transmission protocol, Time slot 1 used BS → RS, applied in the existing Protocol 3, while Time slot 2 used RS → MS, applied in the Protocol 2, in order to minimize path loss or power dissipation and guarantee optimal communication without having to experience waste of resource at BS, RS, MS and pass through multiple relays. Accordingly, a relay station in which BS is exploited as relay is chosen at Time slot 1, while data is transmitted from relay to MS at Time slot 2. Furthermore, LTE downlink is set as the default along with the selective use of OFDMA and SC-FDMA according to BS and relay conditions. In the event that transmitter and receiver are near to one another, OFDMA is recommended to amplify frequency efficiency and cell capacity. However, when they are distant from one another, SC-FDMA used in LTE uplink is recommended to remedy defects of OFDMA and maximize the system performance.

3.3 Frame Structure

Frame structure of the proposed mode is set forth in Fig. 3, BS-RS and RS-MS links use different carrier frequencies, and correct synchronization is necessary between BS and RS. RS is presumed to implement BS mode and MS mode alternately in a time domain. BS mode of RS refers to the cases of which signals received from BS are resent to MS or RS and of which uplink signals are received from either MS or RS. MS mode refers to the cases of which signals are received from BS and of which uplink signals received from MS or RS are resent to BS.

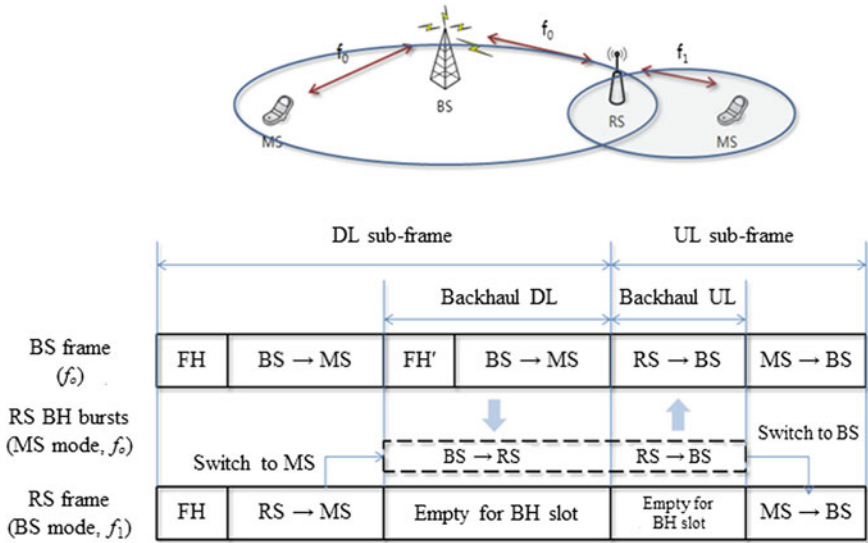


Fig. 3 Frame structures using RS

Frame header refers to control signals for MS transmitted without relay base and is comprised of Preamble, FCH, DL/UL-MAP, D/UCD, etc. FH refers to the same frame header information sent to all RS and sends rather simplified information than FH. In addition, back-hole here refers to communication between BS and RS [14, 15].

In the back-hole section, when RS receives or transmits signals from or to BS, the frame switches to the MS mode. When RS transmits signals from BS to MS in the service zone or receives signals from MS, it switches to the BS mode. In other words, the up/down-link signals transmitted to each RS take different positions in the time domain [16].

3.4 Performance Gap Between OFDMA and SC-FDMA

The switching distance that generates performance gap between OFDM and SC-FDMA can be simulated. Therefore, simulation is the basic element of the method, as suggested in this paper, to select the transmission method according to the conditions of and distance between BS and MS. Because the performance gap between OFDM and SC-FDMA, caused due to a difference of power efficiency can be a critical element, it is required to consider the back-off value in transmission of the OFDM signal.

$$back - off = (A^2)/(MAX_input) \tag{1}$$

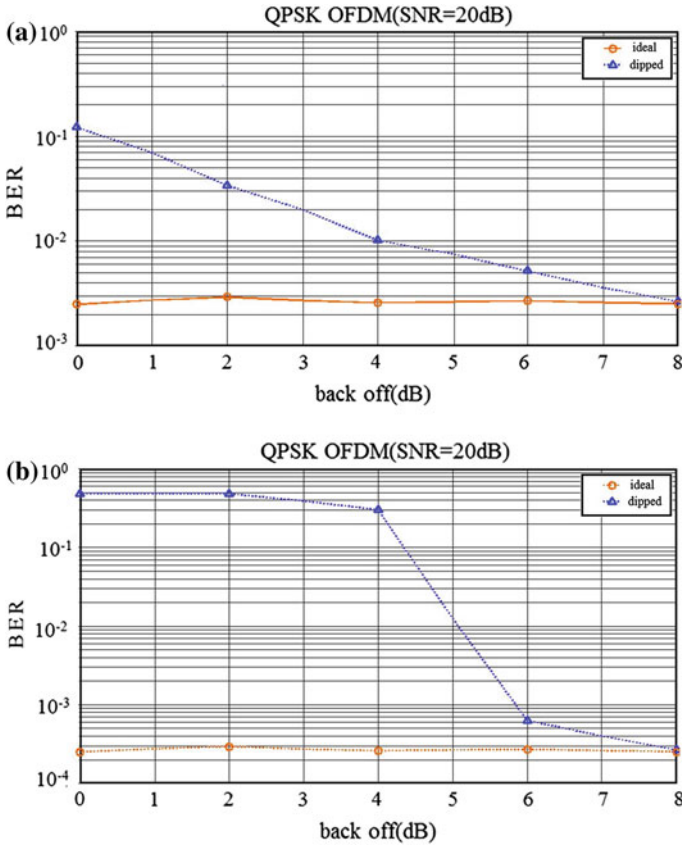


Fig. 4 OFDM performance comparison of back-off values **a** QPSK modulation, **b** 16QAM modulation

Equation (1) represents the back-off value. A back-off is defined as the clipping power divided by the maximum transmission signal power. This formula provides you the size of clipping of the signal ($\pm A$), and by applying it to the simulation, you can compare the performance to which the back-off is applied with the performance of the ideal case. Figure 4 shows the ideal BER of OFDM by using a fixed SNR value. The simulation result shows maintain the ideal or similar OFDM performance at the back-off value of approx. 6–8 dB.

Figure 5 shows the comparison of BLER performance between OFDM and SC-FDMA by distance. The performance was simulated with increasing distance of MS from BS. The simulation result shows that SC-FDMA provides better performance than OFDM after 210–200 m when the back-off value is 6 dB, and the result is inverted at 160–170 m when the back-off value is 8 dB. This result indicates that, if RS is used between BS and MS, performance may vary depending on the selected

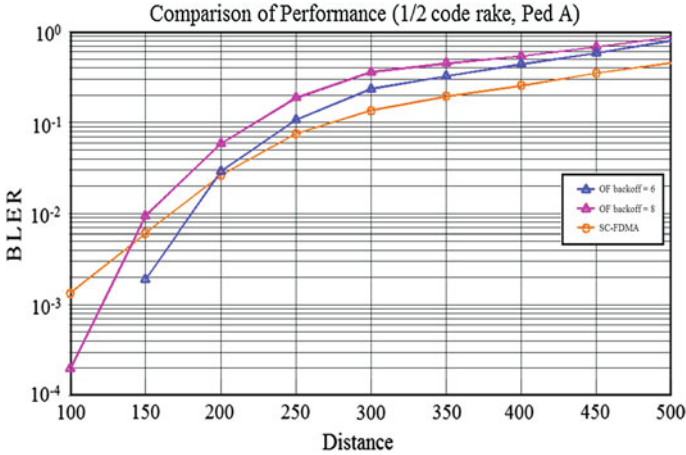


Fig. 5 Performance comparison of OFDM and SC-FDMA by distance

transmission method by locations of MS and BS, though there may be a minor difference depending on the back-off value of OFDM.

4 Simulation

In the event that MS is 500 m away from BS, reception performance of selected transmission mode is displayed according to the location of RS. In this case, provided that the central region covers 250 m, it is found that BER of OF-SC (OFDMA-SCFMA) is better than that of OF-OF (OFDMA-OFDMA) when RS and RS are near to one another or RS and MS are about 100–200 m away from one another. In spite of OFDMA mode in BS, a higher level of performance is presented with SC-FDMA mode in RS, because RS and MS are rather distant. In the central region with stable power efficiency, the performance of OF-OF is better than that of OF-SC. In the meantime, it was found that the performance got better in OFDMA than in SC-FDMA, when the distance between RS and MS is closer. It was also indicated that OF-OF presented great performance from the central region, where the power of BS and RS is stable, to the 300–400 m region, where BS is far from RS and RS is near to MS (Fig. 6).

Figure 7 is the reception performance of transfer modes selected based on RS’s location assuming distance between BS and MS at 1,000 m. The simulation fixed OFDMA as the transfer mode from BS and RS and tested both OF-OF and OF-SC by switching transfer mode from RS to MS to OFDMA and SC-FDMA. OF-SC performed better than OF-OF when RS was located in 100–300 m but both modes hardly differed when BER of both modes were compared from 350 to 500 m in the center.

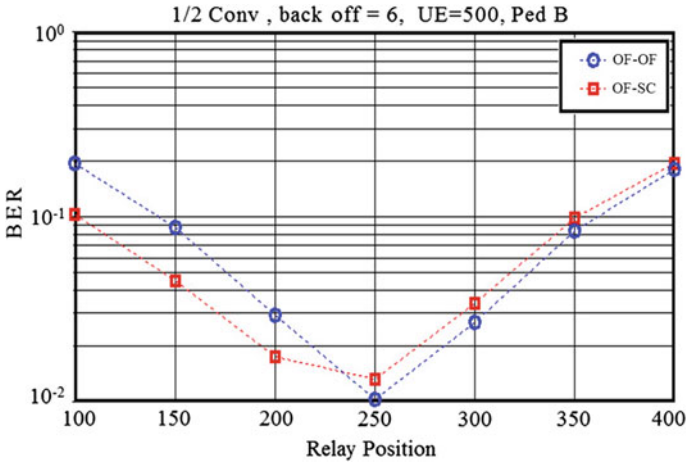


Fig. 6 BER in MS location of 500 m and OFDMA transfer mode between BS and MS

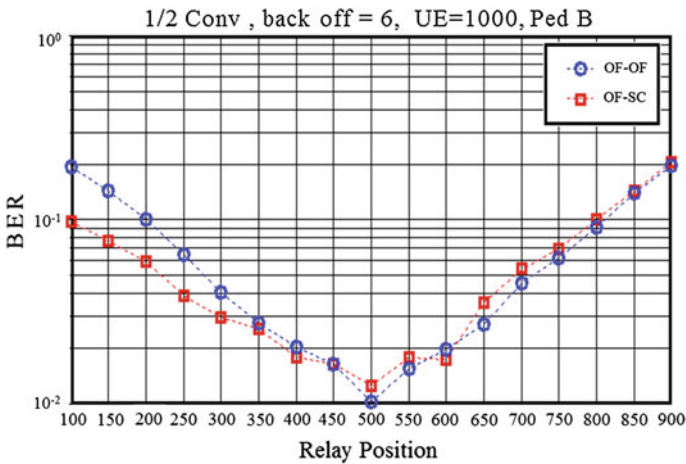


Fig. 7 BER in MS location of 1000 m and transfer mode between RS and MS

5 Conclusions

In this paper, we set up the distance between base station and terminal as 500 m, 100 m, in order to improve the performance difference between OFDMA and SC-FDMA, and to increase the performance of the system and its coverage, then a relay was installed between base station and terminal, and, on the base of the base station and the relay already installed, the we conduct a test for select proper way of the transfer mode of OFDMA and SC-FDMA. Also this paper proposed a combination of two transfer modes to fill up the performance gap between

OFDMA and SC-FDMA and to improve reception performance of LTE system's downlink transfer mode. The paper also proposed setting up RS in between BS and MS to enhance system's performance and coverage, and ran a simulation of the proposed idea. A simulation to appropriately select one out of OFDMA and SC-FDMA based on RS set with BS was carried out by setting the location between BS and MS at 500 and 1000 m, respectively, and setting RS in between BS and MS. Simulation revealed that OFDMA is a better option in BS and SC-FDMA in RS when RS was more closely located to BS. This was just the opposite as in the longer distance between BS and RS where SC-FDMA performed better in BS and OFDMA in RS. In the center between BS and MS, an improvement in the system's reception performance was foreseeable by selecting transfer mode befitting particular situation. It is expected that the results from this paper would be able to provide for desirable communication services by improving the receiving performance of the whole system, while reducing isolated area that is being appeared in LTE downlink system, as well as enlarging cell coverage of base station simultaneously.

References

1. 3rd Generation Partnership Project (3GPP) (2009) Evolved universal terrestrial radio access (E-UTRA); physical channels and modulation. <http://www.3gpp.org/ftp/Specs/html-info/36211.html>
2. Holma H (2009) LTE FOR UMTS—OFDMA and SC-FDMA based radio access. John Wiley & Sons Ltd, Chichester, pp 56–72
3. Zhang J, Huang C, Liu G, Zhang P (2006) Comparison of the Link Level Performance between OFDMA and SC-FDMA. In: Proceedings of the IEEE CNF, 25 Oct 2006
4. Myung HG, Lim J, Goodman DJ (2006) Single carrier FDMA for uplink wireless transmission. *IEEE Veh Technol Mag* 1(3):30–38
5. Pabst R et al (2004) Relay-based deployment concepts for wireless and mobile broadband radio. *IEEE Commun Mag* 42(9):88–89
6. Chuang J, Cimini LJ, Li G, Lin L, McNair B, Sollenberger NR, Suzuki M, Zhao H (1999) Highspeed wireless data access based on combining EDFE with wideband OFDM. *IEEE Commun Mag* 37:92–98
7. Genc V, Murphy S, Yu Y, Murphy J (2008) IEEE 802.16j Relay-based wireless access networks: an overview. *IEEE Wirel Commun* 15(5):56–63
8. Myung HG, Lim J, Goodman DJ (2006) Single carrier FDMA for uplink wireless transmission. *IEEE Veh Technol Mag* 1(3):231–239
9. Lightfoot L, Zhang L, Ren J, Li T (2009) Secure collision-free frequency hopping for OFDMA-based wireless networks. *EURASIP J Adv Signal Process* 2009:11–15 361063
10. Ghosh A, Wolter DR, Andrews JG, Chen R (2005) Broadband wireless access with WiMax/802.16: current performance benchmarks and future potential. *IEEE Commun Mag* 43(2):129–136
11. Zhang C, Wang Z, Yang Z, Wang J, Song J (2010) Frequency domain decision feedback equalization for uplink SC-FDMA. *IEEE Trans Broadcast* 56(2):253–257
12. Berardinelli G, Priyanto BE, Sorensen TB, Mogensen P (2008) Improving SC-FDMA performance by turbo equalization in UTRA LTE uplink. *IEEE Proc VTC* 54(4):2557–2561

13. Cover T, Gamal AE (1979) Capacity theorems for the relay channel. *IEEE Trans Inf Theory* IT-25(5):572–584
14. Holma H, Toskala A (eds) (2006) *HSDPA/HSUPA for UMTS-high speed radio access for mobile communications*. JohnWiley & Sons Ltd, Chichester, pp 125–157
15. Nagata S, Ofuji Y, Higuchi K, Sawahashi M (2006) Optimum resource block bandwidth for frequency domain channel-dependent scheduling in evolved UTRA downlink OFDM radio access. In: *Proceedings of IEEE vehicular technology conference (VTC)*, vol 1 no 1. Melbourne, pp 206–210
16. Goschini GJ, Gans MJ (1998) On limit of wireless communication in a fading environment when using multiple antenna. *Wireless Pers Commun* 6(3):331–335

Performance Improvement Using Single Carrier-FDMA in Relay Based LTE Uplink System

Chanhong Park

Abstract In this paper, we proposed to install a relay between base station and terminal (or user equipment) in order to improve receiving performance of the downlink transfer mode of the Long Term Evolution (LTE) system, which is the next-generation mobile communication technology standard that has been proceeding in the 3rd General Partnership Project (3GPP). In 3GPP LTE-advanced, hybrid OFDMA/SC-FDMA is recommended for its technological capability to make up for the performance gap between OFDMA and SC-FDMA and to make PAPR more efficient. OFDMA is used in LTE downlink in order to make frequency more efficient and raise cell capacity but OFDM struggles with a high PAPR owing to its use of multi carrier wave. LTE uplink, on the other hand, employs SC-FDMA, which is similar to OFDMA, but makes up for OFDM's big PAPR in mobile stations where electricity efficiency is critical. We conduct a paper to improve receiving performance having selected OFDMA and SC-FDMA, as the transfer mode of base station and relay based on the distance of the relay that has been installed and location of base station. The paper found SC-FDMA to be better choice in RS when it is closer to BS and OFDMA to be a better choice in RS when the distance between BS and RS is farther. The system's reception performance improved when the most appropriate transmission method fitting the circumstances was used in the middle between BS and MS.

Keywords LTE · AF & DF · Multi-hop relay · Relay protocol

C. Park (✉)

SamSun Technology Research Co. Ltd, Company-affiliated Research,
564-6, Sang-dong, Wonmi-gu, Bucheon-si, Gyeonggi-do, Korea
e-mail: iparka1028@gmail.com

1 Introduction

Long Term Evolution (LTE) is to further solidify 3G Communication Service on the strong basis of GSM market and existing HSPA(HSPA+). Devoid of the backward compatibility that HSPA+ was able to support, LTE features whole new mobile communication service, overcoming limitations from 3G Communication Service and is aiming the long-term solution for years to come by way of wider coverage, higher system capacity, data transmission rate, transmission continuity (latency improvement), cost-efficiency, and overall service quality improvement. Developing upon the existing communication network, LTE seamlessly supports HSDPA and WCDMA network as well [1]. LTE incorporates 3GPP Rel-8 Standard Technologies of OFDM and MIMO to feature uplink and downlink data transmission rate of 86.4 and 326.4 Mbps. LTE accesses channels by way of Uplink Channel Access and Downlink Channel Access, each incorporating SC-FDMA, for better PARR efficiency, electricity (battery) management and simplification of mobile unit design, and OFDMA, for better frequency efficiency with disturbance. Note that such an OFDMA and MIMO are working reciprocally. Note further that LTE best operates while user is abeyant or in slow motion (at 0–15 km/h), featuring quality services when in faster motion at 15–120 km/h or even faster at 120–350 km/h, up to 500 km/h. In 3GPP LTE-advanced, hybrid OFDMA/SC-FDMA is recommended for its technological capability to make up for the performance gap between OFDMA and SC-FDMA and to make PAPR more efficient. OFDMA is used in LTE downlink in order to make frequency more efficient and raise cell capacity but OFDM struggles with a high PAPR owing to its use of multi carrier wave. LTE uplink, on the other hand, employs SC-FDMA, which is similar to OFDMA, but makes up for OFDM's big PAPR in mobile stations where electricity efficiency is critical [2]. This paper proposed relay to make up for the performance gap between OFDMA and SC-FDMA and performed research to improve reception performance by using OFDMA and SC-FDMA as transfer modes for stations and relays depending on the distance between station and relay. Stated hereunder is to discuss downlink data transmission of LTE System, as a standardized mobile communication means improving the communication disturbance between BS and MS by way of SC-FDMA, in compensation for the higher PARR caused out of a carrier wave. Featuring lower PARR as said, SC-FDMA better transfers data long-range, without respect to the coverage of BS by installing Relays for better MS operation. Such means of OFDMA and SC-FDMA are to be incorporated selectively by how BS, RS, and MS operate [3, 4].

The result of the paper is expected to improve reception performance in the overall system and provide smooth communications service by reducing shadow domain and expanding BS' cell coverage. [Section 2](#) describes the multi-hop relay system. [Section 3](#) describes relay mode and as the means enhance LTE performance. [Section 3](#) describes the relay location, transmission set-up, and Relaying performance by simulation. The conclusions are given in [Sect. 4](#).

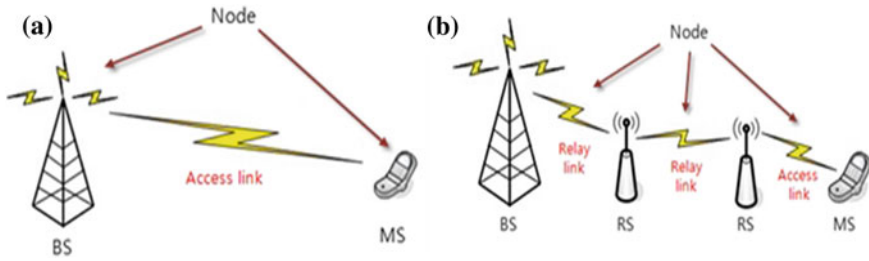


Fig. 1 Relay techniques of single-hop and multi-hop. a Single-hop, b multi-hop

2 Multi-hop Relay

Figure 1a represents direct communication between BS and MS without RS, and Fig. 1b represents communication between BS and MS via RS. Case (a) is generally called a single hop, and (b), though the number of hops may vary, a multi-hop. Basically, a multi-hop relay method requires a channel between BS, which transmits data, MS, which receives data, and RS, which gets involved in the communication and enhances performance, as illustrated in Fig. 2. BS delivers data to RS as well as MS, and RS transmits the received signal to MS at the next time slot by processing it as required. This method enables transmission of data via RS if the BS → MS channel is not in good condition, or if the BS → MS channel is in good condition, MS can combine signals received through the two channels, to provide the benefit of diversity [5, 6].

As illustrated in Fig. 2, RS plays the role of expanding the cell coverage and improving the cell performance. As illustrated in Fig. 3, in order to expand the cell coverage, RS may be used to provide service to shadow zones or cell boundaries where signals of BS are not reached, or outside of cell boundaries where service is not provided [7].

For MS which is outside of a cell, as illustrated in Fig. 3, a modulation technique with low transfer rate, such as QPSK, must be used because of high path loss of the signal from BS. In order to prevent deterioration of performance in the entire cell due to MSs on the cell boundary, relays may be used to provide the service for these MSs with the modulation method with high transfer rate, enhancing performance of the entire cell.

2.1 Multi-hop Relay Technique

Relay techniques are divided into fixed relay and selective relay depending on the method of relaying data. In the fixed relay technique, RS always relays data from BS to MS, regardless of condition of the channel. In the selected relay technique,

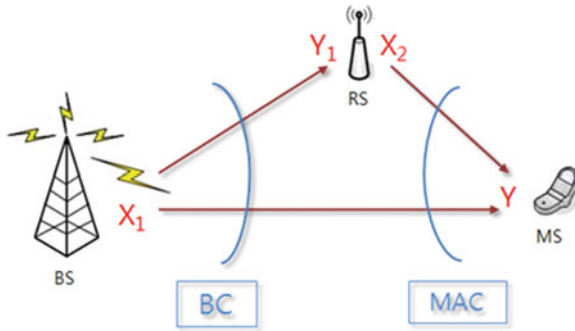


Fig. 2 Relay channel

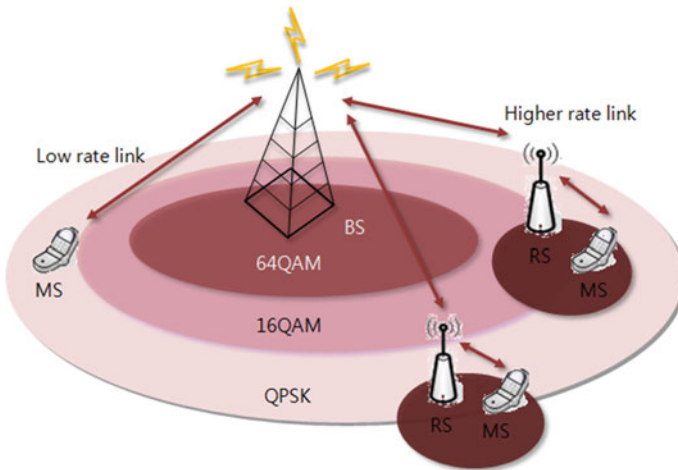


Fig. 3 Example of improve of cell throughput using relay

RS determines whether to transmit the data from BS in consideration of the channel gain [8]. This study deals with fixed relay techniques (AF and DF).

2.1.1 AF Technique

The AF technique is used in the existing relay system. In this technique RS only amplifies power of the received signal before relaying it. This technique regulates power of the received signal, and amplifies it to the level RS can transmit. It is relatively simple to implement, but has a defect that noise is also amplified [9]. Where the signal transmitted from BS is 'x', the signal transferred to MS through RS using AF technique is as follows:

$$y_D = \sqrt{\frac{E_{SR}E_{RD}}{|h_{SR}|^2 E_{SR} + N_0}} h_{SR} h_{RD} + \sqrt{\frac{E_{RD}}{|h_{SR}|^2 E_{SR} + N_0}} h_{RD} n_R + n_D \quad (1)$$

h_{SR} and h_{RD} represent BS \rightarrow RS and RS \rightarrow MS channel, respectively. n_R and n_D are the noises added to the receiving antennas of RS and MS, respectively. The noises have the same distribution but are independent from each other. The effective SNR and p_{eff} of the received signal in Eq. (1) are expressed as follows:

$$p_{eff} = \frac{p_{SR} \cdot p_{RD}}{1 + p_{SR} + p_{RD}} \quad (2)$$

Equation (1) shows that effective SNR gets lower than p_{SR} and p_{RD} if AF technology is used. If p_{SR} equals to p_{RD} , and SNR is higher, the effective SNR becomes $p_{SR}/2$, resulting in the loss of 3 dB in terms of performance. This is because amplification of power in RS also increases power of noise. From the above result, the channel capacity of AF technique is as follows:

$$C_{AF} = \frac{1}{2} \log_2 \left(1 + \frac{p_{SR} \cdot p_{RD}}{1 + p_{SR} p_{RD}} \right) \quad (3)$$

2.1.2 DF Technique

The DF technique decodes the received signal into bits, and transmits the encoded and modulated signal [10]. The DF technique is more complicated than the AF technique in terms of the amount of operation, but is practical to implement considering the fact that most communication terminals are equipped with modulator/demodulator and encoder/decoder. RS, which uses the DF technique, decodes the received signal, and transmits re-encoded and re-modulated signal to MS. The signal received in D is as follows:

$$y_D = h_{RD} \hat{x} + n_D \quad (4)$$

where, \hat{x} is the signal re-encoded and re-modulated by DF and transmitted by RS. The channel capacity of the system with the DF technique equals to the capacity of BS \rightarrow RS or RS \rightarrow MS channel, whichever has the lower SNR, and is expressed as follows:

$$C_{DF} = \min \left\{ \frac{1}{2} \log_2(1 + p_{SR}), \frac{1}{2} \log_2(1 + p_{RD}) \right\} \quad (5)$$

If the effective SNR is equivalent between the two channels, the channel capacity of the DF technique has more gain than the AF technique.

3 Proposed on Relay Structure for Improved Receiver Performance of LTE Downlink

Among all the currently used LTE systems, OFDMA mode, the LTE downlink system, as well as SC-FDMA mode, the LTE uplink system, are currently in use. SC-FDMA mode out of LTE uplink systems is employed as the standard. LTE uplink often uses CAZAC sequence for implementation of CDM to distinguish signals from each terminal in sending reference signal and control channel. CAZAC sequence, which maintains a uniform amplitude in the time/frequency domain, is suitable for increasing coverage by lowering PAPR of terminal. However, SC-FDMA mode, which has lower PAPR than OFDMA mode, is more advantaged in the boundary between the cells, but shows poor performance in the high-order modulation, has disadvantages for multiple antenna system, lacks in flexibility of resource allocation, and has difficulty in creating free pilot pattern. This paper propose a way of opting for either SC-FDMA mode or OFDMA mode, according to each terminal environment to acquire the best performance in the LTE environment.

3.1 Relay Mode

The relay operation mode used in the suggested method is assumed to be the non-transparent mode as defined in IEEE 802.16j. In the non-transparent mode, BS and relay transfer control information to MS at the same time point, so that MS can synchronize with BS or RS, and acquire the frame information. In other words, RS has different frame control information from BS and adjacent RSs, and transmits unique preamble and MAP information at the starting part of DL sub-frame. Therefore, MS recognizes RSs of non-transparent mode as a single BS. RS, however, relays all data and control information between BS and MS. RS can be controlled by BS through the centralized scheduler, or can be the subject of distributed scheduling [11].

Figure 4 illustrates an example of non-transparent mode. Different PN sequences are used between BS and adjacent RSs. RS receives R-MAP and R-DL-Burst from upper-layer RS or BS, generates own frame control information and allocate burst at the next frame. This structure enables easy expansion of coverage, but complicates scheduling in BS because delay occurs in transfer of data by 1 frame per hop. Because the non-transparent relay is useful to provide MS out of the coverage of broadcasting of BS, however, this study suggests the RS mode for the purpose of expansion of communication coverage [12].

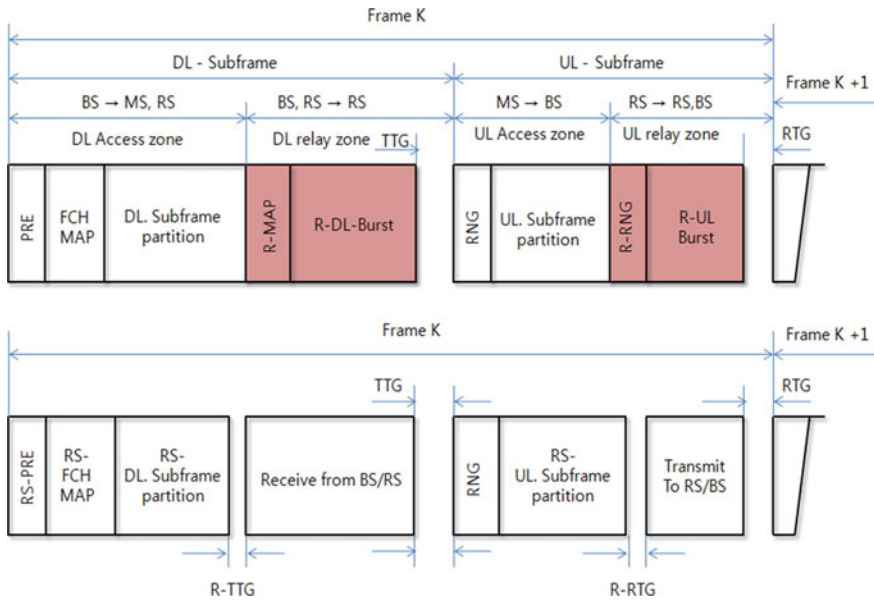


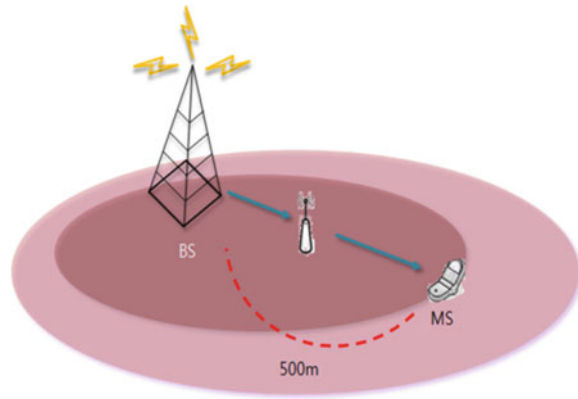
Fig. 4 Example of non-transparent mode

3.2 Receiving Performance Improvement of LTE Using Multi-hop Relay

The mode proposed in this paper facilitates the communication environment by installing RS between BS and MS to guarantee the optimal performance in LTE downlink environment, and selects either SC-FDMA or OFDMA for the location of BS and RS or RS and MS. The simulation environment shall be set forth as Table 1.

Table 1 Simulation parameter

Parameter	OFDM	SC-FDMA
FFT	256	256
Guard Period	FFT/4	FFT*Q/4
Modulation	QPSK, 16QAM	QPSK, 16QAM
Pathloss model (NLOS)	$27.7 + 40.2\log_{10}(d)$	$27.7 + 40.2\log_{10}(d)$
Tx power (dBm)	27 (back off = 6,8 dB)	27
Noise power (dBm)	-114	-114
Coding	Convolution (1/2, 1/4)	Convolution (1/2, 1/4)
Channel compensation	ZF	ZF

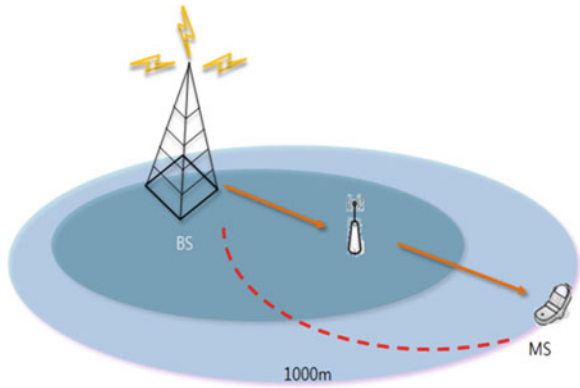
Fig. 5 MS location of 500 m

On the basis of the aforementioned simulation environment, MS and RS are properly located with BS as the center, while FFT of both OFDM and SC-FDMA is set as 256. When it comes to the protection period, it is desirable for active symbol period to be longer than protection period to minimize the SNR loss caused from protection period. As the active symbol period has to be at least four times longer than the protection period, OFDM mode is set for $\text{FFT}/4$, whereas SC-FDMA mode is set for $\text{FFT} \cdot Q/4$ to be of the same condition as OFDM. As a means of modulation by the distance between BS and MS, QPSK and 16QAM are employed. Meanwhile, back-off value is set as 6–8 dB to identify the difference in performance and distance by SNR of OFDM and SC-FDMA. MS and RS remained without power control to send data with the same transmission power. RS is then grounded in DF to get the gain in channel capacity. Lastly, a simulation was made without power control for MS and RS to send data with the same transmission power.

Figure 5 represents the case that MS is 500 m away from the center of BS. It is used to enhance performance with RS when MS is relatively near BS or is on the coverage boundary. If MS is near BS but is in the shadow area, the receiving performance is deteriorated. If MS is on the cell coverage, this MS deteriorates the performance of the entire cell. Therefore, RS is installed by distance. To find the optimum receiving performance by RS location, the researcher simulated service with diverse locations of RS between BS and MS, and compared performance gap between OFDMA and SCFDMA selectively used by MS and RS.

Figure 6 shows the case that MS is 1000 m away from the center of BS. It is used to enhance performance with RS when MS is relatively far apart from BS, and is out of the cell coverage. In this case, because of large distance between BS and RS, path loss of signal transmitted from BS increases, causing deterioration of receiving performance. The researcher placed RS between BS and MS, and compared performance gap between OFDMA and SC-FDMA selectively used to show that loss of data and deteriorated receiving performance of existing OFDM are minimized.

Fig. 6 MS location of 1000 m



4 Simulation

Figure 7 shows performance of RS transfer mode under SC-FDMA mode in BS. SC-SC outperformed SC-OF when RS location is in between 100 and 200 m since BS is closer to RS while RS is farther to MS. Performance difference is hardly observed from when RS distance was farther than 200 m. From 250 m on, SC-OF exhibited strong performance and outperformed at 300–400 m when distance between BS and RS gets farther and that between RS and MS gets closer.

Figure 8 shows receiving performance of transfer modes selected on the basis of RS location when distance between BS and MS is 1000 m. SC-SC outperformed when RS location was in between 100 and 400 m since distance between

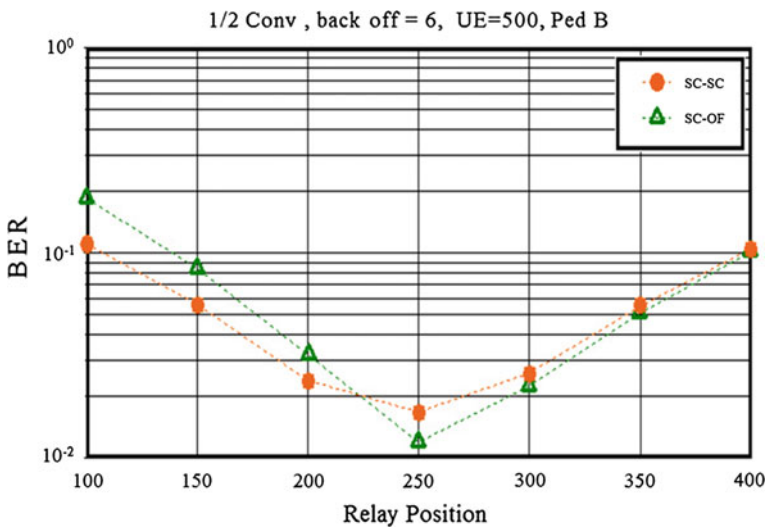


Fig. 7 BER in MS location of 500 m and OFDMA transfer mode between BS and MS

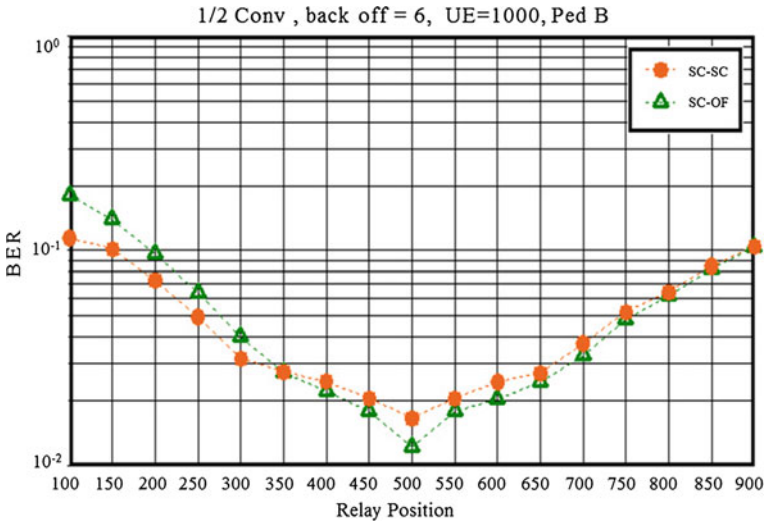


Fig. 8 BER in MS location of 1000 m and transfer mode between RS and MS

BS and RS was shorter than that between RS and MS. SC-SC was found to be outperforming SC-OF when RS is in 500, the central zone. SC-OF exhibited outstanding performance up to 600–700 m when distance between BS and RS gets farther and that between RS and MS gets shorter but the two transfer modes did not show big difference at 800–900 m. SC-FDMA transfer from BS to RS could seriously distort signal itself and RS passing through IDFT by using frequency domain equalizer spreads the impact of signal distortion, which deteriorates performance. Hence, performance does not improve just by passing through RS.

5 Conclusion

This paper proposed a combination of two transfer modes to fill up the performance gap between OFDMA and SC-FDMA and to improve reception performance of LTE system's downlink transfer mode. The paper also proposed setting up RS in between BS and MS to enhance system's performance and coverage, and ran a simulation of the proposed idea. A simulation to appropriately select one out of OFDMA and SC-FDMA based on RS set with BS was carried out by setting the location between BS and MS at 500 and 1000 m, respectively, and setting RS in between BS and MS. Simulation revealed that OFDMA is a better option in BS and SC-FDMA in RS when RS was more closely located to BS. This was just the opposite as in the longer distance between BS and RS where SC-FDMA performed better in BS and OFDMA in RS. In the center between BS and MS, an

improvement in the system's reception performance was foreseeable by selecting transfer mode befitting particular situation.

References

1. Dahlman E, Parkvall S, Skold J, Beming P (2008) 3G evolution: HSPA and LTE for mobile broadband, 2nd.edn. Academic Press, Elsevier, pp 46–81
2. 3GPP TSG RAN WG1 (2004) 3GPP TR 25.892 v6.0.0; feasibility study for orthogonal frequency division multiplexing (OFDM) for UTRAN enhancement (Rel-6), June 2004
3. 3GPP, TR 36.913 (2008) Requirements for further advancements for E-UTRA (LTE-advanced), V8.0.0, June 2008
4. 3GPP, RP-091005 (2008) Proposal for candidate radio interface technologies for LTE-advanced based on LTE. Release 10 and beyond, June 2008
5. IEEE 802.16 Broadband Wireless Access Working Group (2001) Channel models for fixed wireless applications. [Online] http://www.ieee802.org/16/tg3/contrib/802163c-01_29r4.pdf
6. Cover T, El Gamal A (1979) Capacity theorems for the relay channel. *IEEE Trans Inf Theory* IT-25:572–584
7. Hasna MO, Alouini MS (2004) Harmonic mean and end-to-end performance of transmission system with relays. *IEEE Trans Commun* 52(1):130–135
8. Laneman J, Tse D, Wornell G (2004) Cooperative diversity in wireless networks: efficient protocols and outage behavior. *IEEE Trans Inf Theory* 50:3062–3080
9. IEEE 802. 16j-06/026r2 (2007) Baseline document for draft standard for local and metropolitan area networks, part 16: air interface for fixed and mobile broadband wireless access system. Multihop relay specification, FED
10. van Nee R, Prasad R (1999) OFDM for wireless multimedia communications. Artech House, Norwood
11. Pabst Ralf et al (2004) Relay-based deployment concepts for wireless and mobile broadband radio. *IEEE Comm Mag* 42(9):80–89
12. Cover T, El Gamal A (1979) Capacity theorems for the relay channel. *IEEE Trans Inf Theory* IT-25(1):572–584

An Efficient High-Speed Traffic Control Scheme for Real-Time Multimedia Applications in Wireless Networks

Moonsik Kang

Abstract In this paper, an efficient traffic control scheme for real-time multimedia services is proposed with the use of the IEEE 802.11 WLAN for wireless access part in wireless networks, which is based on both the traffic estimation of the mean rate and header compression. The network model consists of the core RTP/IP network as well as the wireless access parts with lots of mobile hosts. This model is designed to include the means of provisioning Quality of Service (QoS) strategy according to the requirements of each particular traffic flow as well as the header compression method for real-time multimedia applications. Performance evaluation is carried out for showing the effectiveness of the proposed traffic control scheme.

Keywords High-speed traffic control · QoS strategy · Traffic estimation · Header compression · Wireless networks

1 Introduction

Most recently, the astonishing growth of the wireless mobile network technology have been facilitating new demands for multimedia applications over the Internet including both wired and wireless parts [1–3]. In order to cope with this increasing traffic requirements, several research topics have aimed at providing users for the required Quality of service (QoS) at different network layers [1, 4], in the sense

M. Kang (✉)

Department of Electronic Engineering, School of Engineering, Gangneung-Wonju National University, Gangneung, South Korea

e-mail: mskang@gwnu.ac.kr

that QoS is considered as 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. Also Real Time Protocol (RTP) is widely used as the protocol for various real-time communications and multimedia services using IP Networks, such as Multi-User online games and video/audio broadcasting, etc. The packets transferred in the RTP, when audio/video data are transferred through cable or wireless networks, are encapsulated into an IP header (20 octets) or RTP (12 octets), with the total header size of 32 (52 bytes in case of IPv6) bytes. In addition, RTP can also be used for video/audio streaming or remote video conferencing and remote patient care. So header compression, the process of reducing the header size by eliminating increased packet header overheads, is necessary for the efficient high-speed multimedia transmission. On the one hand it includes two representative methods of IPHC and ROHC. These utilize the repetitiveness of header fields, and they enable more efficient multimedia communications.

Till now we have seek for the appropriate service method which is adequate to meet diverse multimedia traffic requirements, such as efficiency and scalability. As a result we propose an efficient traffic control method using both the traffic estimation of the mean rate for QoS control and Header compression over wireless networks, which is the effective solution for the required QoS from one end of the network to the other. This may include the connection between the access point and the router, which lies at the border of the core-IP(/RTP) network. There are lots of papers which is referred that both core and wireless LAN QoS methodologies try to provide a better service for specific classes of traffic [3, 5], and not for the particular end-to-end flows [5]. Also, Cognitive Networks are promising to be the major step towards efficient and autonomic management of increasing complexity of communication networks. In this sense, our overall framework may be more specifically considering as a Class of Service (CoS) optimization for such networks. 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 of mean rate using moving average scheme. Our scheme is designed to do that the real-time traffic transmission is carried out at the Resource Control Agent (RCA) part [1].

2 Core Networks and Header Compression Method

In order to study an efficient header compression method of the RTP protocol in Multi-User online game's peer-to-peer multimedia services or video streaming services using IP wireless networks, we have executed various experiments on the MPEG streaming data with RTP/IP packets in IP Wireless Networks. Meanwhile the purpose of QoS scheme is to provide the prioritized services by the control of the bandwidth, the controlled jitter and latency, and the 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. From the point that the QoS scheme

enables us to provide a better service to certain flows, it is done by either raising the priority of a flow or limiting the priority of another flow. Referring to congestion management we try to raise the priority of a flow by queuing and servicing queues in different ways. The queue management method for congestion avoidance adapts priority by dropping lower-priority flows before servicing higher-priority flows.

The core IP network model is an appropriate architecture for implementing a scalable service differentiation in the Internet by aggregating traffic classification state [1, 6]. Instead of maintaining state information, the DS applies different Per Hop Behaviors (PHBs) that are specified by DS Code Point (DSCP) of IP header [2]. The Traffic Conditioning Framework (TCF) consists of two parts such as the traffic classifier and the traffic conditioner. The former is used to select packets from incoming packet stream according to predefined rules. In addition, two kinds of classifiers are defined in the DS model, which may be located at the ingress nodes or at interior nodes in the DS domain. 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. Two kinds of routers are identified in core network, i.e., border routers and core routers. Border routers exchange packets with other domains and perform traffic conditioning, which are allowed to keep per-flow information. 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 WLAN standard proposes an Enhanced Distributed Co-ordination Function (EDCF) for wireless access. That is, the 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).

In order to compress the header with the highest compression rate for the SN and the TS, which increases dynamically and constantly among the many fields in the RTP header, we propose the efficient scheme with some features as follows; First, the increase of SN and TS among the packets sent and received between the Compressor and the Decompressor should be uniformly maintained in the following way; The SN increases one by one if Basic Compression Bit (BCB) is determined as the demanded bit value resulted from the compression. In addition to that, the TS increases by the PiCTure clock interval (PCTSI) times. Second, the uncompressed SN and TS comprise a total of 48 bits. The Compressor and the Decompressor first decide how many bits these 48 bits should be compressed. It can be decided as the 3-bits as Basic Compression Bit. If this value is inadequate, the value of basic compression bit decides by way of adjustment between compressor and decompressor. It is the Negotiation Compression Bit (NCB). The NCB is equal to BCB if 3-bit is adequate. Third, the Compressor compresses the total 48-bits of the SN and the TS into n NCB bits, of which the n bits are transferred to

the Decompressor. The Decompressor, calculates the SN by $SN_{n-1} + 1$ with the previous values and restores the TS as $PCTSI + TS_{n-1}$. Finally, the 48-bits (16-bit SN and 32-bit TS) can be compressed into n bits. Here, let's consider that the BCB of SN and the TS between the compressor and the decompressor nodes can be set as 3-bits. However, the compressor and the decompressor decides the final NCB according to the characteristics of the payload traffic transferred through the RTP, and the SN and the TS are compressed based on this result. When the NCB value is determined as n , it is decided whether to compress the 48-bits of SN and the TS in 3-bits (BCB) or to compress them with an additional n bits in $3 + n$ (NCB). As a result, the 48-bits of SN and TS value is sent from the Compressor after being compressed in NCB (BCB + n bits, $0 \leq n < 48$). This means that long streamed or large video traffic can be compressed with a value larger than 3-bits, while small files or small traffic is compressed by the BCB, 3-bits. Here, the n NCB is decided upon by calculating basic and extended bits in order to not only increase the compression rate, but also to prevent cases where the decompressor itself has problems of restoring the lost consecutive packets more than $2^n - 1$. Considering the ability to restore consecutive packet loss, we set the n -bit size according to depending on the negotiation results of the compressor and the decompressor. Consequently we suggested that the number of compressed and sent bits be extended to $NCB(BCB + n)$.

3 The Proposed Traffic Control Strategy

3.1 Network Model for QoS Traffic Control

The access part of the proposed network model will also support legacy 802.11e users, which should be capable of interfacing with the rest of the QoS network. In this regard, the AP becomes the end point for PHB operation as Service Level Specification (SLS). The proposed model is shown in Fig. 1. Here, the BR establishes the cognitive requirements based on traffic measurement by getting the moving average of rate from the recent observed traffic load, and then accordingly asks for traffic condition from the network. Another important function of the BR is the marking of packets so that the core network may easily recognize it, which is important for the translation of information between the two networks. The BR and boundary entity of the core network are the critical elements of this integration, which consists of a translation between these UP tags and the DSCP. Four classes of the 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.

The boundary entity is co-located at the ingress router to the core part as 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). The BR is in charge of receiving

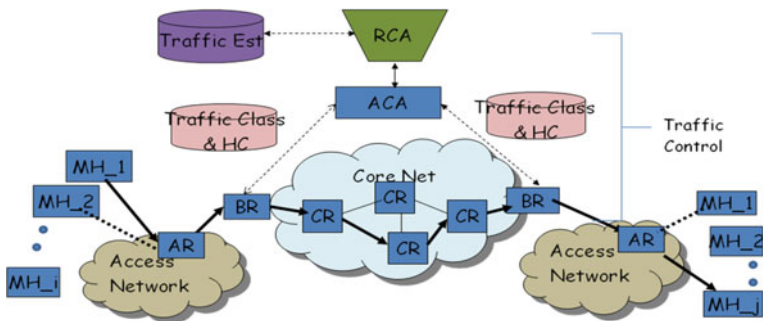


Fig. 1 Network model for high-speed traffic control

packets from the AR and marking them with an appropriate DS code point (DSCP), which 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 and should provide policing to account for falsification. RCA may be the central management entity. It is in charge of traffic monitoring, dynamically provisioning the core network based on the recent load and Time of Day, as well as indirectly controlling the admission control of BR. It may be supported by a database to store SLS information, as well as traffic condition and measurement information.

3.2 SN and TS Compression Method in Core Network

An existent method for the sequence Number (SN) in the RTP is to increase the sequence numbers of the consecutive packets by one. That is, the RTP packet with a 16-bit sequence number increases by 1 from 0 to 65535. This is because it is easy to detect loss of packet in the part of the reception. However, let's consider the case of achieving a higher compression rate. When the n th packet's SN is 3000, the SN of the $N + 1$ th SN is $3000 + 8(2^3) = 3008$. Here, SN_n is the original sequence number in the RTP packet, and also the NCB is the basic 16-bit SN compression bits. Therefore, if $n = 3$, this means that the 16-bit SN is compressed in 3-bits. When the range is expanded in the proposed method, the SN can be represented in Eq. (1), where SN_n is the n th packet's sequence of RTP and SN_{n-1} is the most recently decoded sequence number.

$$SN_n = SN_{n-1} + 1 \tag{1}$$

If the SN is compressed in such a way, the SN compression rate of the RTP header can be calculated by the following relationship.

$$C_{SN} = \frac{SN - SN_{proposed}}{SN} = 1 - \frac{SN_{proposed}}{SN} \quad (2)$$

Here, the SN is the uncompressed packet header size (byte), and the $SN_{proposed}$ is the compressed size of the SN. The C_{SN} is the compression rate of the SN, namely the relative ratio of the compressed SN size. As with the constitution of the SN, in order to achieve a higher compression efficiency, changes in the current RTP Time Stamp (TS) value should be set at a constant multiple value. We propose that the TS value of the RTP header increases by the Picture clock frequency (PCF) value. In a video stream, within the same picture the difference in TS becomes 0. The TS value for the following B or the original Intra(I) or Inter(P) pictures can be many times larger than PCF or bigger than 0. The PCF of H.261, H.263 ver 1 is 29.97 Hz, and this means that the increase between the two coded pictures is 3003. The following represents why such PCTSI value was calculated. When the video is compressed using RTP, a packet may not include more than 1 picture. A single picture may be divided into two or more packets. The header compression profile regarding video information uses 90 kHz as the reference clock.

Also, the PCF of both the H.263 ver. 2 and the MPEG-4 become 25 and 30 Hz, respectively, of which values show the increase multiple in TS of 3600 and 3000, respectively. Even if there is loss in the transfer process between the Compressor and the Decompressor, since the TS increases by the constant multiple of PCTSI, this is immediately restorable using the TS of the previous packet. Like this, if the TS value of the RTP header is configured to increase by the constant multiple of PCTSI, the TS value can be simply compressed and restored using the following formula.

$$TS_n = TS_{n-1} + PCTSI \quad (3)$$

If TS_n is the time stamp of the n th packet of the RTP, and TS_{n-1} is the most recently decoded time stamp, when the range of the time stamp is expanded by the suggested method, it can be represented by Eq. (3). If the change in TS between two packets is configured to be the multiple of PCTSI, and the PCTSI value is shared between the compressor and the decompressor, the decompressor can use the previously mentioned BCB n -bit to easily restore the TS value. Using this method, the Compressor may use either 3-bits or 3+ extended_n bits to compress and restore the 16-bit SN and the 32-bit TS. The following formula shows what the compression rate is in the TS field.

3.3 Bandwidth Allocation Method

The ingress traffic can be continuously monitored and measured by obtaining moving average of rate. Also the network continuously collects data traffics from each link. Over time this collection of data is used to characterize the behavior of

the traffic, for instance, which what kind of traffic dominates at a particular time of day. This enables us to compile data about traffic patterns over time, which helps us to define the required parameters in the core of the network to support this variation of traffic over time. The basic concept within a CR is that of four priority queues—One for Expedited Forwarding (EF) traffic, two queues for AF4x and AF2x (i.e., Assured Forwarding traffic), and the fourth for BE (i.e., Best effort traffic). The different priorities for weighting algorithms such as Priority Queuing (PQ) govern each of these queues. The Traffic1 queue, the highest priority one, passes through a single weighing stage of PQ, while Traffic2 and Traffic3 traffics pass through two levels of weighing stage of WFQ. Thus, the S1 has highest precedence and the least number of weighing stages. Like this, the network service maps to different Traffic Classes. For instance, the real time voice traffic is for Traffic1, which is implemented using EF and so on. The traffic matrix is specified based on traffic patterns according to traffic measurement method. In the normal, we can safely provision the network according to our pre-specified matrix by allocating the optimal bandwidth. 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 enters an abnormal state. In this case, the network reacts by further changing the weight in discordance with the matrix above. As soon as the network returns to a normal state, parameters are brought back to the recommended values.

4 Performance Evaluation

Here, we consider some QoS parameters, such as delay, packet loss, and throughput, in order to evaluate the performance of the proposed model. The delay parameter may refer to either propagation delay or round-trip delay. The packet loss parameter refers to the ratio between the number of lost packets from source to destination. The Maximum Transmit Unit (MTU) in the RTP layer is 4000 bits. This means that the video frames larger than 500 bytes are segmented and are transported after being divided into two or more RTP packets. However, in this experiment, it assumed that the RTP packets are divided in a set size, and the algorithm and the process of the MPEG4 stream being divided into RTP packets are not considered. In this experiment, only the constant size RTP packets are used as test subjects. Two channels were used between the two nodes: the data channel where data are transferred between the Compressor and the Decompressor, and the feedback channel where the ACK or the NAK signal is transmitted from the Decompressor. The results show that, for smaller lengths of consecutive packet loss, the success rate for the 3-bit compression method is increased, while the error rate for the method of large compression bit number of the header is decreased as the length of consecutive packet loss gets longer. In addition, for 3-bit compressions, we could see that the difference between consecutive error rates of 10 and 50 was more than 8 times, while the 9-bit

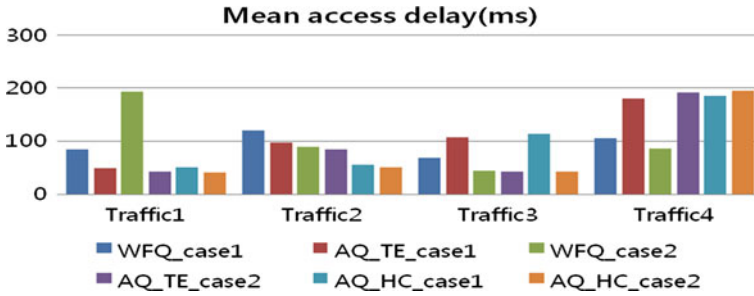


Fig. 2 Mean access delay according to traffic types for two cases

compression showed not much effect from the consecutive error lengths, as the error rate was maintained from 0.4 to 0.6.

Also our network model is described as the collection of the individual CR nodes and link states, of which traffic consists of entities in the core network as well as access part. The traffics are represented explicitly as they determine the behavior of nodes and links at a particular point in time. In our simulation, the performance is evaluated from the view of the network efficiency by the mean delay. Also, the traffic arrival process is selected as Poisson model with mean throughput of 150 Mb/s for the simulation network model. In the simulation work, the proposed priority queuing based on traffic estimation (AQ_TE) based on header compression (AQ_HC) and Weighted Fair queuing (WFQ) functions are implemented respectively between BR and the CR in the core network. Figure 2 shows the comparison of mean access delay according to each traffic class for two cases. Case 1 means that the traffic amount of each class occurs at the same rate, meanwhile case 2 means that the traffic occurrence rates for each class are 38, 22, 15, and 25 % in order for Traffic1, Traffic2, Traffic3, Traffic4, respectively. These results show that the delay performance of Traffic1 (with relative high priorities) of AQ_TE is much improved as compared with other lower priority traffics, and much better performance for Traffic2 (i.e. video traffic) of AQ_HC for the case 2.

5 Conclusion

In this paper, an efficient traffic control scheme based on traffic estimation of the mean rate as well as header compression technique is introduced for the multimedia applications in wireless networks. This scheme can provide the efficient end-to-end QoS performance between mobile hosts over IP/(RTP) core network including wireless access parts. All the data traffics are classified as four different types and then they are accordingly processed differently passing through BR, and CR through both the wireless access part and the core network. The performance of the proposed scheme is evaluated for the different traffic patterns in simulation network model in some aspect of mean access delay. Simulation results show that

the proposed control scheme has the better performance of the low delay. This solution will be also a scalable one in the core network because of the low-overhead performance because of both the behavior aggregation and header compression. For further study, we will study the optimal adaptive resource control strategy to cooperate with various wireless access networks over cognitive environments.

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. *IEEE Commun Mag* 45(2):96–103
3. 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
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):1005–1009
6. Qin D, Shhroff N (2004) a predictive flow control scheme for efficient network utilization and Qos. *IEEE/ACM Trans Netw* 12(1):161–172

Interference Assessment on the Circuit Domain in GSM-R Networks by Grey Clustering and Analytic Hierarchy Process

Si-Ze Li, Zhang-Dui Zhong, Yuan-Yuan Shi, Bo Ai, Jian-Wen Ding and Si-Yu Lin

Abstract In the high-speed railway communication networks, interference is quite complicated and serious interference can even ruin the security of trains. So how to evaluate the effect is becoming more and more important. Both grey clustering evaluations and analytic hierarchy process are an effective comprehensive evaluation theory. In this paper, we apply the theory of grey clustering evaluations and analytic hierarchy process to assess the interference of the circuit domain in GSM-R networks and make a comprehensive evaluation on interference. Based on the theory of grey clustering evaluations, the interference is sorted into rough groups. Then further classification of the interference can be obtained. The results show that grey clustering evaluations combined with analytic hierarchy process can provide the reliable interference evaluation in the railway services.

Keywords Quality of service · Interference · Circuit domain · Grey clustering evaluations · Analytic hierarchy process

S.-Z. Li (✉) · Z.-D. Zhong · Y.-Y. Shi · B. Ai · J.-W. Ding · S.-Y. Lin
State Key Laboratory of Rail Traffic Control and Safety, and the Department of Mathematics, Beijing Jiaotong University, Beijing 100044, China
e-mail: lsz00@163.com

Z.-D. Zhong
e-mail: zhdzhong@bjtu.edu.cn

Y.-Y. Shi
e-mail: 09111053@bjtu.edu.cn

B. Ai
e-mail: boai@bjtu.edu.cn

J.-W. Ding
e-mail: jwding@bjtu.edu.cn

S.-Y. Lin
e-mail: 07111017@bjtu.edu.cn

1 Introduction

In the railroad communications system, it is generally known that the security and reliability are the key factors of the QoS (Quality of Service QoS), and they are the first principle during the period of railway construction. However, the infrastructure of the GSM-R networks is in a complex electromagnetic environment, and the GSM-R communication system used for Chinese Railway stays at special band. So, the interference is more complicated than the other mobile communication systems ruining the security of trains [1]. It is very important and necessary to identify the interference source, analyze the interference state, and evaluate the level of the interference.

At the present time, the related research is focused on analyzing the interference, and aimed at the statistical characteristics of the interference signals [2], such as the probability density function, the n th order moment or the characteristic function of the interference signals. Moreover, the evaluation algorithm just used only one parameter. However, the QoS of the communication system is a comprehensive concept, if we just focus on single parameter, the overall performance may be neglected. So we should evaluate the interference base on the comprehensive QoS parameters, and make an assessment on the interference by the measured value of QoS parameters.

In this paper, we apply the theory of grey clustering evaluations and analytic hierarchy process on the interference of the circuit domain in the GSM-R networks, and provide reliable interference level classification.

2 Interference in the GSM-R Networks

GSM-R (GSM for railway, GSM-R), which is based on GSM, the regarded most mature and commonly-used public wireless communication system on global, is an integrated digital communication network used specially in railway system. Nowadays, GSM-R system plays a more and more important role in the development of the railway, and it has become the direction of the railway communication.

From a frequency band perspective, there are three main kinds of interference for GSM-R. First, GSM-R is a frequency limited system, and how to reuse the finite frequency is the focus of the communication engineers and scholars. If the frequency is not reused reasonably, and the cells which use the same frequency band are not far away enough, then the interference will be generated. This kind of interference is called co-channel interference. Second, sometimes the power of the interferer signals will fall into the passband of the receivers which use the adjacent frequency band, and then we call this kind of interference adjacent frequency interference. At last, nature interference is the interference which caused by the natural phenomenon and heavy weather. In communication systems, interference

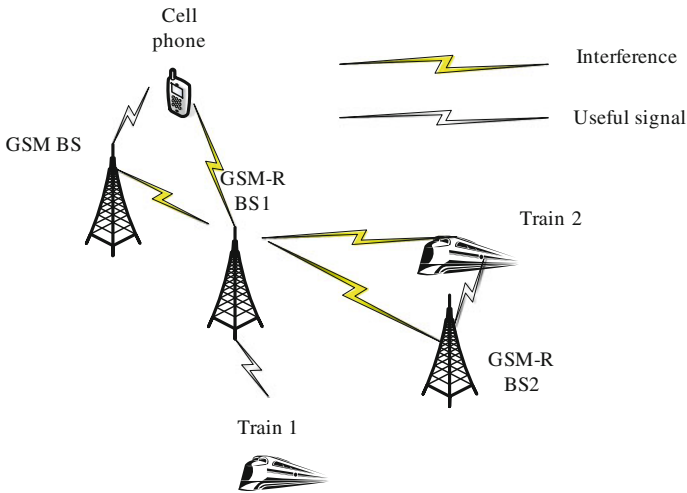


Fig. 1 Different kind of interference in GSM-R networks

is one of the main factors that degrade the quality of service, sometimes even interrupt the communication, and thus, it will bring serious trouble to the security and reliability.

From the point view of the interference source, the interference in the GSM-R networks can be divided into inner-system interference and inter-system interference. Inner-system interference is mainly caused by unreasonable frequency planning in GSM-R networks, while inter-system interference includes the interference between GSM-R and other public communication networks, such as GSM and WCDMA. Figure 1 shows some kinds of interference in the GSM-R networks.

3 Application of Grey Clustering Evaluations in Interference Assessment

Grey clustering evaluation, proposed by Mr. Deng Julong, is an effective evaluation method to make full use of known information to get the unknown information [3, 4]. One of the strengths of grey clustering evaluation is to obtain the classification of the observational objects.

Our analysis is based on six scenarios samplings (I, II, ..., VI) of the circuit domain QoS attributes, and we can get the interference level by analyzing the QoS parameters in the circuit domain. To evaluate the interference of the circuit domain in GSM-R networks, we propose seven evaluation parameters based on the theory of grey clustering evaluations: data call setup times, data call setup failure rates, frame delivery delay, data transmission link outage probability, interference times, error-free times, network registration delay, and all the sampling values have been

Table 1 The evaluation parameters

Object	Attribute (j)						
(i)	Data call setup times (<8.5 s)	Data call setup failure rates	Frame delivery delay (≤0.5 s)	Data transmission link outage probability	Interference times (<1 s)	Error-free times (>20 s)	Network registration delay (≤30 s)
I	0.930	0.003	1.000	0.001	0.995	0.960	0.950
II	1.000	0.000	1.000	0.001	0.990	0.950	0.910
III	0.990	0.001	0.990	0.000	0.940	0.900	0.940
IV	0.920	0.006	0.940	0.003	0.890	0.960	0.910
V	0.940	0.005	0.990	0.004	0.930	0.950	0.960
VI	0.980	0.003	0.990	0.000	0.990	0.950	0.990

Table 2 The threshold value

Interference grey class (k)	Threshold value(C _{jk})						
	Data call setup times (<8.5 s)	Data call setup failure rates	Frame delivery delay (≤0.5 s)	Data transmission link outage probability	Interference times (<1 s)	Error-free times (>20)	Network registration delay (≤30 s)
1 Good	0.970	0.003	0.995	0.003	0.995	0.970	0.950
2 Acceptable	0.950	0.010	0.990	0.010	0.990	0.950	0.920
3 Unacceptable	0.930	0.030	0.970	0.020	0.950	0.900	0.900

shown in Table 1. In Table 1, the QoS attribute is noted j, j = 1, 2, 3, 4, 5, 6, 7. The threshold values of them which are noted C_{jk} have been shown in Table 2, and the C_{jk} means the value of the jth attribute which belongs to the kth grey classification. Further results are presented as follows, where k = 1 represents “minor interference”, and the interference state is good; k = 2 represents “acceptable interference”, and k = 3 represents “serious interference”, and it means the interference state is unacceptable.

We process the data with normalization of Table 1 to lead to following matrix:

$$(x_{ij})_{6 \times 7} = \begin{pmatrix} 0.930 & 0.997 & 1.000 & 0.999 & 0.995 & 0.960 & 0.950 \\ 1.000 & 1.000 & 1.000 & 0.999 & 0.990 & 0.950 & 0.910 \\ 0.990 & 0.999 & 0.990 & 1.000 & 0.940 & 0.900 & 0.940 \\ 0.920 & 0.994 & 0.940 & 0.997 & 0.890 & 0.960 & 0.910 \\ 0.940 & 0.995 & 0.990 & 0.996 & 0.930 & 0.950 & 0.960 \\ 0.980 & 0.997 & 0.990 & 1.000 & 0.990 & 0.950 & 0.990 \end{pmatrix} \quad (1)$$

And we can get the critical values as Table 3 too.

We can obtain the following tables (Table 4) by

Table 3 The critical value

Grey class	Critical values						
	λ_1^k	λ_2^k	λ_3^k	λ_4^k	λ_5^k	λ_6^k	λ_7^k
k = 1	0.970	0.997	0.995	0.997	0.995	0.970	0.950
k = 2	0.950	0.990	0.990	0.990	0.990	0.950	0.920
k = 3	0.930	0.970	0.970	0.980	0.950	0.900	0.900

Fig. 2 The first grey class

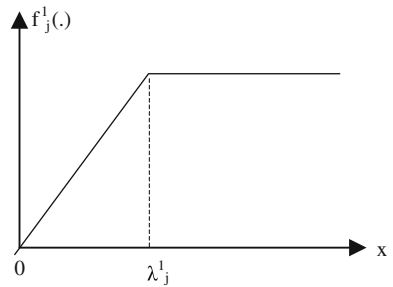
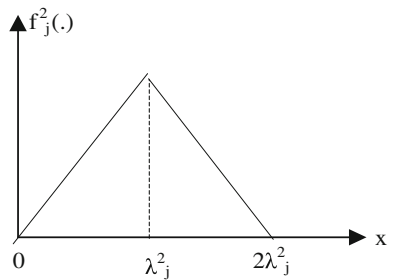


Fig. 3 The second grey class



$$\eta_j^k = \lambda_j^k / \sum_{j=1}^m \lambda_j^k \tag{2}$$

where is η_j^k called the weight of the j th criterion with respect to the k th subclass.

Corresponding whitenization weight functions are as follows: (Figs. 2, 3, 4)

The cluster coefficient of variable weight for object i belonging to the k th grey class is defined as

$$\sigma_i^k = \sum_{j=1}^m f_j^k(x_{ij}) \cdot \eta_j^k \tag{3}$$

If

$$\sigma_i^{k*} = \max_{1 \leq k \leq s} \{\sigma_i^k\} \tag{4}$$

Fig. 4 The third grey class

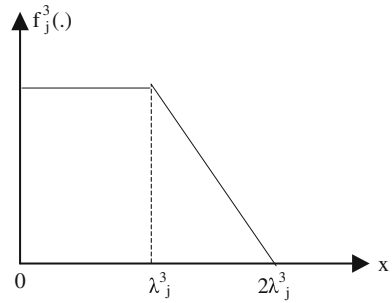


Table 4 The weights for grey clustering

Grey class	Weights						
	η_1^k	η_2^k	η_3^k	η_4^k	η_5^k	η_6^k	η_7^k
k = 1	0.1411	0.1450	0.1447	0.1450	0.1447	0.1411	0.1382
k = 2	0.1401	0.1460	0.1460	0.1460	0.1460	0.1401	0.1357
k = 3	0.1409	0.1470	0.1470	0.1485	0.1439	0.1364	0.1364

Table 5 The list of the cluster coefficients

σ_i^k	k = 1	k = 2	k = 3	$\sigma_i^{k*} = \max_{1 \leq k \leq 3} \{\sigma_i^k\}$	Grey class
I $i = 1$	0.9925	0.9865	0.9651	$0.9925 = \sigma_1^1$	$k^* = 1$
II $i = 2$	0.9904	0.9868	0.9624	$0.9904 = \sigma_2^1$	$k^* = 1$
III $i = 3$	0.9794	0.9735	0.9745	$0.9794 = \sigma_3^1$	$k^* = 1$
IV $i = 4$	0.9616	0.9688	0.9833	$0.9833 = \sigma_4^3$	$k^* = 3$
V $i = 5$	0.9819	0.9820	0.9727	$0.9820 = \sigma_5^2$	$k^* = 2$
VI $i = 6$	0.9954	0.9839	0.9551	$0.9954 = \sigma_6^1$	$k^* = 1$

we can say that object i belongs to the grey class k^* . The results are shown in the Table 5.

The results show that I, II, III, and VI belong to the first grey class; V belong to the second grey class; and IV belong to the third grey class.

According to the results of the evaluation of the interference of the circuit domain in GSM-R networks, we can know the interference state clearly and comprehensively. For example, in the fourth sampling scenario, we evaluated the interference belonging to the third grey, and the results told us that the interference

is very serious and it cannot acceptable. Then we need to take effective measures to eliminate the interference signals.

4 Further Interference Classification by Applying Analytic Hierarchy Process

On the basis of the result of the above analysis, I, II, III, and VI belong to the same grey class. In this section, a further classification of these four scenarios is made by applying Analytic Hierarchy Process.

The hierarchy tree is shown in Fig. 5. In the hierarchy tree, there are three layers. The top layer means the goal of the assessment problem, i.e., interference evaluation in the circuit domain, labeled as A. The middle layer is the criteria layer, including data call setup times, data call setup failure rates, frame delivery delay, data transmission link outage probability, interference times, error-free times, network registration delay, labeled as B₁, B₂,..., B₇. The bottom layer is the alternative layer, and there are four alternatives to be evaluated.

According to [5] and [6], the comparison matrix for all the criteria as respect to the goal can be obtained as A.

$$A = \begin{pmatrix} 1 & 1/3 & 1/5 & 1/5 & 1/7 & 3 & 3 \\ 3 & 1 & 1/2 & 1/2 & 1/3 & 5 & 5 \\ 5 & 2 & 1 & 1 & 1/2 & 7 & 7 \\ 5 & 2 & 1 & 1 & 1/2 & 7 & 7 \\ 7 & 3 & 2 & 2 & 1 & 9 & 9 \\ 1/3 & 1/5 & 1/7 & 1/7 & 1/9 & 1 & 1 \\ 1/3 & 1/5 & 1/7 & 1/7 & 1/9 & 1 & 1 \end{pmatrix} \tag{5}$$

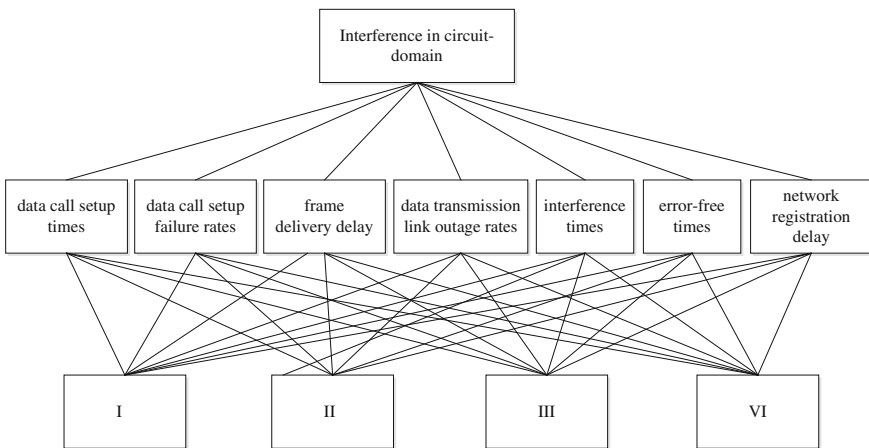


Fig. 5 Hierarchy tree

Considering $[Ax = \lambda_{\max}x]$, x is the eigenvector of size 7×1 , also called the priority vector (by normalizing); λ_{\max} is the eigenvalue, $7 \leq \lambda_{\max} \in R$.

By computing, we obtain eigenvalue $\lambda_{\max} = 7.1409$, eigenvector $x = (0.0554, 0.1249, 0.2109, 0.2109, 0.3425, 0.0276, 0.0276)^T$, i.e., criteria weights vector is $(0.0554, 0.1249, 0.2109, 0.2109, 0.3425, 0.0276, 0.0276)^T$.

Analytic Hierarchy Process evaluations are based on the assumption that the decision maker is rational. So, it is very important to judge the consistency of the comparison matrix A . According to Saaty in [5], the consistency of the comparison matrix can be measured by Consistency Ratio (CR), and $CR = CI/RI$, where $CI = (\lambda_{\max} - n)/(n - 1)$; n is the order of the comparison matrix; RI is the corresponding index of consistency for random comparison matrix. The matrix is of consistency when $CR < 0.1$.

In the above case: $CR = CI/RI = 0.0178 < 0.1$, so the evaluations are consistent.

Similarly, we can get the comparison matrix for the alternatives with respect to B_1, B_2, \dots, B_7 criteria respectively. Then seven matrixes can be obtained, labeled as matrix B_1, B_2, \dots, B_7 .

$$\begin{aligned}
 B_1 &= \begin{pmatrix} 1 & 1/7 & 1/3 & 1/5 \\ 7 & 1 & 5 & 3 \\ 3 & 1/5 & 1 & 1/3 \\ 5 & 1/2 & 3 & 1 \end{pmatrix} & B_2 &= \begin{pmatrix} 1 & 1/3 & 1/2 & 1/3 \\ 3 & 1 & 2 & 1 \\ 2 & 1/2 & 1 & 1/2 \\ 3 & 1 & 2 & 1 \end{pmatrix} \\
 B_3 &= \begin{pmatrix} 1 & 1 & 3 & 3 \\ 1 & 1 & 3 & 2 \\ 1/3 & 1/3 & 1 & 1 \\ 1/3 & 1/2 & 1 & 1 \end{pmatrix} & B_4 &= \begin{pmatrix} 1 & 1 & 1/5 & 1/5 \\ 1 & 1 & 1/3 & 1/5 \\ 5 & 3 & 1 & 1 \\ 5 & 5 & 1 & 1 \end{pmatrix} \\
 B_5 &= \begin{pmatrix} 1 & 3 & 7 & 5 \\ 1/3 & 1 & 9 & 1 \\ 1/7 & 1/9 & 1 & 1/7 \\ 1/5 & 1 & 7 & 1 \end{pmatrix} & B_6 &= \begin{pmatrix} 1 & 1/5 & 3 & 1/9 \\ 5 & 1 & 6 & 1/4 \\ 1/3 & 1/6 & 1 & 1/7 \\ 9 & 4 & 7 & 1 \end{pmatrix} \\
 B_7 &= \begin{pmatrix} 1 & 3 & 2 & 1 \\ 1/3 & 1 & 2 & 1 \\ 1/2 & 1/2 & 1 & 1/2 \\ 1 & 1 & 2 & 1 \end{pmatrix}
 \end{aligned}$$

We can obtain eigenvalues $\lambda_{\max}(B_1) = 4.2049$, $\lambda_{\max}(B_2) = 4.0104$, $\lambda_{\max}(B_3) = 4.0206$, $\lambda_{\max}(B_4) = 4.0328$, $\lambda_{\max}(B_5) = 4.3084$, $\lambda_{\max}(B_6) = 4.3097$, $\lambda_{\max}(B_7) = 4.1545$; eigenvectors (weights vector) $x_1 = (0.0539, 0.5550, 0.1139, 0.2773)^T$, $x_2 = (0.1091, 0.3509, 0.1890, 0.3509)^T$, $x_3 = (0.3828, 0.3475, 0.1276, 0.1420)^T$, $x_4 = (0.0864, 0.0990, 0.3828, 0.4319)^T$, $x_5 = (0.5586, 0.2188, 0.0394, 0.1832)^T$, $x_6 = (0.0792, 0.2526, 0.0475, 0.6208)^T$, $x_7 = (0.3763, 0.2132, 0.1368, 0.2736)^T$. And the evaluations are consistent.

We can obtain the criteria weights vector as following:

$$(x_1^T, x_2^T, x_3^T, x_4^T, x_5^T, x_6^T, x_7^T)x = (0.3195, 0.2565, 0.1561, 0.2677)^T$$

Since $0.3195 > 0.2677 > 0.2565 > 0.1561$, we obtain a sort of I, II, III, and VI such as I, VI, II, III. The results mean that the interference in the scenario I is the weakest, and the scenario III is suffering the stronger interference than scenario I, VI and II, though I, II, III, VI belong to the same grey glass.

5 Conclusions

This paper presents the theory of grey clustering evaluations and analytic hierarchy process applied in interference of the circuit domain in GSM-R networks. A rough sorting is obtained by grey clustering evaluations, while we get further classification of the interference by analytic hierarchy process. The results can be used as the basic theory to provide the reliable communications in the railroad services.

Acknowledgments This work was supported by the State Key Lab of Rail Traffic Control and Safety under Grant RCS2010ZT012, the Key Project of State Key Lab of Rail Traffic Control and Safety under Grant RCS2008ZZ006, RCS2008ZT005 and RCS2010K008, the NSFC under Grant 60830001, the Program for Changjiang Scholars and Innovative Research Team in University under Grant No. IRT0949 and the Program for New Century Excellent Talents in University under Grant NCET-09-0206.

References

1. An H (2007) An new method of positioning interference in GSM-R network in China. In: 2007 International symposium on electromagnetic compatibility, Beijing, pp 83–87
2. Shi Y, Chen X Zhu G (2011) Interference evaluation in wireless communication system. In: IEEE international conference on service operations, logistics and informatics, Beijing, pp 438–442
3. Liu SF, Dang YG, Fang ZG, Xie NM (2010) The theory and application of grey system. Science Press, Beijing, pp 108–118
4. Liu SF, Xie NM (2011) Novel models of grey relational analysis based on visual angle of similarity and nearness. *Grey Syst Theory Appl* 1(1):8–18
5. Saaty TL (2000) Fundamentals of decision making and priority theory with the analytic hierarchy process, vol 6. RWS Publications, Pittsburgh
6. Ramanathan R (2001) A note on the use of the analytic hierarchy process for environmental impact assessment. *J Environ Manag* 63:27–35

Design of Transducer Interface Agent and its Protocol for WSN Middleware

Surgwon Sohn

Abstract This paper presents a simple design and implementation of a transducer interface agent. The agent communicates with a middleware for wireless sensor networks, and can be installed on the PC or any small gateway devices. The agent's chief end is to replace the network capable application processor (NCAP) and wireless transducer interface module (WTIM) with the agent itself in the IEEE 1451 family of standards. In order to support various non-standardized transducers, we propose to use a variety of configuration template files instead of modifying the agent itself. While a wrapper of the transducer interface agent manages various types of sensors, the controller deals with many actuators. In order to justify the usefulness of the proposed agent and its protocol, we carried out health monitoring experiments in a mock-up railway system and verified the efficiency.

Keywords Transducer interface agent • Wireless sensor networks • Middleware

1 Introduction

In wireless sensor networks, a lot of transducers such as sensors and actuators provide self-defined application programming interfaces (APIs) by each manufacturer. These give rise to increase development costs of application programs excessively and decrease the interoperability between transducers and sensor network systems [1]. In order to maintain device independence and compatibility between networks and

S. Sohn (✉)
Graduate School of Venture, Hoseo University, 9 Banpo-daero,
Seocho-gu, Seoul, South Korea
e-mail: sohn@hoseo.edu

transducers such as sensors and actuators, the National Institute of Standards and Technology (NIST) and the IEEE have offered the IEEE 1451 family of standards, which provide a common interface protocol. To achieve this end, many studies have explored this smart transducer interface. Song and Lee have published many works which include an implementation of IEEE 1451 standards [2, 3]. Moreover, some research has suggested that a web service for transducers is embedded in the network capable application processor (NCAP) [4–6].

Standardized transducers follow the family of IEEE 1451 specifications, but complying with the standard interface protocol requires a great deal of work and is expensive. Moreover, many sensors and actuators on the market do not comply with the IEEE 1451 standards at this time.

Recently, widely used TinyOS in the wireless sensor networks is combined with the IEEE 1451 standard capabilities to use different data formats and protocols which cause the most critical bottleneck in the network [7]. To overcome this situation, we present a simple design of a transducer interface agent which communicates with the middleware, replacing both the NCAP and a wireless transducer interface module (WTIM) with the agent for wireless sensor networks. We are basically motivated by a software defined radio in a radio communication system in that hardware can be replaced with software [8]. In this paper, the proposed transducer interface agent and its protocol for WSN middleware present the relevant solution to settle the previous defects. Especially, since XML-based transducer interface protocols support standard-compliant transducers as well as non standard-compliant ones, a high practical use of the proposed technique is expected.

2 Design of Transducer Interface Agent and Its Protocol

The IEEE 1451.5 standard defines a set of wireless sensor interfaces for the IEEE 1451 family. It introduces the concept of the WTIM, which is a module containing a radio running either the Zigbee (IEEE 802.15.4), WiFi (IEEE 802.11) or Bluetooth (IEEE 802.15.1) wireless communications protocol. The IEEE 1451.5 standard also establishes a set of specifications for communication between the WTIM and the NCAP, as shown in Fig. 1. Through the NCAP, sensed data can be available for the network and the web access.

We propose a simple design and implementation of a transducer interface agent (TIA) in Fig. 2 which replacing both the NCAP and the WTIM in Fig. 1. The agent communicates with the middleware for the wireless sensor networks, and provides APIs to the non-standardized sensors and actuators. The proposed TIA replaces both the WTIM and the NCAP for non-standardized transducers. Sensors and actuators connect directly to the TIA and send/receive data to/from the agent. The agent has many template files for configuring the transducers. Each XML-based configuration template file can be understood as emulating the applicable transducer electronic data sheet (TEDS).

Fig. 1 IEEE 1451.5-Zigbee standard with NCAP and WTIM

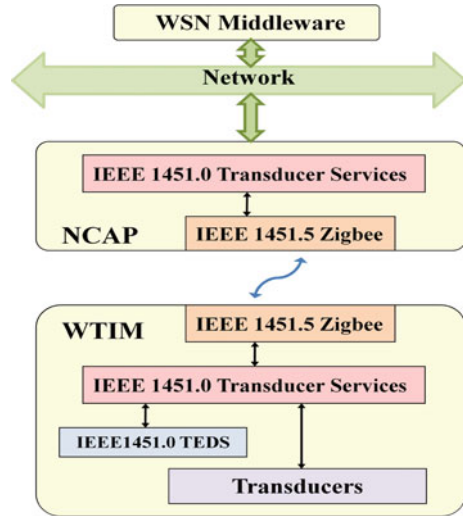
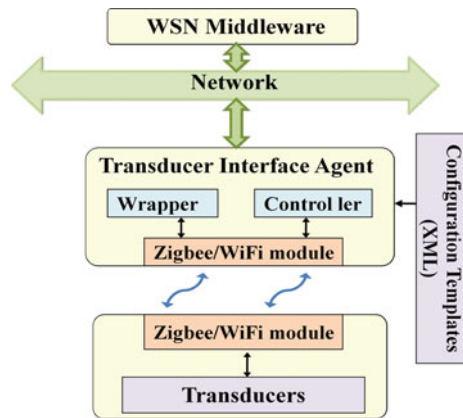


Fig. 2 The architecture of the transducer interface agent



The functional diagram in Fig. 2 presents our proposal graphically. The agent includes a communication module which can be implemented on the PC or a small gateway device on the network such as a WiFi access point, a PDA, or even a smart phone. While the TIA has a wrapper manager to handle various non-standardized sensors, the controller manager is embedded into the agent to deal with many different actuators. For our functional test, we implemented the agent on the PC.

The agent communicates with the WSN middleware on top of an operating system such as Windows. The middleware runs on Microsoft .Net Framework, which provides a high-performance virtual machine [9, 10]. The detailed structure of WSN middleware consists of seven modules, one message table, and one user interface (UI) as described in Fig. 3. In terms of layer, the middleware is largely divided into an agent IO layer, a context processing layer, and an application layer.

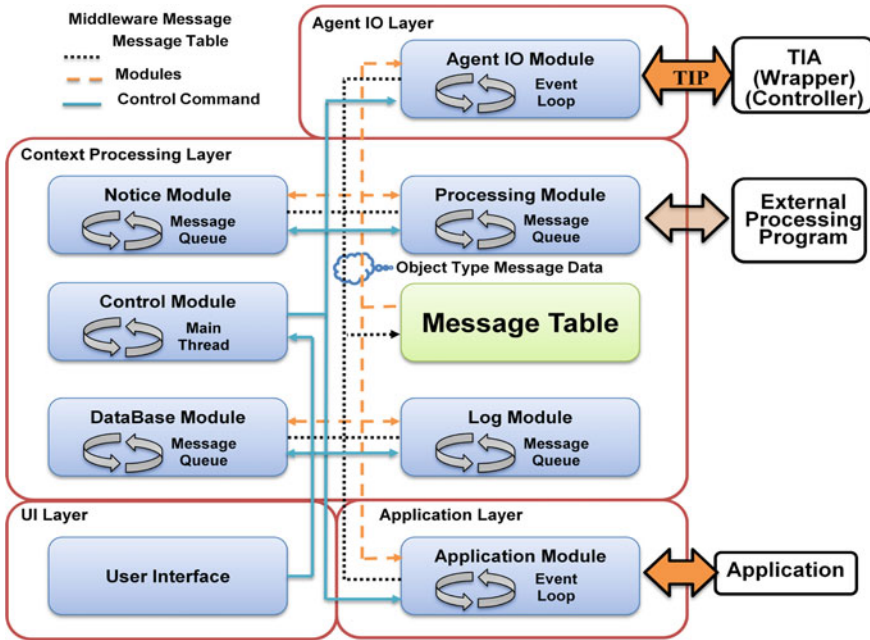


Fig. 3 The structure of WSN middleware

The agent IO layer is for communication with TIA. A data packet is converted into a middleware message in the agent IO module, and then the message is marked with the destination module and stored in the corresponding module queue of the message table. The application layer is for communication with user programs. If the application module receives a data request from a user program, it processes a middleware message in accordance with the external user’s request and then transmits the message.

The notice, processing, database, and log modules of the context processing layer read corresponding middleware messages from the module queues and execute the messages. If necessary, we can change or modify only the function of the corresponding module because every module except UI and the message table is structured independently in the form of dynamic linking library (DLL). For example, if the database is changed from Oracle to MS-SQL, we change only the database module.

The communication between the middleware and the agent uses the transducer interface protocol (TIP) composed of a header and data. Table 1 shows the 20-byte TIP header.

While the source address uses 1 byte, the destination address uses 2 bytes: one for the relaying address, and the other for the destination address. The usual destination is the WSN middleware. The packet length is the total size of the header and data combined. The data packet can be a maximum of 4 GB because the data type of a

Table 1 Transducer interface protocol header

Byte offset	Field name	Size (byte)
0	Source address	1
1	Destination address	2
3	Group address	1
4	Packet length	4
8	Sequence number	2
10	Date (YYMMDD)	4
14	Time (HHMMSS)	3
17	Time (10 ms)	1
18	Template index	2
20	Data	Object

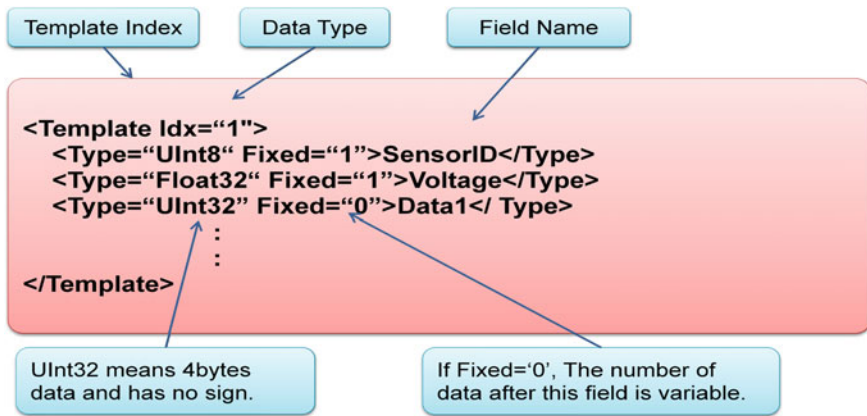
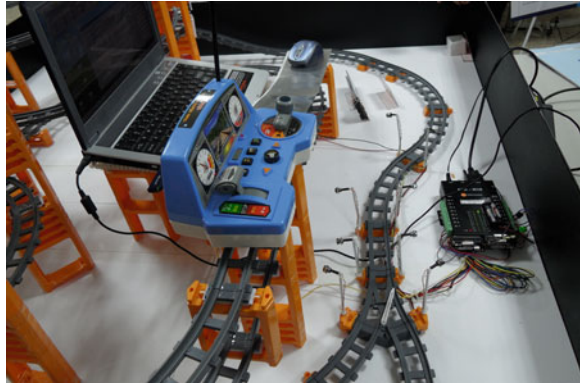


Fig. 4 Transducer data schema defined by XML

packet length is 4 bytes long. The sequence number is the order of header transmitted from the agent. The template index to configure the transducers is 2 bytes long, thus we can recognize a total of 65,536 different transducer data formats; these formats are defined by XML for each transducer, (see Fig. 4, below) [11].

Figure 4 represents a template index from an arbitrary sensor. The template is composed of field names such as Sensor ID, Voltage, and Data1. Each field name has its own properties like “Type” and “Fixed”. The type property of the Sensor ID field is one byte long, thus the template can support up to 256 different sensors of the same type. The second field name, Voltage, is 4 bytes long and is used to hold the voltage value measured from the specific sensor. The total size of Data1 is not fixed but variable; each piece of data is 4 bytes long.

Fig. 5 A mock-up railway for monitoring and controlling



3 Experiments

As an experiment to justify the usefulness of the proposed TIA and XML-based configuration template files for the wireless sensor networks, we built a mock-up railway system on which to exploit some accelerometers and PZT sensors, and control some DC motors. Figure 5 shows the experimental setup to monitor the structural health of the mock-up railway system.

For the experiment, Imote2 sensor nodes with three-axis accelerometers and piezoelectric (PZT) sensors are used. The DC motors are controlled wirelessly from the agent through a WiFi-to-RS232 remote controller [12]. The transducer interface agent makes a packet from sensor data measured with the accelerometers and the PZT sensors on the mock-up railway. By using the transducer interface protocol, the agent sends the packet to the middleware. Then, the middleware analyzes the packet, converts it into a middleware message, and interworks with the external processing program. In this case, the external processing program is the structural health monitoring MATLAB codes and communicates through the processing module of the middleware.

The accelerometers measure the acceleration on the mock-up railway system, and the MATLAB health monitoring codes computes the correlation function to get the natural frequency and mode shape. The natural frequency and mode shape are used to determine the structural health because the mode shape is an indicator of the structural properties. We also install the PZT sensors on three areas of the mock-up railway in order to monitor the state of fastening bolts on the bridge and to determine the degree of damage.

The sampling rate of the accelerometers and the PZT sensors is 250 Hz for each sensor. Thus, the transducer interface agent collects data at a rate of 250/sec from the accelerometers and PZT sensors and sends them to the agent IO layer of the middleware. We demonstrate the effectiveness of the template files by supporting these transducers without modifying the TIA. Because accelerometers and PZT sensors are different in data type, we define their data types respectively in Figs. 6, 7.

Fig. 6 An example of XML definition of configuration template files for accelerometers

```
<Template Idx="3">
< Type="UInt8" Fixed="1">Sensor_ID</Type>
< Type="Float32" Fixed="1">ValX </Type>
< Type="Float32" Fixed="1">ValY </Type>
< Type="Float32" Fixed="1">ValZ </Type>
</Template>
```

Fig. 7 An example of XML definition of configuration template files for PZT sensors

```
<Template Idx="4">
< Type="UInt8" Fixed="1">Sensor_ID</Type>
< Type="UInt8" Fixed="1">Count </Type>
< Type="Float32" Fixed="1">Value </Type>
</Template>
```

Configuration template index #3 applies to the accelerometers and index #4 applies to the PZT sensors. Figures 6, 7 show the XML definition of configuration template index #3 for accelerometers and index #4 for PZT sensors respectively.

We use a total of three non-standardized accelerometers and three PZT sensors to obtain vibration data. Because the sampling rate was 250 Hz for each sensor, a total of 1,500 pieces of sensor data for one second are transmitted to the WSN middleware by way of the agent.

4 Conclusions

We propose the concept of a transducer interface agent(TIA) for wireless sensor networks that is capable of supporting the various non-standardized sensors and actuators in the market. The agent can also serve as a replacement for the NCAP and the WTIM when it complies with the IEEE 1451.5 standard. The configuration template files connected to the agent work as TEDS in the family of IEEE 1451 standards. Providing many configuration template files permits the agent to deal with many different non-standardized transducers. We verified the usefulness of the TIA and the configuration template files through successful experiments receiving data from a mock-up railway rigged with three accelerometers, three PZT sensors, and DC motors as transducers.

Our future works will focus on refining the configuration template files to exactly follow the forms of TEDS in the IEEE 1451 standards. This will make a software defined IEEE 1451 standards like a software defined radio in a radio communication system [8].

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2010-0024242)

References

1. Lee JH (2009) Design of the middleware for wireless sensor node based on IEEE1451.5. In: 2009 ICROS-SICE international Joint conference (ICROS-SICE 2009), pp 1972–1975
2. Song EY, Lee KY (2006) Smart transducer web services based on the IEEE 1451.0 standard. In: 2006 IEEE sensors and applications symposium (SAS 2006), pp 72–77
3. Lee K, Song EY (2007) Wireless sensor network based on IEEE 1451.0 and IEEE 1451.5-802.11. In: The eighth international conference on electronic measurement and instruments (ICEMI 2007), pp 7–11
4. Song EY, Lee K (2007) An implementation of the proposed IEEE 1451.0 and 1451.5 Standards. In: 2007 instrumentation and measurement technology conference (IMTC 2007), IEEE (2007), pp 1–6
5. Song EY, Lee K (2008) STWS: a unified web service for IEEE 1451 smart transducers. IEEE Trans Instrum Meas 57:1749–1756
6. Sadok EF, Liscano M (2005) A web-services framework for 1451 sensor networks. In: 2005 instrumentation and measurement technology conference (IMTC 2005), IEEE (2005), pp 554–559
7. Guevara J, Vargas E, Brunetti F, Barrero F, Aranda L (2011) A framework for WSN using TinyOS and IEEE 1451 standard. In: 2011 IEEE latin-American conference on communications, pp 1–5
8. Ishikawa H (2012) Software radio technology for highly reliable wireless communications. Wirel Pers Commun 64(3):461–472
9. Viegas V, Pereira JMD, Girao PS (2008) Next generation application processor based on the IEEE 1451.1 standard and web services. In: International instrumentation and measurement technology conference (I2MTC 2008), IEEE (2008)
10. Viegas V, Pereira JMD, Girao PS (2007) NET framework and web services: a profit combination to implement and enhance the IEEE 1451.1 standard. IEEE Trans Instrum Meas 56:2739–2747
11. Viegas V, Pereira JMD, Girao PS (2006) IEEE 1451.1 standard and XML web services: a powerful combination to build distributed measurement and control systems. In: 2006 instrumentation and measurement technology conference (IMTC 2006), IEEE (2006), pp 2373–2377
12. Sollae Systems. <http://www.sollae.co.kr/en/products/>

Collaboration of Thin-Thick Clients for Optimizing Data Distribution and Resource Allocation in Cloud Computing

Pham Phuoc Hung and Eui-Nam Huh

Abstract Mobile cloud computing is growing rapidly because its device (i.e., smart phone) is becoming one of the main processing devices for users nowadays. Due to the limitation of resources such as battery life time, CPU and memory capacity, etc., a mobile device cannot satisfy some applications which usually demand more resources than it can afford. To alleviate this, the mobile device should collaborate with external resources to increase its capacity. In order to address these problems, we introduce a collaboration of thin-thick clients which enhances thin client capacities. We further propose a strategy to optimize the data distribution, especially big data in cloud computing. Moreover, we present an algorithm to allocate resources to meet service level agreement (SLA) and conduct simulations to evaluate our approach. Our results evaluation shows that our approach can improve resource allocation efficiency and has better performance than existing approaches.

Keywords Cloud computing · Data distribution · Thin · Thick client · Resource allocation

P. P. Hung · E.-N. Huh (✉)

Department of Computer Engineering, Kyung Hee University, Suwon, South Korea
e-mail: johnhuh@khu.ac.kr

P. P. Hung

e-mail: hung205a2@yahoo.com

1 Introduction

In recent years, mobile devices (thin clients) are more and more popular. As stated by the latest report, there were 5.6 billion mobile devices, a number that is expected to grow more [1]. Along with it, mobile service is overwhelming, including various areas such as Mobile Application, Mobile Commerce, Mobile Healthcare, Mobile Computing, etc. However, most of them still lack in resources such as battery life time, CPU and memory capacity, etc., to process heavy computing or higher data transferring. One of the ways to alleviate this issue is by using mobile cloud computing [2] or collaborating with external devices to get more resources. That can be released with minimal management effort or service provider interaction.

Currently, there are several researches dealing with this problem but they still have some limitation. For example, in [3], Gonzalo Huerta-Canepa presents guidelines for a framework to create virtual mobile cloud computing providers. This framework uses thin clients in the vicinity of users although thin clients have some capacity restriction and the bandwidth for thin clients and cloud is low. Therefore, thin clients had better connect with other more powerful processing devices (thick clients) like personal computers, laptops... which have more capacities and much higher bandwidth for thick clients. Good bandwidth is very important, the higher bandwidth we have the higher quality of services we get [4].

Thus, in this paper, we propose approach that is a collaboration of thin-thick clients which enhances thin client capacities. We further propose a strategy to optimize the data distribution, especially big data in cloud computing. Moreover, we present an algorithm to allocate resources to meet service level agreement (SLA) and conduct simulations to evaluate our approach. From our result evaluation, we see that our approach can improve resource allocation efficiency and has better performance than other existing approaches.

The paper is organized in following way. In [Sect. 2](#) related work is discussed. [Section 3](#) illustrates the motivating scenario. System architecture is presented in [Sect. 4](#). [Section 5](#) discusses our system implementation and analysis and the paper will end with conclusion.

2 Related Work

There have been a numerous studies which tried to solve some parts of above problems. In [5], authors propose a new approach for efficient cloud- based synchronization of an arbitrary number of distributed file system hierarchies. They use master slave architecture for propagation data to reduce traffic. In [6], researchers demonstrates some resource scheduling techniques can be effective in mitigating the impacts that cause a negatively impact to application response time and system utilization. Andreolini [7] and Fan [8] introduces the impact of the data transfer

delay on the performance but they do not reckon to use bandwidth efficiently. Gueyoung Jung, Nathan [9] presents a way to make a parallel process to a big data to increase performance in Federated Clouds but they do not consider how much resources should be used. For resource allocation, Ye Hu [10] shows that SA is superior to RA but the author do not conduct experiment with an arbitrary number of SLAs. In [11], Lenk provides service to a large amount of SLAs but performance difference between SA and RA is difficult to obtain.

Similar with our approach, other research efforts have made to integrate between mobile devices and cloud computing. In [12], Luo suggests an idea of using cloud to improve mobile device's capability. Marinelly [13] innovates Hyrax, which allows mobile devices to use cloud computing platforms. The researcher introduces the idea of using mobile devices as resource providers, but experiment is not integrated.

3 Motivating Scenario

The following scenario reflects the benefits of collaboration of thin-thick clients.

“A woman goes to a restaurant. She takes a picture of her favorite food by her mobile phone and notes the food. At home, she wants to cook that food. But her hand is sweat at that time. So, it is not convenient to use her mobile phone. She uses the refrigerator alternatively to access her mobile phone to get the picture then connect to the internet to search the way of cooking the food. Unfortunately, the bandwidth of the refrigerator to connect to the internet directly is too weak. Instead she connects the refrigerator to her personal computer then request the computer to access the internet to search the information she needs. After getting information, the computer returns the results to the refrigerator. Now she can see the information she have found on the refrigerator screen”. This scenario can be described through the left of Fig. 1 and each part of it corresponds to a numerated section in the sequence chart on the right of the figure.

The scenario shows the potential benefit of collaboration of thin-thick clients in cloud computing. Collaboration increases the opportunity of using resources efficiently. However, if bandwidth is too weak but the search result is too heavy, the issues here are how to optimize the data distribution and increase the efficiency of using resources? How to allocate resource to satisfy variety of SLAs?

4 System Architecture

In this section, we will describe our system architecture to solve above problems. Figure 2 illustrates the general design of the proposed system. Different with other approaches that follow a 1/m/1 model, our system can be viewed as a 1/m/m/1 model. It means that from the source, data is divided into multi blocks which are

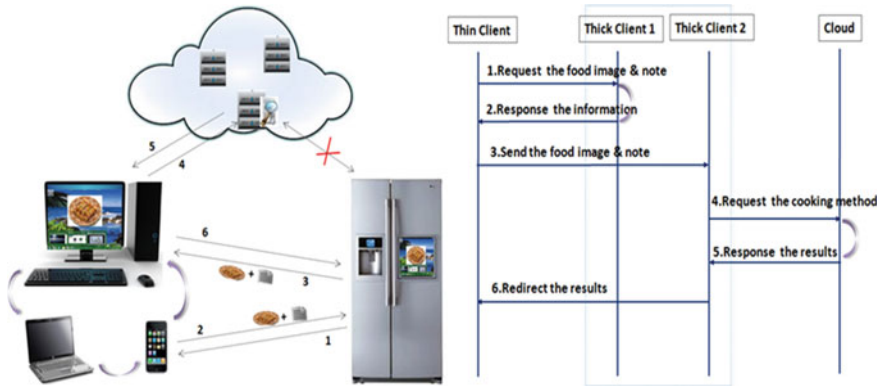


Fig. 1 Scenario of getting cooking method

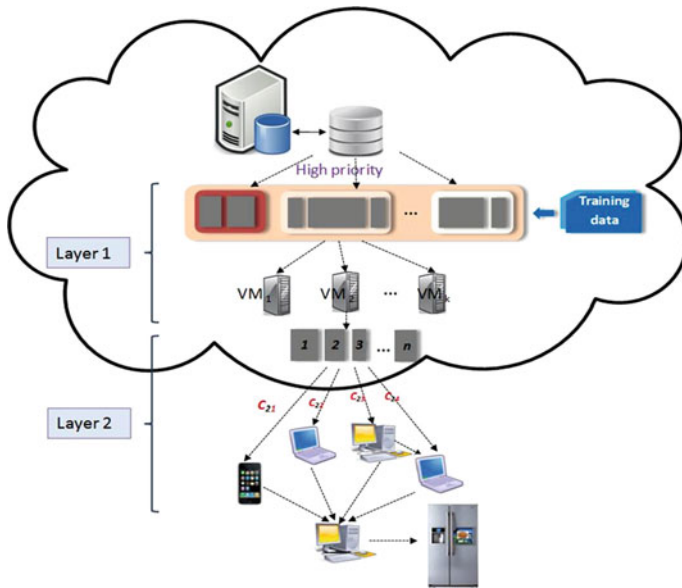


Fig. 2 System architecture

assigned to VMs; each block is divided into multi chunks and those chunks are transferred to multi processors. After receiving the data, processors combine the data into one then send it to a destination. The system consists of the following 2 layers:

Layer 1: In this layer, it involves: (a) Determining the minimum number of virtual machines (VMs) can meet the SLAs of cloud customers; (b) classifying, splitting and assigning data according to VM capacities.

- (a) Determining the minimum number of VMs: The algorithm 1 intends to determine the minimum number of VMs depending on SLA. In this algorithm, the cumulative distribution function (CDF) $F(x)$ of response time is available in [14] and the minimum number of VMs m increases until $F(x)$ reaches the target probability. At that time, we get m required for SLA. $F(x)$ is described as following:

$$F(x) = \text{Probability}(\text{response time} < x) = \begin{cases} 1 - e^{-\mu x} - k\mu e^{-\mu x} & \text{for } \sigma = m_i - 1 \\ 1 - e^{-\mu x} - k\mu e^{-\mu x(m_i - \sigma)} \left[\frac{1 - e^{-\mu x(1 - m_i + \sigma)}}{1 - m_i + \sigma} \right] & \text{for } \sigma \neq m_i - 1 \end{cases}$$

where

$$\begin{aligned} \sigma &= \lambda / \mu \\ k &= P(O) \frac{\sigma^{m_i} - \mu}{m_i!} * \frac{m_i}{(m_i - \sigma)} \\ P(O) &= \left(\sum_{n=0}^{m-1} \frac{\sigma^n}{n!} + \frac{m\mu^m}{m!(m - \sigma)} \right)^{-1} \end{aligned}$$

λ is the arrival rate and μ is the service rate

Algorithm 1 determine minimum of VMs

Input: λ //Arrival rate ; μ : Service rate
 $SLA(x, z)$ // x : response time; z : target probability
Output: m // minimum number of VMs required
float $\sigma = \lambda / \mu$
function determineMinVM(σ, μ, x, z) {
 if ($\sigma == (int) \sigma$) $m = (int) \sigma$;
 else $m = (int) \text{Math.floor}(\sigma) + 1$;
 while $F(x) <= z, m++$;
 return m ;
}

In general, a cloud computing infrastructure may provide services to a large number of SLAs with FCFS scheduling. Thus, we should allocate the VMs into two groups. One is Shared Allocation (SA) m_{shared} , the other is Reserved Allocation (RA) $m_{reserved}$. For SA, arrival jobs (SLA) are combined into a single stream and served by m VMs. For RA, each arrival job has its own dedicated VMs. This is illustrated in Fig. 3.

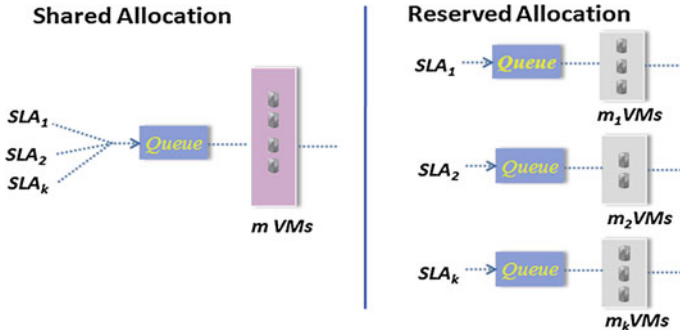


Fig. 3 Resource allocation strategy

With Shared Allocation, all SLAs have the same CDF of response time and arrival rate $\lambda = \sum_{i=1}^n \lambda_i$. Therefore, the minimum number of VMs m_{Shared} to meet k SLAs is given by: $m_{Shared} = \max(m_1, \dots, m_i, \dots, m_k)$ where m_i is the number of VMs required for SLA_i of $user_i$. Let $m_{Reserved}$ be the smallest number of VMs required to meet k SLAs in Reserved Allocation. So $m_{Reserved}$ is given by:

$$m_{Reserved} = \sum_{i=1}^k m_i$$

(b) Classifying data and splitting data according VM capacity: We use training data to classify the data in order to set priority for the data. The one that has a high priority can be transferred first and vice versa. This data may be divided into blocks $\{bl_1, bl_2, \dots, bl_n\}$ with different sizes. Then the best VMs are selected according to their capacity using the Greedy algorithm. At last, we assign big blocks to VM which has higher capacity.

Layer 2: In this layer we consider; (a) how to distribute data to processors with different capacities and (b) combine the data then transfer it to a thin client.

(a) Distributing blocks of data to processors with different capacities: For clarity, we give the important definition and assumption for our system. First, split each block into chunks $\{ch_1, ch_2, \dots, ch_n\}$ with different sizes depending on bandwidth. $w(ch_i)$ is the size of a chunk ch_i . b_i is bandwidth from a VM to a processor. Therefore, time spent for the transferring a chunk ch_i from VMs to a processor is $w(ch_i)/b_i$. For parallelization, the time to transfer chunks to processors should be equal.

$$\frac{w(ch_1)}{b_1} = \frac{w(ch_2)}{b_2} = \dots = \frac{w(ch_i)}{b_i} = t$$

$$\text{Set } S = w(\text{block}) = \sum_{i=0}^n w(ch_i) = t \sum_{i=0}^n b_i$$

$$\text{Thus, } w(ch_i) = t * b_i = \frac{S}{\sum_{i=0}^n b_i} * b_i$$

According to this value, we can determine the size of each chunk to adapt with bandwidth. Next step is to sort the processors $\{p_1, p_2, \dots, p_p\}$ depending on their capacities. The processor which has higher capacity will receive big chunks and the processor which has lower capacity will receive small chunks.

- (b) Combining the data then transfer it to a thin client: After receiving data from cloud service, instead of peer-to-peer synchronization between all processors which make communication more complex, we consider that a processor acts as a master which receive data from others processors to decrease the complexity due to firewall between processors.

5 Implementation and Analysis

This section will demonstrate the implementation result of our system. The result shows that SA and RA have almost the same impacts when they meet the same SLA with different arrival rate λ or different response time x or different target probability y , respectively, except the last case where we consider SA and RA that have to meet larger than one SLA. Through intensive testing, it was found that the minimum number of VMs SA required is smaller than the one RA required. The implementation results are presented in Fig. 4.

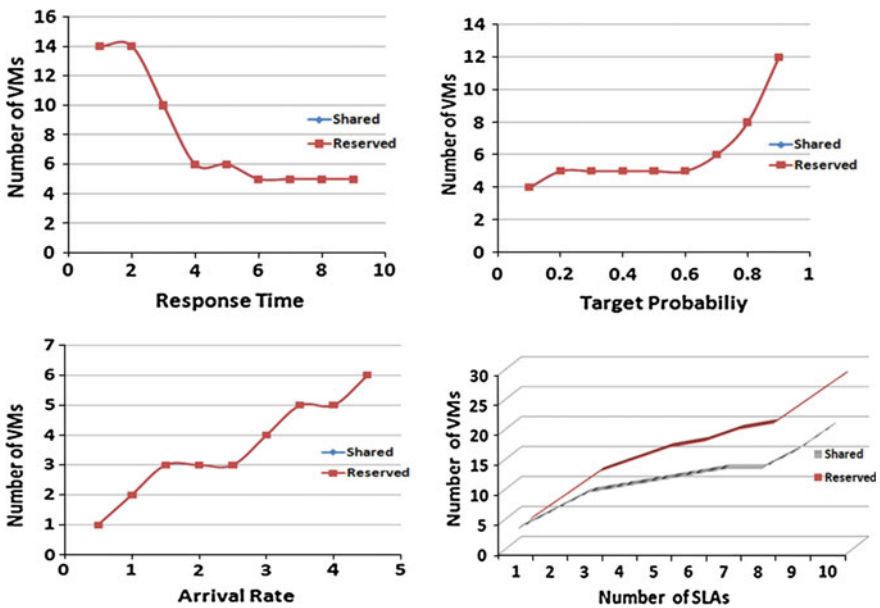


Fig. 4 SA and RA with different cases

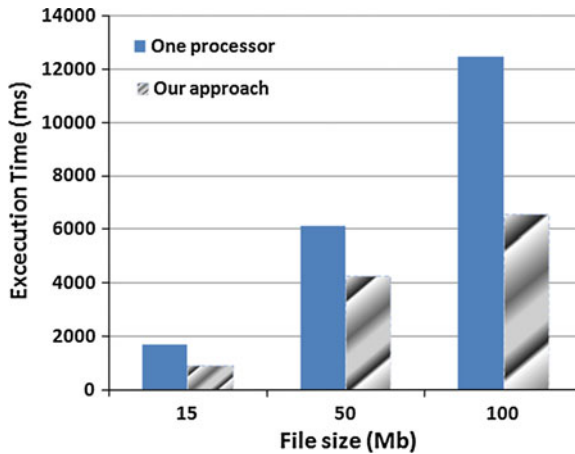


Fig. 5 Comparison our approach with other approach

In addition, we further compare the processing time of transferring a big data from a source to a destination of our system and other approach. Through experiment result in Fig. 5, we have seen that our approach has a better performance.

6 Conclusion

In this paper, we presented a system which is a collaboration of thin-thick clients to enhance thin client capacities. We also proposed a strategy to optimize the data distribution, especially big data in cloud computing. Moreover, we presented an algorithm to allocate resources to meet service level agreement and conducted simulations to evaluate our approach. Through implementation of our approach, we have seen that our approach can improve resource allocation efficiency and has better performance than existing approaches.

Acknowledgments This research was supported by Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0006418). The corresponding author is Eui-Nam Huh.

References

1. Wikipedia (2012) Mobile phone—Wikipedia, the free encyclopedia http://en.wikipedia.org/wiki/Mobile_phone. Accessed Sept 2012
2. Kumar K, Yung-Hsiang L (2010) Cloud computing for mobile users: can offloading computation save energy? *IEEE Comput* 43(4):51–56

3. Huerta-Canepa G (2010) A virtual cloud computing provider for mobile devices. In: MCS'10, San Francisco, 15 June 2010
4. Nguyen T-D (2012) Service image placement for thin client in mobile cloud computing. In: 2012 IEEE. doi:[10.1109/CLOUD.2012.39](https://doi.org/10.1109/CLOUD.2012.39)
5. Plastira Av N (2010) Cloud-based synchronization of distributed file system hierarchies. In: 2010 IEEE
6. Delgado J, Fong L (2011) Efficiency assessment of parallel workloads on virtualized resources. In: 2011 fourth IEEE international conference
7. Fan P (2011) Toward optimal deployment of communication-intensive cloud applications. In: Proceedings of international conference on cloud computing
8. Kwok M (2006) Performance analysis of distributed virtual environments. Ph D Thesis, University of Waterloo, Ontario
9. Nathan JG (2012) Synchronous parallel processing of big-data analytics services to optimize performance in federated clouds. In: 2012 IEEE, USA
10. Hu Y (2009) Resource provisioning for cloud computing. In: CASCON '09
11. Li J (2009) Fast scalable optimization to configure service systems having cost and quality of service constraint. In: IEEE 2009
12. Luo X (2009) From augmented reality to augmented computing: a look at cloud-mobile convergence. In: International symposium on ubiquitous virtual reality
13. Marinelli E (2009) Hyrax: cloud computing on mobile devices using MapReduce. Master Thesis Draft, Computer Science Department, CMU
14. Andreolini M (2008) Autonomic request management algorithms for geographically distributed internet-based systems. In: Proceedings of International Conference 2008

Mutual Exclusion Algorithm in Mobile Cellular Networks

Sung-Hoon Park and Yeong-Mok Kim

Abstract The mutual exclusion (MX) paradigm can be used as a building block in many practical problems such as group communication, atomic commitment and replicated data management where the exclusive use of an object might be useful. The problem has been widely studied in the research community since one reason for this wide interest is that many distributed protocols need a mutual exclusion protocol. However, despite its usefulness, to our knowledge there is no work that has been devoted to this problem in a mobile computing environment. In this paper, we describe a solution to the mutual exclusion problem from mobile computing systems. This solution is based on the token-based mutual exclusion algorithm.

1 Introduction

The wide use of small portable computers and the advances in wireless networking technologies have made mobile computing today a reality. There are different types of wireless media: cellular (analog and digital phones), wireless LAN, and unused portions of FM radio or satellite services. A mobile host can interact with the three different types of wireless networks at different point of time. Mobile systems are more often subject to environmental adversities which can cause loss of messages or data [1]. In particular, a mobile host can fail or disconnect from the rest of the network. Designing fault-tolerant distributed applications in such an

S.-H. Park · Y.-M. Kim (✉)
School of Electrical and Computer Engineering, Chungbuk National University Cheongju,
Chungbuk 361-763, Korea
e-mail: spark@chungbuk.ac.kr

environment is a complex endeavor. In recent years, several paradigms have been identified to simplify the design of fault-tolerant distributed applications in a conventional static system. Mutual exclusion, simply MX, is among the most noticeable, particularly since it is closely related to accessing shared resource called the critical section (CS) [2], which (among other uses) provides an exclusive access basis for implementing the critical section.

The mutual exclusion problem [3] requires two properties, safety and liveness, from a given set of processes. The problem has been widely studied in the research community [4–8] since one reason for this wide interest is that many distributed protocols need an mutual exclusion protocol. However, despite its usefulness, to our knowledge there is no work that has been devoted to this problem in a mobile computing environment.

The aim of this paper is to propose a solution to the mutual exclusion problem in a specific mobile computing environment. This solution is based on the token-based mutual exclusion algorithm that is a classical one for distributed systems. The rest of this paper is organized as follows: In Sect. 2, a solution to the mutual exclusion problem in a conventional synchronous system is presented. Section 3 describes the mobile system model we use. A protocol to solve the mutual exclusion problem in a mobile computing system is presented in Sect. 4. We conclude in Sect. 5.

2 Mutual Exclusion in a Static System

2.1 Model and Definitions

We consider a synchronous distributed system composed of a finite set of process $\Pi = \{p_1, p_2, \dots, p_n\}$ connected by a logical ring. Communication is by message passing, synchronous and reliable. A process fails by simply stopping the execution (*crashing*), and the failed process does not recover. A correct process is the one that does not crash. Synchrony means that there is a bound on communication delays or process relative speeds. Between any two processes there exist two unidirectional channels. Processes communicate by sending and receiving messages over these channels.

The mutual exclusion problem is specified as following two properties. One is for *safety* and the other is for *liveness*. The *safety* requirement asserts that any two processes connected the system should not have permission to use the critical section simultaneously. The *liveness* requirement asserts that every request for critical section is eventually granted. A mutual exclusion protocol is a protocol that generates runs that satisfy the mutual exclusion specification.

2.2 Token-Based Mutual Exclusion Algorithm

As a classic paper, the token-based mutual exclusion algorithm, which was published by Raynal, specifies the mutual exclusion problem for synchronous distributed systems with crash failures and gives an elegant algorithm for the system; this algorithm is called the token-based MX Algorithm [4]. The basic idea in the token-based MX algorithm is that the any process holding the token can use the critical section exclusively. The token-based MX algorithm is described as follows.

- A distributed system is connected by a logical ring. Each process has a unique ID that is known by its neighborhood processes.
- The CS is exclusively used by the process holding the token.
- The token is circulated on the logical ring. If a process wants to use the CS, then it just waits until receiving a token from its neighborhood. Only when it has received the token, it has a right to use the CS exclusively.

3 Mobile System Model

A distributed mobile system consists of two set of entities: a large number of mobile hosts (MH) and a set of fixed hosts, some of which act as mobile support stations (MSS_s) or base stations. The non MSS fixed hosts can be viewed as MSS_s whose cells are never visited by any mobile host. All fixed hosts and all communication paths connect them from the static network. Each MSS is able to communicate directly with mobile hosts located within its cell via a wireless medium. A cell is the geographical area covered by a MSS . A MH can directly communicate with a MSS (and vice versa) only if the MH is physically located within the cell serviced by the MSS . At any given instant of time, a MH can belong to one and only one cell. In order to send message to another MH that is not in the same cell, the source MH must contact its local MSS which forwards the messages to the local MSS of the target MH over the wireless network. The receiving MSS , in its turn, forwards the messages over the wireless network to the target MH . When a MH moves from one cell to another, a *Handoff procedure* is executed by the MSS_s of the two cells.

4 Mutual Exclusion in a Mobile System

In the following, we consider a broadcast group $G = (G_{MSS}, G_{MH})$ of communicating mobile hosts, where G_{MH} and G_{MSS} are respectively a set of m mobile hosts roaming in a geographical area (like a campus area) covered by a

fixed set of n MSS_s . In so far, local mobile hosts of base station MSS_i , which currently residing in MSS_i cell, will refer to mobile hosts that belong to group G .

A mobile host can move from one cell to another. If its current base station fails, the connection between the mobile host and the rest of system is broken. To recover its connection, a mobile host must move into another cell covered by an operational or correct base station. So, unless it crashes, a mobile host can always reconnect to the network. A mobile host may fail or voluntarily disconnect from the system. When a mobile host fails, its volatile state is lost. In this environment, the mutual exclusion problem is defined over the set G_MH of mobile hosts. When a mobile host h_k wants to use the CS , it sends the request message to a MSS . In this case, the mobile host eventually should get the permission from the MSS and use the CS .

4.1 Principle

The mutual exclusion protocol proposed in this paper is based on the solution described by Raynal in Token-based MX algorithm [4]. The outlines of their protocol have been described in Sect. 2. In this section, we give an overview of our protocol and identify the major differences compared with the original token-based MX algorithm. We assume that the mutual exclusion is initiated by a mobile host which requests its current base station a token to use the CS . The contacted base station saves the request into the queue until it receives the token from its neighborhood.

During the mutual exclusion, each base station on one hand interacts with the mobile hosts located in its cell to gather the request of each mobile host for CS and on the other hand interacts with the other neighboring base stations to send and receive a token. In our approach, a base station MSS which participates in the mutual exclusion protocol, always acts on behalf of a subset of mobile hosts. More precisely, the initial value of $Token_Holder_k$ is false but the value of it is changed true as a mobile host h_k that resides in MSS_i receives the token from its MSS_i . After returning the token to its base station, the mobile host h_k changes the value of its $Token_Holder_k$ into false again. The mutual exclusion protocol in such an environment consists of two cases depending on who the token holder is. As the first case, that is when a base station received a token from its neighboring base station or its mobile hosts. When it received the token from its neighboring base station, then it just sends the token to a mobile host with highest priority among the mobile hosts connected to the base station. In case of returning the token from its mobile hosts, it just sends the token to the next base station.

4.2 Protocol

The protocol is composed of three parts and each part contains a defined set of actions. Part A (Fig. 1) describes the role of an arbitrary mobile host h_k . Part B (Fig. 2) presents the protocol executed by a base station MSS_i . Part B is related to the interactions between a base station and its local mobile hosts on one hand and the other base station on the other hand. Thus, Part B is based on the traditional Token-based MX protocol adapted to our environment.

Finally, the part C of the protocol is the handoff protocol destined to handle mobility of hosts between different cells. In Fig. 1, the three actions performed by an arbitrary mobile host are:

- (1) A mobile host executes this action when it receives a request from an upper application program to initiate a mutual exclusion.
- (2) Token message is sent to a mobile host h_k by the mobile support systems MSS_i when it had requested a token from the local base station where it resides. Upon receipt of such a message, the mobile host gets into the *Critical Section*.
- (3) When the application program terminates the mutual exclusion protocol, the Token is released to the mobile support system, MSS_i .

Actions of the protocol in Fig. 2 numbered from (4) to (7) are executed a mobile support system, i.e., a base station MSS_i . They have the following meaning:

- (4) When a base station is asked by a mobile host to send a Token, it inserts the request into the rear of its queue.
- (5) In case of receiving a Token from other base station, the base station checks its queue My_Queue_i to confirm whether the queue is empty or not. If the queue is not empty, then the base station sends the Token to the mobile host that is positioned at the front of the queue. And it deletes the element from the queue and sets its status to true that means it holding Token, i.e., $My_Status_i := 1$. But if the queue is empty, then the base station just passes the Token to the next base station.
- (6) When a base station receives a Token from a mobile host h_k , it checks its queue and status. If both $(Phase_i = 0 \wedge My_Queue_i \neq \emptyset)$ are true, which means that it does not hold the token and at the same time the queue is not empty, then the base station sends the Token to the mobile host that is the front element of the queue. And it deletes the element from the queue and sets its

Fig. 1 Protocol executed by a mobile host h_k (part A)

```

% Mobile host  $h_k$  is located in  $MSS_i$  cell %
(1) Upon receipt of the request for CS from the application
    Send Req_Token to  $MSS_i$ 
(2) Upon receipt of Token from  $MSS_i$ 
% The mobile host ( $h_k$ ) gets into CS %
    CS ( $h_k$ )
(3) Upon receipt of the release for CS from the application
    Send Release_Token to  $MSS_i$ 
    
```

Fig. 2 Protocol executed by a mobile support station MSS_i (part B)

```

My_Statusi := 0;
My_Queuei := ∅;
Cobegin
(4) || Upon receipt of Req_Token( $h_k$ )
      insert Req_Token( $h_k$ ) to rear (My_Queuei);
(5) || Upon receipt of Token ( $MSS_{i-1}$ )
      if My_Queuei ≠ ∅ then
        My_Statusi := 1;
        send Token to front (My_Queuei);
        delete front (My_Queuei);
      else
        send Token to  $MSS_{i+1}$ ;
      end-if
(6) || Upon receipt of Token ( $h_k$ )
      if (Phasei = 0 ∧ My_Queuei ≠ ∅) then
        My_Statusi := 1;
        send Token to front (My_Queuei);
        delete front (My_Queuei);
      else
        My_Statusi := 0;
        send Token to  $MSS_{i+1}$ ;
      end-if
(7) || Upon receipt of Req_Token ( $MSS_j$ )
      insert Req_Token( $h_k$ ) to Rear(My_Queuei);

```

status to true. Otherwise it sends the Token to the next base station and sets its status to false.

- (7) On receiving the Token request message from other mobile support system, the MSS_i insert the request message into its queue. As shown in Fig. 3, the handoff protocol is described.
- (8) When a mobile host h_k moves from MSS_j cell to MSS_i cell, the handoff protocol execution is triggered. Mobile host h_k has to identify itself to its base station by sending a message GUEST(h_k, MSS_j).
- (9) Upon receiving this message, MSS_i learns that a new mobile host h_k , coming from MSS_j cell has entered in its cell. With BEGIN_HANDOFF(h_k, MSS_i)

Fig. 3 Handoff procedure (part C)

```

Cobegin
% Role of  $h_k$  %
(8) || Upon entry in  $MSS_i$  cell
      send Guest( $h_k, MSS_j$ ) to  $MSS_i$ 
% Role of  $MSS_i$ 
(9) || Upon receipt of GUEST( $h_k, MSS_j$ )
      Local_MHi := Local_MHi ∪ { $h_k$ };
      send BEGIN_HANDOFF( $h_k, MSS_j$ ) to  $MSS_j$ ;
% Role of  $MSS_j$ 
(10) || Upon receipt of BEGIN_HANFOFF( $h_k, MSS_j$ )
      Local_MHj := Local_MHj - { $h_k$ };
      if (Req_Token( $h_k$ ) ∈ My_Queuej) then
        send Req_Token( $h_k$ ) to  $MSS_i$ ;
        delete Req_Token( $h_k$ ) from My_Queuej;
      end-if

```


message, MSS_i informs MSS_j that it removes h_k from the set of mobile hosts that reside in its cell.

- (10) Upon receiving such a message, MSS_j checks its queue to confirm that the token request of h_k is in the queue. If it is in its queue, then it transfers the token request to MSS_i and deletes the token request from the queue.

4.3 Correctness Proof

As our protocol is based on the Token-based logical ring algorithm proposed by Raynal, some statements of lemmas and theorems that follow are similar to the ones encountered in [4].

Theorem 1 No two different processes can have permission to use the critical section simultaneously (safety property).

Proof (proof by contradiction). Let assume that there exist two mobile hosts to get a permission to use the critical section. A mobile host can use the CS only if it received a permission token from the MSS of the cell to which it belonging (action 2). In this case, the assumption means that there exist two MSS s holding the token or one MSS sends the token twice to two different mobile hosts each. The first case is false since there is only one token circulating under the logical ring. The second case is also false since the MSS holding the token sends it to mobile host h_k only once (action 5). So it is a contradiction. \square

Theorem 2 Every request for the critical section is eventually granted (liveness property).

Proof If a mobile host sends a message to request a token (action 1), at least one MSS eventually receives it and inserts it into the queue (action 4). After that, there are two cases. In first case, if the mobile host h_k sent the message does not move to other cell, then the message Req_Token eventually will be positioned at the front of the queue and the MSS received the message sends the token. Thus, the mobile host sent the message eventually receives the token and uses the CS. In a second case, when the mobile host h_k sent a message Req_Token moves from MSS_j cell to another MSS_i cell before receiving the token, then the handoff protocol execution is triggered (action 8–10). Mobile host h_k has to identify itself to its base station by sending a message GUEST(h_k, MSS_j). In this case, by (action 10) the request message will be transferred to the MSS of the cell to which the mobile host has moved. Consequently, the mobile host will receive the Token and use the CS when the MSS sends the Token. \square

5 Conclusion

The communication over wireless links are limited to a few messages (in the best case, three messages: one to request a token and the others to get the token and release the token respectively) and the consumption of mobile hosts CPU time is low since the actual mutual exclusion is run by the base stations. The protocol is then more energy efficient. The protocol is also independent from the overall number of mobile hosts and all needed data structures are managed by the base stations. Therefore, the protocol is scalable and can not be affected by mobile host failures.

In addition, other interesting characteristics of the protocol are as follows. (1) During the mutual exclusion period, a base station should keep track of every mobile host within its cell to manage the request messages and the token. (2) In such a mobile computing environment, a handoff algorithm is needed to perform mutual exclusions efficiently and correctly, but it is not needed in static distributed systems.

The mutual exclusion algorithm in a mobile computing environment consists of two important phases. One is a local mutual exclusion phase in which a mobile host holds and uses the CS. The other phase is a global mutual exclusion phase in which each *MSS* takes part in the mutual exclusion by passing the token to another *MSS*.

References

1. Pradhan DK, Krichna P, Vaidya NH (1996) Recoverable mobile environments: design and tradeoff analysis. FTCS-26
2. Lodha S, Kshemkalyani AD (2000) A fair distributed mutual exclusion algorithm. IEEE Trans Parallel Distrib Syst 11(6):537–549
3. Agrawal D, Abbadi AE (1991) An efficient and fault-tolerant solution for distributed mutual exclusion. ACM Trans Comput Syst 9(1):1.20
4. Raynal M (1986) Algorithms for mutual exclusion. MIT Press, Cambridge
5. Maekawa M (1985) A \sqrt{N} algorithm for mutual exclusion in decentralized systems. ACM Trans Comp Syst 3(2):145–159
6. Manivannan D, Singhal M (1994) An efficient fault-tolerant mutual exclusion algorithm for distributed systems. In: Proceedings of the ISCA international conference on parallel and distributed computing systems, pp 525–530
7. Vidyasankar K (2003) A simple group mutual *l*-exclusion algorithm. Inform Process Lett 85:79–85
8. Singhal M (1993) A taxonomy of distributed mutual exclusion. J Parallel Distrib Comput 18(1):94–101
9. Alagar S, Venkatesan S (1994) Causally ordered message delivery in mobile systems. In: Proceeding of workshop on mobile computing systems and applications, Santacruz
10. Badache N (1995) Mobility in distributed systems, technical report #962. IRISA, Rennes, Octor
11. Badrinath BR, Acharya A, Imielinski T (1993) Impact of mobility on distributed computations. ACM Oper Rev 27(2):15–20

An Effective Personalized Service Provision Scheme for Ubiquitous Computing Environment

Chung-Pyo Hong, Cheong-Ghil Kim and Shin-Dug Kim

Abstract In a ubiquitous computing environment, one basic parameter is whether all the components in any specific environment are able to connect with each other and users can utilize them at anytime, anywhere. To address above issues, we introduce an efficient resource management scheme based on polymorphism. In this paper, resource objects are presented in the unified form to cooperate each other based on a common interface defined in it. The unified form of resource objects are called as Virtual Object (VO). With the VO, we propose a mechanism to provide services to users. The proposed mechanism is based on the profiles, which represents the situational information of any certain user. In this paper, services will be transformed based on the varying profile and it is called polymorphic services.

Keywords Ubiquitous computing · Personal space · Polymorphism

C.-P. Hong (✉) · S.-D. Kim (✉)
Department of Engineering, Yonsei University, 5-4, Sinchon-dong,
Seodaemun-gu, Seoul, South Korea
e-mail: hulkboy@yonsei.ac.kr

S.-D. Kim
e-mail: sdkim@yonsei.ac.kr

C.-G. Kim (✉)
Department of Computer Science, Namseoul University, 21, Mae Ju-ri,
Seonghwan-eup, Seobuk-gu, Cheonan-si, Chungnam 331-707, South Korea
e-mail: cgkim@nsu.ac.kr

1 Introduction

The term ubiquitous computing means that the computer is not visible and would be integrated in our daily lives. So, users do not need to concern how the computer is operated as doing any specific process. In a ubiquitous computing environment, one basic parameter is whether all the components in any specific environment are able to connect with each other and users can utilize them at anytime, anywhere.

We have found two major issues which are needed to establish a service provision framework for the ubiquitous computing. One is the use of distributed resource objects and the application of various contexts. In this paper, resource objects are presented in the unified form to cooperate each other based on a common interface defined in it. The context is divided into several categories to express the situation and characteristics of each component related to a user. And also, the context is presented in the form of well-designed profiles. With these profiles and unified form of resource objects, we propose a mechanism to provide services to users. Thus the proposed mechanism is based on the profiles, which represents the situational information of any certain user, we begin our design process with the inventing a model of personal service space. And also, it transforms services in accordance with the context change for the better user experience.

The rest of this paper, we introduce the ubiquitous computing researches in [Chap. 2](#). In [Chap. 3](#), we describe out the concept of personal service space. [Chapter 4](#) explains several evaluation results. Conclusion is provided in [Chap. 5](#).

2 Related Work

In this chapter, we introduce some related work. Systems for providing ubiquitous computing are analyzed and present some drawbacks. Also, resource management scheme is introduced.

Location-based service is a basic service model [1–3]. It tracks user's location and displays location of the user. Then, It provides location-specific services to the user by using this information [2, 3]. As location tracking technologies are being advanced, we will apply them to various practical areas. However, these service models have many limitations, namely it is difficult to define proper location range to provide services and cannot concern many situations except specific user's location [4–6].

GAIA is one of the most famous ubiquitous computing frameworks [7]. It provides dynamic services based on profiles generated by environmental contexts, e.g., network information, user movement information, and environmental brightness. All the services presented by GAIA are based on the assumption that all the objects are connected and operational. The GAIA team also defines a personal active space, which is a cluster of devices that are connected to the user's

device via wireless communication. In their system, the user's device behaves as a coordinator. In [7], they suggest a scheme for handling devices around a user, but this scheme does not mention how to control remote devices. The paper simply assumes that every device needed for user services is located around a user. Another framework, Aura [8], provides an abstract characterization of services to search for appropriate matches among available resources. Aura thus overcomes the problem of heterogeneity. However, [8] does not provide a detailed explanation regarding how to support management of users' personal information and environment.

These studies address more user-centric services, but the common problem is that they do not overcome the limitations of the location-based service mechanism. So, services are not available if there are proper resources which can provide any specific function.

3 The Conceptual Model of Polymorphic Services

3.1 Personal Service Space

To realize the personal Personal service space (PSS) concept, we define detailed elements, which construct the PSS. Virtual object is the first one, which is required for our designed personal space model. Virtual Object (VO) is the software component that uniformly represents resource object participating ubiquitous computing. VO can be contained in a physical object such as display panel, printer, and any other objects, and also it can be a part of software component of any running application on the computing system. Thus, the VO is in charge of providing information regarding physical or software object and manages functions on the corresponding objects. And also, it provides unified interfaces among VOs.

With the functions provided by VO, services are defined. Services are application that performs operations for user intention. By utilizing the unit functions offered by VO, and organizing them together, user's task can be achieved. Thus, the organized form of VOs is a model of service. The relationship between VO and service is described in Fig. 1.

3.2 Service Provision Flow

The polymorphic service is designed based upon the basic concept of PSS, which is already proposed by my colleagues and me. We propose an overall operational flow. So, the conceptual flow of service provision can be simply described as shown in Fig. 2. In 2, the service is executed as analyzing the user profile. Based on the profile PSS select the appropriate services and propose it to the user. If any

Fig. 1 PSS service model

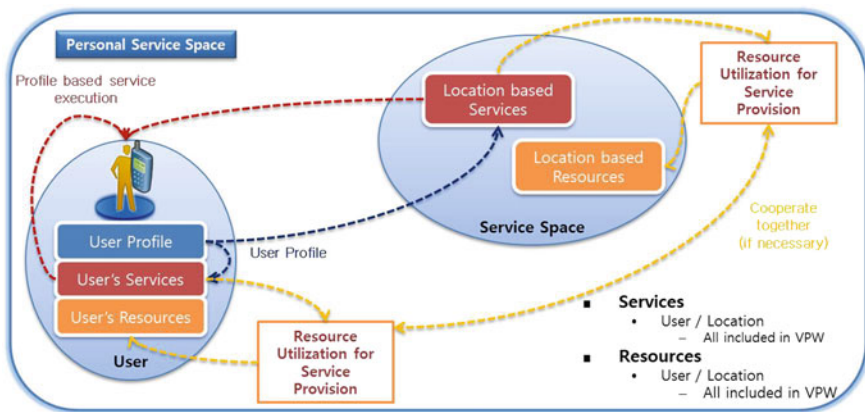
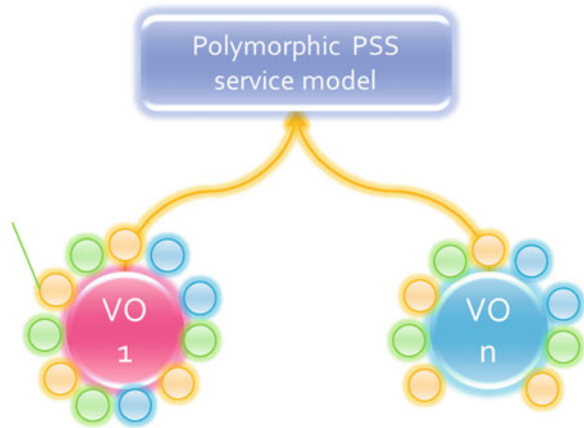


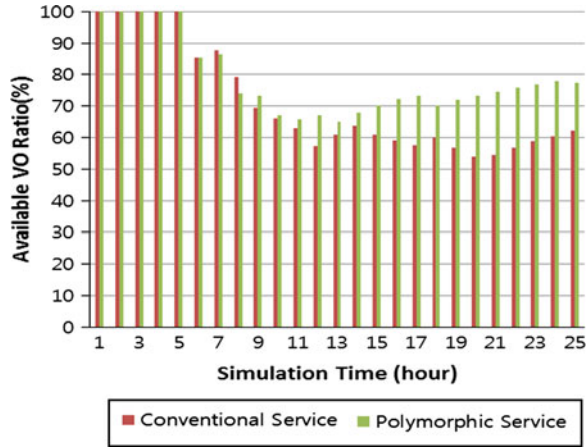
Fig. 2 Overall flow of proposed service provision platform

user decide to run a service, then the PSS engine choose proper VOs by the analysis result of the service description. The service description is a specification of a service, and it represents required VOs and service execution step. As running the services, the PSS examine any transition of user profile, which can be made by user's interaction. If there is any factor which can cause the environmental change for the running service, the PSS transforms the running service.

4 Evaluation

As service is running, the environment is changed continuously, the availability of required VOs changes. After a resource environment is changed and the required VOs are not satisfied, the proposed service model can change its service

Fig. 3 Ratio of available VOs to the required resource



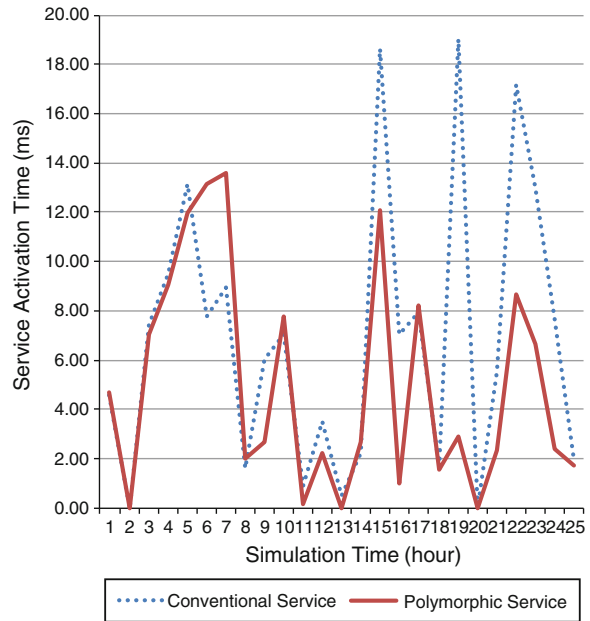
description or the VOs can be transformed into another form. With the given characteristics, the ratio of the available VOs to the required VOs can be increased and the service provision can be continued.

As shown in Fig. 3, the ratio of the available VOs to the required resource of the proposed scheme became higher at the 5th hour from the simulation beginning. Consequently, the ratio is grown up to be about 29 % by the polymorphic service framework.

When Service is activated, VOs, which are described in service description, is needed. So, the preparation time is required. In the conventional model of ubiquitous service, the service cannot be activated without the exact resources, but the proposed polymorphic service model can run any given services with the transformed VOs or modified service description. These aspects guarantee the smaller time to activate any services in comparison with the conventional model.

For the simulation, we assume that the basic delay time for the VO connection is from 0 to 3 ms. And also, the service activation delay is set as from 3 to 10 ms. Figure 4 shows the windowed activation time based on the elapsed simulation time. It takes less time in activation of the polymorphic model of user service than conventional service model because the possibility of required VO existence is higher in the proposed scheme. The difference is not significant in beginning part and after middle part it increases definitely. As the proposed scheme running, the set of VOs are extended. And the possibility of the VO transformation and the VO matching probability increases. Consequently, the windowed-sum of activation time decreases slowly.

Fig. 4 Service activation time



5 Conclusion

In this paper, we propose a mechanism to provide services to users with VO and well-designed profile. Thus the proposed mechanism is based on the profiles, which represents the situational information of any certain user, we begin our design process with the inventing a model of personal service space. And also, it transforms services in accordance with the context change for the better user experience.

As a result, the ratio of the available VOs to the required resource of the proposed scheme became higher at the 5th hour from the simulation beginning. Consequently, the ratio is grown up to be about 29 % by the polymorphic service framework. And also, it takes less time in activation of the polymorphic model of user service than conventional service model because the possibility of required VO existence is higher in the proposed scheme.

References

1. Rajkumar R, Lee C, Lehoczy J, Siewiorek D (1997) A resource allocation model for QoS management, Real-time systems symposium, proceedings, The 18th IEEE 2–5 Dec 1997 pp 298–307
2. Mohapatra D, Suma SB (2005) Survey of location based wireless services, personal wireless communications, 2005. ICPWC 2005. 2005 IEEE international conference on 23–25 Jan 2005 pp 358–362

3. Schilke SW, Bleimann U, Furnell SM, Phippen AD (2004) Multi-dimensional-personalisation for location and interest-based recommendation. *Internet Res Electr Netw Appl Policy* 14(5):379
4. Scholtz J, Consolvo S (2004) Toward a discipline for evaluating ubiquitous computing application. Technical report, intel research
5. Mori G, Fabio P, Santoro C (2004) Design and development of multi device user interfaces through multiple logical descriptions. *IEEE Trans Softw Eng* 30(8):507–520
6. Ballagas R, Borchers J, Rohs M, Sheridan JG (2006) The smart phone: a ubiquitous input device. *Pervasive Comput IEEE* 5(1):70–77
7. Roman M, Hess C, Cerqueira R, Ranganathan A, Campbell RH, Nahrstedt K (2002) Gaia: a middleware infrastructure to enable active spaces. *IEEE Pervasive Comput Mag* 1(4):74–82
8. Sousa JP, Garlan D (2002) Aura: an architectural framework for user mobility in ubiquitous computing environments. In: *Proceeding of the 3rd workshop IEEE/IFIP conference on software architecture*, pp 29–43

Fast Data Acquisition with Mobile Device in Digital Crime

Chang-Woo Song, Jeong-Hyun Lim, Kyung-Yong Chung,
Ki-Wook Rim and Jung-Hyun Lee

Abstract As mass storage media is recently becoming more common due to spread of smart phones in which new technologies have been applied, data collected from digital crime has been increased a lot. At the time, if the investigator did not conduct the initial response properly, we may lose the value as the evidence. Thus, the collection of digital evidence in a short time is required. Therefore, this paper proposes the methods to collect data rapidly at the scene of the crime based on table related to criminal charges. Implemented application can collect data with a consideration of each feature of software and provide the rapid results through a pattern search.

Keywords Mobile device · Forensic · Evidence collection · Data acquisition

C.-W. Song · J.-H. Lee
Department of Computer Science and Engineering, Inha University,
Yonghyeon, 1,4-dong, Nam-gu, Incheon, Korea
e-mail: ph.d.scw@gmail.com

J.-H. Lee
e-mail: jhlee@inha.ac.kr

J.-H. Lim
School of International Affairs and Information, Dongguk University,
26, Pil-dong, Jung-gu, Seoul, Korea
e-mail: rockq81@dongguk.edu

K.-Y. Chung (✉)
School of Computer Information Engineering, Sangji University,
83 Sangjidae-gil, Wonju-si, Gangwon-do, Korea
e-mail: dragonhci@hanmail.net

K.-W. Rim
Department of Computer Science and Engineering, Sunmoon University,
Galsan-ri, Tangjeong-myeon, Asan-si, Chungcheongnam-do, Korea
e-mail: rim@sunmoon.ac.kr

1 Introduction

When the incident has occurred, we are often dispatched to the scene. The initial activity made in the scene is very important. The traditional forensics preserves analog evidence such as fingerprints, bloodstains and footprints without being compromised by external impacts using police line. According to “Locard’s exchange principle”, it can be determined whether the case is solved by how good the scene is preserved. Criminal and in general illegal activities are no exception [1]. It is a big key to how you could collect digital evidence in the initial crime scene. Mobile devices are rooted in their own operating systems, file systems, file formats and methods of communication. Dealing with these devices creates unique problems for examiners [2]. Performing a forensic exam on a mobile device requires special software and special knowledge of the way these devices work, as well as where possible evidence could be stored [3, 4].

This paper proposes the method to rapidly collect the digital evidence from mobile devices by using tables related to criminal charges and examines it through actual implementation. The rest of this paper is organized as follows. [Section 2](#) describes the data acquisition method for movie device. [Section 3](#) describes the fast data acquisition with mobile device in digital crime. The conclusions are given in [Sect. 4](#).

2 Data Acquisition Method for Mobile Device

Supply of various mobile devices makes it difficult for detectives to collect the information. Typically, the OS systems of iPhone and Android have different data storage structures. Also, since most of the mobile device information is modified frequently, the verification of overlapping of collected data is required. Therefore, most of all, since the immediate data collection is very important at the crime scene, the collecting method fits to each of mobile device property is required [5, 6].

2.1 *Collecting Information in Android OS*

Contents Provider is an application data sharing interface. Applications can store the data to or bring the data from other applications using Contents Provider. Also, each application can make its own Contents Provider and then open it to outside. During this process, data information can be loaded using two different actions such as “action_pick” and “action_get_content”. The difference between those two actions is that “action_pick” uses the actual data address and “action_get_content” uses the type of data. Generally, the data being used in various applications uses

“action_get_content”, and the data being used in various applications such as images and videos brings them by finding out the corresponding addresses using “action_pick”.

Rooting means the method to get the authority of administrator by attacking the weak points of Android operating system. Android OS was developed based on the Linux and uses the same user’s authority system. Though the “root” means the authority of administrator in Linux, and it is existed in Android OS, it doesn’t include the “su” command which can be elevated using this authority. Therefore, rooting means the insert of the “su” command after obtaining the authority of administrator temporarily through the attack of weak point. Authority of administrator can be obtained anytime afterward using the inserted “su” command, and the information will be available through the following operation. Dumpsys provides the detailed information of service, memory and system information as follows. The verification of the details of Time stamp or the boot-up time recorded in Kernel log is available (if it was not booted recently, it may not be recognized). If it is long, it can be verified by using pipeline (|) or redirection (>). Data can be measured in the crime scene and can be prioritized during the storage. For your reference, one of the modifications of Android 4.0 is that the back-up and restoration of applications is available without rooting.

2.2 Collecting Information in iOS

Using Jailbreak, the Firmware version of iPhone can be analyzed. This seems not enough to be a forensic method that can be used as evidence in the court but is suitable to be used with the purpose for the confession of suspect and internal/external audit. This can be the best alternative in securing the critical evidence and it also is partly accepted legally. If Jailbreak is applied on the iPhone that was secured as an evidence, the internal data can be collected by approaching to the file system using the programs such as iFunbox [4], DiskAid [5] etc.

There is the method that can obtain the logical copy using iTunes Sync program in the computer of suspect. Though it has advantage with the speed, it is difficult to restore the deleted files or data, and the suspect’s computer needs to be secured. Since this is non-jailbreaking type system, the data such as intentionally deleted files or slack space of folders cannot be found. Back-up data does not have tree structure by each application program and all of the back-up subject files will be stored in one folder. Therefore, each data needs to be distinguished by analyzing the Meta files of back-up data per each of the iOS version.

3 Forensic Evidence Collection in Mobile Device

Packing the Evidential Matter: The apparatus in question shall be packed together with a data cable, driver for connecting to computer, battery charger, external memory etc. using the frequency isolating envelop and device and sealed with the sealing paper. Date of confiscation, name of executing institution, personal data and information of the apparatus (Model name, Serial Number, etc.) shall be recorded on the label of evidential matter [7].

Completion of Evidence Collection and Transfer: The collector and transferor shall transfer evidential matters after checking if collected evidences are in accordance with the list. Evidential matters shall be safely packed using anti-electricity protection film, shock buffering materials or anti-shock protective box etc. [8, 9].

Preparation of Result Report of Evidence Collection: A result report of the evidence collection shall be prepared containing the execution institution, outline of accident, date of collection, personal data, list of evidential matters and procedure of collection after completion of evidence collection (Table 1).

Table 1 Type of criminal and digital evidential matter

Contents of talk	Details of receipt and transmission of talks and absences of the user of Smart phone before confiscation could be seen and an indirect inference could be made for the activities in question
Contact list	Images of the contact list could be identified with Smart phone, and data which may find persons surrounding the offender and the conspirator using the account synchronization of SNS could be presented and social position and relation of the offender could be inferred
Short message/ Email	They may provide information such as criminal intent of the user and circumstances
Media (video, audio)	Media files may provide information such as date and time and may provide a location clue of the image shot by storing GPS coordinate which is one of functions of Smart phone
Record of web search	It may provide a clue related to the accident utilizing the websites which the user of the mobile device enjoys or web search log
Chatting log	Several chatting applications could be installed in mobile device and criminal intent and clue could be inferred through these chatting logs
Social network	A potential criminal clue could be inferred by understanding writings of the user, images, videos posted in Social Network and relation with surrounding persons.
Schedule table, note	A criminal clue could be inferred on the basis of past, present and future plan of the user based on the schedule table and memos etc
Network connection	Internet ID or nickname of the user could be inferred through SSID of Wi-Fi connection etc
Map service	A potential evidential matter could be inferred with log data of places where the user used to visit and search frequently
Other installation app.	A potential evidential matter such as character and hobby of the user on the basis of account and applications which the user used frequently

4 Implementation of Fast Data Acquisition with Mobile Device

Figure 1 shows the fast data acquisition with mobile device in digital crime. After implemented application is connected to smart phones, if dump image is extracted and analyzed, items are shown on the screen as shown in Fig. 1. Pre-analysis is carried out based on table related to criminal charge. These methods can reduce the entire analysis time. Data such as smart phone basic information, a list of contacts, call list/log, SMS/MMS, memos/events and traces of using the Internet (URL, cookies, and bookmarks) can be found easily using the pattern search. Among them, the most time-consuming task is the physical memory dump. In spite of the most time, the reason why the physical memory is placed in the front is that all information collected from the rear eventually comes from the physical memory. If the physical memory dump is located in the rear, the trace of physical memory can be damaged while collection command is running [10].

Since the inactive data can be originally maintained after the power is turned off, it is common to analyze it after storage media imaging is completed. However, as mass storage media is recently becoming common, it is not suitable for rapid incident response to wait until the storage medical imaging is completed. Among

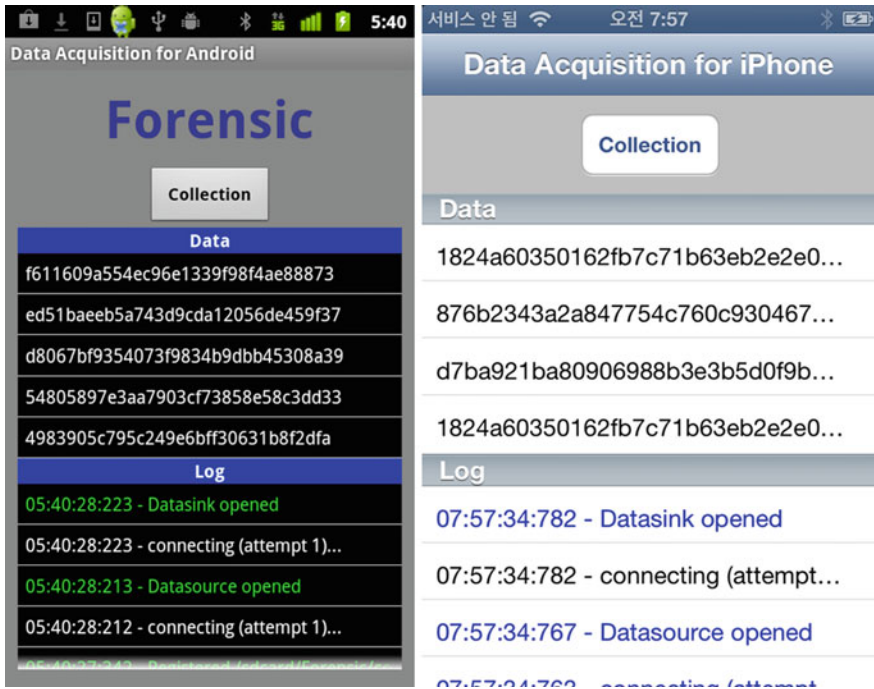


Fig. 1 Fast data acquisition with mobile device

inactive data, there are some data which can be utilized very importantly for analysis even if the size of data is not large. If these data are collected together when live forensics are carried out, pre-analysis can be performed while the storage media imaging is being carried out. If the system is not in active state, the entire analysis time can be significantly reduced if the major inactive data are collected at first in the state where the write-protector is installed before imaging.

All commands should be run independently. Since the basic command or library could be infected by malignant code, command statically compiled in advance and brought here should be used. Since the version of target operating system can be various, you should prepare a list of appropriate commands after it is examined whether it works properly by testing it based on version of each operating system in advance.

5 Conclusions

In this paper, we proposed the fast data acquisition with mobile device in digital crime. A pattern search based on table related to criminal charges has advantage to immediately provide investigators with necessary information more rapidly. However, since a pattern search has different methods to import data and print it out based on features of smart phones, the disadvantages should be supplemented by using a variety of methods. However, actions to collect evidences using all possible methods are not good from the perspective of collection time or evidence damaging. Therefore, table related to criminal charges proposed in this paper must be constantly updated and be supplemented to perform a variety of pattern searches.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (No. 2012-0004478).

References

1. Kubi AK, Saleem S, Popov O (2011) Evaluation of some tools for extracting E-evidence from mobile devices. In: Proceedings of the international conference on application of information and communication technologies, pp 1–6
2. Bertè R, Dellutri F, Grillo A, Lentini A, Me G, Ottaviani V (2009) Fast Smart phones forensic analysis results through MIAT and forensic farm. *Int J Electron Secur Digit Forensics*, Indersci
3. Lin IL, Chao HC, Peng SH (2001) Research of digital evidence forensics standard operating procedure with comparison and analysis based on smart phone. In: Proceedings of the international conference on broadband and wireless computing, communication and applications, pp 386–391

4. Akkaladevi S, Keesara H, Luo X (2011) Efficient forensic tools for handheld device: a comprehensive perspective. *Softw Eng Res Manage Appl Stud Comput Intell* 377:349–359
5. Andrew H (2011) *Android forensics: investigation, analysis and mobile security for Google Android*. Syngress
6. Andrew H, Katie S (2011) *iPhone and iOS forensics: investigation, analysis and mobile security for Apple iPhone, iPad, and iOS Devices*. Syngress
7. Said H, Yousif A, Humaid H (2011) iPhone forensics techniques and crime investigation. *Proceedings of the international conference and workshop on current trends in information technology*, pp 120–125
8. Zareen A, Baig S (2010) Mobile phone forensics: challenges, analysis and tools classification. In: *Proceedings of the IEEE international workshop on systematic approaches to digital forensic engineering*, pp 47–55
9. Kuntze N, Rudolph C (2001) Secure digital chains of evidence. In: *Proceedings of the IEEE international workshop on systematic approaches to digital forensic engineering*, pp 1–8
10. Raghav S, Saxena AK (2009) Mobile forensics: guidelines and challenges in data preservation and acquisition. In: *IEEE student conference on research and development*, pp 5–8

R²TP: Remote-to-Remote Transfer Protocols and Methods

Tae-Gyu Lee and Gi-Soo Chung

Abstract Recently Internet users can access data to share information in distributed environment by way of FTP, Telnet, Web browser etc. Consequently, the importance of these transfer programs is increasing. This paper proposes new remote transfer methods that can remotely upload or download the searched information into one more computers on the other side in Internet. This is a remote-to-remote transfer protocol (R²TP) for transferring the shared data from a third located computer or to other third located computer in Internet. In addition, this work proposes as well a simple user interface method easy to use as the selective or adaptive scenarios to transfer effectively the large volume of data.

Keywords Mobile computing · Remote file access · File transfer protocols · Wireless data access

1 Introduction

Lately the moving information seekers earnestly require the mechanism to save or transfer new searched data into the remote file system within their computers on remote locations through Internet [1].

T.-G. Lee (✉)

Korea Institute of Industrial Technology, KITECH B-108-1, 143 Hangeulro,
Sangnok-gu, Ansan, Gyeonggi, Korea
e-mail: tglee@kitech.re.kr

G.-S. Chung

Korea Institute of Industrial Technology, KITECH B-415, 143 Hangeulro,
Sangnok-gu, Ansan, Gyeonggi, Korea
e-mail: gschung@kitech.re.kr

Especially, the unbalanced traffic performance between backbone networks and private networks shall be our continuous concern and the unbalance issues need an improved substitute. A conventional data transfer protocols and algorithms including FTP, HTTP, SMTP etc. have a linear (one to one) model that the information is transferred between two computers on Internet [2]. Hence, a moving user with instant terminals such as PDA, KIOSK and mobile terminal cannot avoid downloading or storing some seeking data into temporary terminal storage. Then the searched data should be relay-uploaded to target server with adequate storages. Alternatively, he has memorized once and later will revisit the information site servers through networks. These previous works behave in annoying users and demand storage space for storing these searched data on instant terminal. However, almost instant terminals do not provide users with sufficient data space because of management inconvenience. They also do not give permanent data space in those terminals. Therefore, mobile users cannot escape on remote sites through networks.

To overcome these problems, we propose a new transfer protocol, R^2TP so that mobile users easy and quickly transfer their data from a remote server to the other servers through sever networks in spite of the limitations on that terminal. Additionally this work classifies all transfer application protocols into online or offline protocols. In addition, it provides an intelligent united transfer method, which enlarges the advantages and diminishes the disadvantages of online and offline protocols. It effectively supports large-volume, real-time, fault-tolerance for data transfer. The proposed transfer model reduces cost of data transfer by reforming transfer scenarios of past online protocols. That model shortens time delay by minimizing transfer delay and supporting server transfer scheduling for deterministic data transfer processes.

We will describe as following sequences. Next, [Sect. 2](#) describes some problems and issues about previous data transfer protocols and services. [Section 3](#) illustrates remote-to-remote transfer models as an alternative for the previous problems. [Section 4](#) presents the performance analysis model and the evaluation results of the proposed protocol R^2TP . Finally, [Sect. 5](#) shows conclusions and future directions on remote-to-remote data transfer.

2 Related Works

Internet users frequently access on the various multimedia such as audio, video, image, text and so on using different terminals. A practical range of Internet is increasingly expanded into ocean, heaven and so on from office, home, road, etc. of terrestrial [3, 4]. Internet transfer protocols have sources and destinations as transfer sites, which exchange data with each other. A source is the start system providing data source to transfer protocols. A destination is the target system receiving data source from transfer protocols.

Current transfer protocols can be classified and defined as the sets of ‘*online*’ or ‘*offline*’ transfer protocols dependent on connection status (i.e. real-time or non real-time) on a transfer path between source and destination.

Definition 1 *Online transfer protocols.* An Online Transfer Protocol is a transfer protocol, which keeps direct connection from a start site to target while delivering data sources. Therefore, it processes a real-time transfer of data going through one communication session. For example, FTP, HTTP etc. are included into, this set of online protocols [5, 6]. The online transfer protocols are possible immediately to check the exact delivery of data sources to the last target. Despite large data, those deliver smoothly and rapidly data to the final site.

Definition 2 *Offline transfer protocols.* An Offline Transfer Protocol is a piecewise transfer protocol, which provides the stepwise indirect connections from a start site to the last while eventually delivering data sources. Therefore, it realizes a step-by-step data transfer going through some temporary sites using several communication sessions. The indirect channels organize non real-time links such as from start to temporary server and from temporary to target server. For example, there are email, messenger, etc. [7, 8]. The offline transfer protocols eventually and gradually convey data sources from start to the last target despite non real-time connections of delivery path. Those are used to delivering of short messages or small data.

In online transfer protocols, it is impossible to convey data to destination when network failures happen accidentally on the links or servers between source and destination. Then clients cannot avoid waiting for completing the delivery of data source to destination. These problems exist in online transfer protocols as HTTP, FTP, etc.

Offline transfer protocols are impossible instantly to check the exact delivery of data sources to the last target, and are difficult and inefficient to transmit large data because of the limited space of temporary storages on intermediate paths. Primary offline transfer protocols with the problems are SMTP, Messenger, etc.

3 Remote to Remote Transfer Model

Among R²TP servers and clients, data accesses or transfer processes organize two channels same as all data transfer systems. That is, the connections between clients and servers are made of control and data channels.

We only consider the data time variables as the values of experimental elements except of control time variables. So it is assumed that the control times on every path, a, b and c are constant including setup time and hold time. In Table 1, the primary factors to have an effect on time delays for transferring data are data volume variables and unit speed variables on each path.

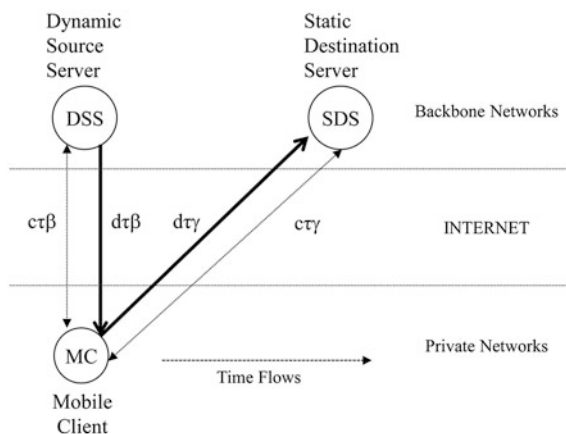
Table 1 Transfer elements and variables

Elements	Variables
Node	<i>DSS</i> (dynamic source servers): Start server with data sources <i>SDS</i> (static destination servers): Target server as data destination <i>MC</i> (R ² TP Clients): R ² TP clients
Link	<i>a</i> : A path between <i>DSS</i> and <i>SDS</i> <i>b</i> : A path between <i>MC</i> and <i>DSS</i> <i>c</i> : A path between <i>MC</i> and <i>SDS</i>
Data volume (Byte unit)	<i>S</i> (<i>x</i>): the data transfer volume (or file size) loaded on the path <i>x</i> by clients
Transfer speed for paths (Mbps unit)	α : Unit transfer speed for the path <i>a</i> β : Unit transfer speed for the path <i>b</i> γ : Unit transfer speed for the path <i>c</i>
Data time delay for data transfer (Sec unit)	<i>d$\tau$$\alpha$</i> : Data propagation time on path <i>a</i> <i>d$\tau$$\beta$</i> : Data propagation time on path <i>b</i> <i>d$\tau$$\gamma$</i> : Data propagation time on path <i>c</i>
Control time (Sec unit)	<i>c$\tau$$\alpha$</i> : The time needed for allocating path <i>a</i> <i>c$\tau$$\beta$</i> : The time needed for allocating path <i>b</i> <i>c$\tau$$\gamma$</i> : The time needed for allocating path <i>c</i>

We split the transfer system models into an existing client-server and a suggested server-server model. The existing transfer system establishes an *indirect* delivery route that a mobile client (MC) transmits data source from a dynamic source server (DSS) to static destination server (SDS) as Fig. 1. The model can be transformed to a trivial model, which directly transfers data between two servers while client keeping control connections to two servers.

The proposal model executes data transfers through the *direct* paths among two or more servers without passing by client. We organize a basic transfer topology to which one client and two servers accomplish data transfer processes. By dilating a

Fig. 1 Client-Server transfer model



basic model, let us obtain the generalized transfer model with three or more servers. The experiments in this paper carry out performance evaluations based on a basic model. Undoubtedly, the proposal model can identically support transfer scenarios of a conventional client–server model. R²TP transfer systems have a tri-transfer structure of client–server–server as Fig. 2.

Figure 1 shows a Client–Server (CS) model, which routes ‘b’ and ‘c’ as data transfer path. Each transfer speed of ‘b’ and ‘c’ is ‘β’ and ‘γ’ respectively. Each data volume is S(b) and S(c) respectively. We assume that a mobile client, MC supports an optimal intermediate path from b to c without user mediation. That is, mobile clients provide automatic transfer channels using its memory buffers for relaying data source through a path ‘b’(DSS to MC) to ‘c’(MC to SDS).

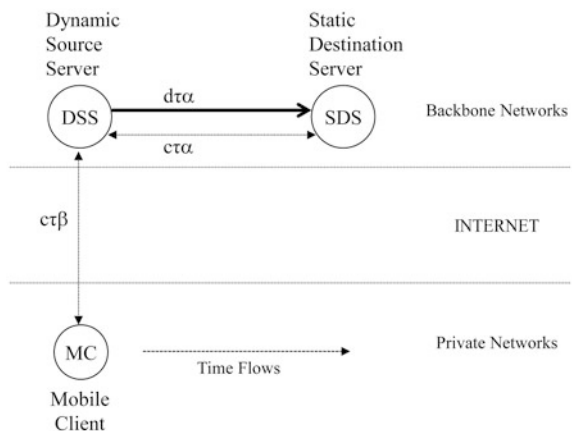
The proposed Server–Server (SS) model assumes that the network speed of data path, ‘a’ is ‘α’ as Fig. 2, and the data size per second on a path, ‘a’ is S(a). The transfer cost of CS method through the paths, ‘b’ and ‘c’ is expressed a cost function, Cost_{cs}, as (1). And the transfer cost of SS method through the paths, ‘a’ is expressed a cost function, Cost_{ss}, as (2).

$$Cost_{cs} = \frac{s(b)}{\beta} + \frac{s(c)}{\gamma} \tag{1}$$

$$Cost_{ss} = \frac{s(a)}{\alpha} \tag{2}$$

R²TP protocol is an application service protocol operated over TCP/IP suit such as HTTP, FTP, SMTP, etc. [9, 10]. The principle system components of R²TP protocols are organized into R²TP source server, R²TP destination server, and R²TP client. The R²TP protocol supports a set of FTP standard commands described in 10(OCT) RFC959 1985 [6].

Fig. 2 Remote to remote transfer model



4 Performance Analysis

The analysis model of this section assumes that a network topology is organized into unbalanced or asymmetric traffic loads between server networks and clients such as Sect. 2. There are α , β and γ as experimental variables. The symbol α indicates a transfer channel speed between source server and destination server. The β is download transfer channel speed between source server and client. Moreover γ is upload transfer channel speed between destination server and client.

4.1 Experimental Analysis in Internet

The experimental model assumes that two R²TP servers and a R²TP client are located in different areas through Internet. Inter-server speeds are firstly very high, secondly client-download speeds are high and finally destination-server upload has low speeds as the condition $\alpha \gg \beta > \gamma$. The data sizes for three transfer paths a , b , c have the same volumes as $S(a) = S(b) = S(c)$.

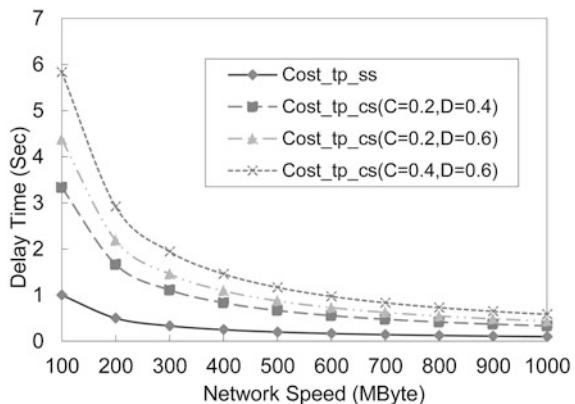
$$Cost_{tp_cs} = \frac{s(a) \times (\beta + \gamma)}{\beta \times \gamma}, \text{ by } S(a) = S(b) = S(c) \tag{3}$$

$$\beta = \alpha - \alpha * C, \gamma = \beta - \beta * D, \text{ by } (C, D : \text{coefficients})$$

$$Cost_{tp_ss} = \frac{s(a)}{\alpha}, \text{ by Equation(2)} \tag{4}$$

Figure 3 compares time delays of CS and SS models respectively according as variance of network speeds as the conditions followed. Firstly, the sizes of transferred data are the same as $S(a) = S(b) = S(c) = 1$. Secondly, the different coefficients i.e. C , D values express the speed difference between server and client networks.

Fig. 3 Comparisons on asymmetric internet



Let analyze and evaluate a CS transfer cost function dependent on the coefficients i.e. C, D in asymmetric Internet. The asymmetric model, which is almost similar to real Internet, is based on the fact that download speed, ‘β’ is quick in proportion to upload speed, ‘γ’ from server to client. That is, the coefficients (i.e. C, D stand for the differences of the relative network speeds of ‘β’ and ‘γ’ respectively over ‘α’). Generally, because the network speed for downloading is faster than that for uploading in Internet, it is assume that the coefficient D has a large value rather than C. Here are expressions ‘β’ = α - α*C and γ = ‘β’ - ‘β’ *D for CS model in (3) and SS in (4).

Based on Internet environments, the results of performance evaluations show transfer-speed enhancements of minimum twice to maximum seven-fold. As the speed gaps between server and client networks become enlarging, the improvement ratios of the R²TP model over the existing CS model become enlarging more.

4.2 Transfer Time Comparisons by Data Sizes

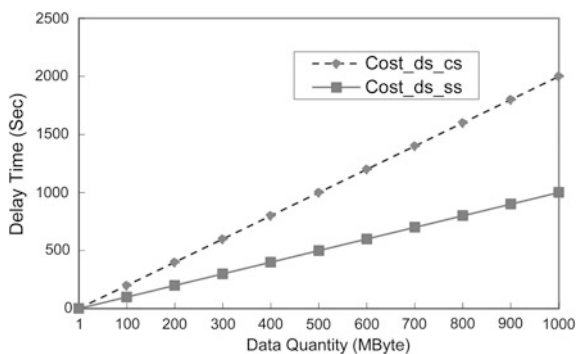
It is assumed that all clients and servers are in networks with the same topology that the network delays or speeds of every communication links have the same values (α = β = γ). The transfer cost functions are analyzed and compared according to the sizes of data sources. The cost functions of the existing CS model and the proposed SS model are as following (5) and (6) respectively.

$$Cost_{ds_cs} = \frac{s(b) + s(c)}{\alpha}, \text{ by } \alpha = \beta = \gamma, \text{ if } \alpha > 0 \tag{5}$$

$$Cost_{ds_ss} = \frac{s(a)}{\alpha}, \text{ by Equation (2)} \tag{6}$$

Figure 4 compares time delays in proportion to variance of S(x) under the conditions that the sizes of transferred data are the same as S(a) = S(b) = S(c). The results indicate that CS models have high time delays over SS models.

Fig. 4 Comparisons by differential data transfer size



Under the condition $S(a) = S(b) = S(c)$, the expanding results of (5) and (6) are a twice difference between the two. Independence of network speeds and data volumes, it is found that SS transfer costs are twice less than that of CS.

If a scope of performance analysis is extended to control information as commands beyond data, the transfer cost of the proposed R²TP model relatively goes down compared to that of a conventional linear model.

5 Conclusions and Future Works

In Figs. 3 and 4 of the previous section, we have shown performance enhancement ranging from minimum twice to maximum seven-fold according to coefficients of C, D of speed difference between server and client networks. The proposed protocols and systems of R²TP provide advantages of getting simple scenarios, safe and fast paths compared to that of traditional search and store systems. R²TP enhances economical and productive activity of information seekers by overcoming time (or transfer speed) and space constraints of accessed data.

Future works will realize multi-load, multi-save, etc. issues considering transfer performance, safety, and security problems in distributed system environments. Also we will research mobile file system, mobile backup system, etc. so that moving users effectively support backup and recovery processes of transferring data.

References

1. Imielinski T, Korth HF Storage alternatives for mobile computers. Architecture measurements, mobile computing, pp 473–503
2. Gunningberg P, Pink S, Bjorkman M, Sjodin P, Nordmark E, Stromquist J-E (1989) Application protocols and performance benchmarks. IEEE communications magazine, pp 30–36
3. Schulze B, Madeira ERM (2000) Migration transparency in agent systems. Ieice Trans Commun E83-B(5):942–950
4. Toshiaki T, Hideo M (1997) Mobile computing using personal handy-phone system (PHS). Ieice Trans Commun E80-B(8):1112–1118
5. Fielding R, Gettys J, Mogul J, Frystyk H, Berners-Lee T (1997) Hypertext transfer protocol—HTTP/1.1. RFC 2068
6. Postel J, Reynolds J (1985) File Transfer Protocol (FTP), RFC 959
7. Postel JB (1982) Simple mail transfer protocol, RFC 821
8. Tzerefos P, Smythe C, Stergiou I, Cvetkovic S (1997) A comparative study of simple mail transfer protocol (SMTP), post office protocol (POP) and X.400 electronic mail protocols. Department of Computer Science, The University of Sheffield, 211 Portobello St., Sheffield S1 4DP, 0742-1303/97 IEEE pp 545–554
9. Douglas E (1991–1993) Comer internetworking with TCP/IP, vols I, II, III. Prentice-Hall, Englewood Cliffs pp 33–51, 169–178, 223–260, 419–459, 161–206
10. Stevens WR (1994–1996) TCP/IP illustrated, vols 1, 2, 3. Addison-Wesley, Reading, pp 89–120, 171–202, 377–385, 391–401, 9–18

The Stack Allocation Technique on Android OS

Yeong-Kyu Lim, Cheong-Ghil Kim, Min-Suk Lee and Shin-Dug Kim

Abstract Garbage collection is one of major reason for performance degradation on Android OS. Escape analysis can be one of techniques to prevent performance degradation and Google has tried to implement scalar replacement through the escape analysis. But it does not become Android default functionality. This paper took it and compared with our proposed stack allocation method. The experimental result shows scalar replacement has no effect at all but stack allocation produce effective results. The CaffeinMark benchmark also shows no performance degradation in spite of additional instructions.

Keywords Dalvik VM · Stack allocation · Android

Y.-K. Lim (✉) · S.-D. Kim
Department of Computer Science, Yonsei University, 134 Shinchon-dong,
Seodaemun-gu, Seoul, South Korea
e-mail: postrain@yonsei.ac.kr

S.-D. Kim
e-mail: sdkim@yonsei.ac.kr

C.-G. Kim
Department of Computer Science, Namseoul University, 91 Daehak-ro
Seonghwan-eup, Cheonan, South Korea
e-mail: cgkim@nsu.ac.kr

M.-S. Lee
School of Computer Engineering, Hansung University, 116 Samseongyoro-16gil
Seongbuk-gu, Seoul, South Korea
e-mail: minsuk@hansung.ac.kr

1 Introduction

As Android OS, one of the major mobile OS together with Apple iOS, is free to everyone and widely used by many smart phone manufacturers [1]. Unlike other mobile OS environments, most applications on Android are written in Java language and require virtual machine (VM) for running [2]. To do this, Google adopted Dalvik VM rather than popular JVM made by SUN because it is adequate for resource constrained mobile devices [3, 4]. Even though Dalvik already has several optimization techniques to run well on mobile hardware devices like phones or tablets, it still has more room for further optimization than JVM. Unfortunately, escape analysis, one of the optimization techniques to be utilized to improve the performance of applications, is not available on the latest Android version 2.3(Gingerbread). The analysis result can be applied to scalar replacement, synchronization removal, and stack allocation method to increase the performance. Gingerbread may try to use scalar replacement technique; unfortunately, it is not included in official Android release.

As mentioned earlier, the purpose of escape analysis is mainly improving application performance and its analysis result can be utilized to reduce the overhead of garbage collection (GC), which is one of performance bottlenecks on Dalvik VM. If GC occurs too many times and takes a long time to run, this can make UI stopping. As a result, user may feel that phone is not working for a while. To overcome this problem, we introduce stack allocation technique through the escape analysis. Escape analysis can help to determine precisely which object can be stack-allocated. The expected result can reduce the frequency of GC, thus this will improve application execution performance [5, 6].

In the next section, we present the proposed stack allocation method working with DX and Sect. 3 describes design and implementation of proposed method. Section 4 explains the experimental results using Android Smartphone. The last section concludes this paper with a summary and suggestion about future works.

2 Proposed Method

2.1 Escape Analysis

GC may give influence on the system performance adversely in Java-based systems where all the objects are dynamically allocated in heap space and never freed explicitly by the programmer. Escape analysis can help to reduce dynamic allocations by analyzing the attributes and life scope of objects. It analyzes whether they escape a method (i.e., whether they should be alive even after the method in which they are allocated) or return to its caller. If an object is confirmed not to escape, the object can be replaced into a scalar data. In Gingerbread, DX only does escape analysis for arrays. Scalar replacement, one of techniques through the



Fig. 1 The process of escape analysis on Dalvik

escape analysis, for array means that only the used elements in a non-escaping array are allocated in virtual registers rather than in heap space, removing the new() operation. Currently, the escape analysis for Gingerbread framework does almost no help to increase the performance, because conditions to be non-escaping limits the number of target object instances. In Gingerbread, the conditions with which the DX decides an array (element) as a non-escaping object are as follows. First, the size of array is declared as a constant. Second, the array object is not referenced from other objects. Third, the elements of array are accessed with constant indices. Fourth, the array is used as neither a return value nor a parameter for other methods (Fig. 1).

2.2 Stack Allocation

Every Java object has its life scope. Some objects live only in a method. Others do in a class or throughout a thread. Dynamic allocation scheme with no explicit freeing in Java allows programmers to program easily but requires GC which consumes considerable amount of CPU resource. GC is one of the key reasons for non-smooth user interface in Android Smartphone.

If the scope of an object is limited to a method or its child methods, it can be safely allocated in the stack of the method rather than the heap area. By lowering the heap requirements, stack allocation can reduce the number of invocations of garbage collector. And the stack allocation reduces the object allocation time because it uses much simpler push operation on behalf of heavy new() operation. And when a method returns, and the objects have the life scope only within the method, the objects in the object allocation stack are automatically discarded while the stack pointer and the frame pointer move to their caller’s positions respectively. However, the heap allocated objects occupy the heap memory until the garbage collector sweeps the objects no longer needed.

The difference between the stack allocation and the scalar replacement in DX for Gingerbread is that our stack allocation is for all possible objects and does not replace the objects into scalar data. Even if objects are allocated in stack, they are still objects in nature. Objects referenced by the stack-allocated objects could reside on heap area, and the stack allocated objects themselves can be implicitly or explicitly referenced by other objects. That means the garbage collector still needs to scan all the stack allocated objects alive.

An object cannot be allocated in stack when at least one of the following conditions is met:

- i. Target object is not used as a parameter of other methods
- ii. Another object is used as a parameter of the method in the target object
- iii. The type of the return value of target object is an object
- iv. The target object is a thread
- v. The target object is assigned to other escaping object.

To implement stack allocation scheme, we introduced a new Dalvik opcode named `STACK_INSTANCE`, which is a stack version of `NEW_INSTANCE` opcode. By adding a new instruction, Dalvik VM also needs to be changed to support `STACK_INSTANCE` instruction.

3 Implementation

3.1 *Interpreter Stack*

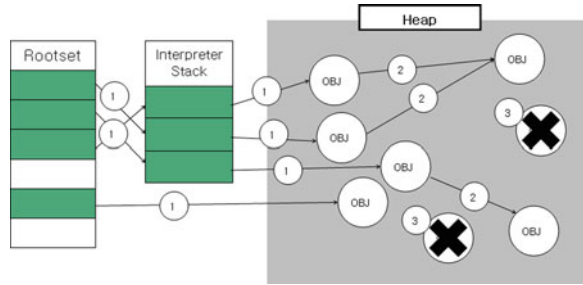
All Dalvik objects have ‘Object’ structure as their first field. To support stack allocation of objects, we added a flag in the ‘Object’ structure to distinguish stack-allocated objects from heap-allocated objects. Interpreter stack is a virtual stack used by Dalvik VM. Dalvik VM manages the interpreter stack as real CPU does. Every method has its own stack frame in the interpreter stack. In the stack frame, there are input parameters, local data, and the information to maintain the stack frame itself and the thread’s object allocation stack.

To allocate objects in stack rather than heap area, we added object allocation stack to each thread. The object allocation stack is separated from interpreter stack though they are managed in a similar way. Initially, the threads are given 16 KB of object allocation stack. If the thread requires more stack area to allocate objects, Dalvik VM allocates more stacks (up to 40) in the unit of 16 KB. When system has no more memory to allocate, Dalvik VM allocates all the objects in heap area by changing `STACK_INSTANCE` opcode to `NEW_INSTANCE` opcode. Whenever a method is called or returned, the object allocation stack of the thread also grows or decreases. Each thread maintains all the information on the object allocation stack on the internal data area of stack frame in the interpreter stack.

3.2 *Garbage Collection*

In Dalvik VM, as a kind of Java VM, GC process is called when threads fail to allocate objects on heap due to the shortage of memory. GC is to collect all the objects which are allocated and expected not to be used anymore. As shown in Fig. 2, the mark and sweep algorithm for garbage collection is the following sequence:

Fig. 2 Garbage collection without stack allocation



1. Suspend threads
2. Mark rootset objects
3. Mark objects recursively referenced by rootset objects
4. Sweep unmarked objects
5. Resume threads.

Rootset objects are directly accessible objects which include the class loader, primitive classes, thread objects, thread stacks, and strings. Garbage collector recursively marks objects referenced by Rootset objects and sweeps the remaining orphan objects.

In stack allocation scheme, one more step is needed to mark the objects in object allocation stack. Though garbage collector also needs to scan the object allocation stack, the total number of objects to be scanned is not changed from the number without stack allocation. This indicates that the performance degradation by scanning the object allocation stack is minimal.

4 Evaluation

To evaluate the performance of stack allocation over heap allocation, we took CaffeinMark benchmark [7] and also used one sample application. The test device is equipped with ARM11 800 MHz CPU clock speed and Gingerbread version of Android OS. Table 1 shows the evaluation result of CaffeinMark benchmark. The experiment was carried out 10 times and each row summarizes the average value.

The stack allocation (shown in the right-hand column of the Table 1) performs slightly better than heap allocation in most cases. Even though *Method* score is lower than heap allocation due to the overhead in the method call, the total score is almost the same because this overhead can be compromised with reducing the GC operation time while application is running. This result can conclude that the overhead of stack allocation is small and can be ignored.

We also evaluated the performance using another sample application, named as *Cyclingbox*. It draws a white box along the boundary of the display screen. It shows frame count in the middle of the screen. The result is summarized on Table 2.

Table 1 CaffeinMark benchmark result

	Heap allocation	Stack allocation
Sieve	2304.0	2369.5
Loop	3899.5	4002.5
Logic	4416.1	4639.7
String	1784.0	1645.5
Float	1680.3	1719.2
Method	1847.5	1792.1
Overall	2455.6	2461.7

Table 2 Evaluation result with *Cyclingbox* application

	Heap allocation	Stack allocation
Frame count (1 min)	996	1785
GC invocation (1 min)	222	4

5 Conclusion

A stack allocation method is proposed to reduce GC effect on Android OS. For this, new STACK-INSTANCE opcode was introduced and Dalvik VM is modified to interpret new opcode. We simulated the proposed method on real Android device with latest version and compare its performance with built-in method to show the efficiency of our approach. The evaluation result shows stack allocation can effectively reduce GC frequency without generating additional overheads on Dalvik VM. Our future work will continue on improving VM performance on latest Android version.

References

1. Morrissey S (2010) iOS forensic analysis for iPhone, iPad, and iPod touch. Springer, New York
2. Hashimi YS, Komatineni S (2009) Pro android. Springer, New York
3. Grønli T, Hansen J, Ghinea G (2010) Android vs Windows Mobile vs Java ME: a comparative study of mobile development environments. Proceedings of the 3rd international conference on Pervasive technologies related to assistive environments. ACM, New York, pp 1–8
4. Bornstein D, Dalvik VM (2010) internals
5. <http://sites.google.com/site/io/dalvik-vm-internals>
6. Blanchet B (1999) Escape analysis for object oriented languages. Application to JavaTM, Proceedings of the 14th ACM SIGPLAN conference on object-oriented programming, systems, languages, and applications, ACM, New York, pp 20–34
7. CaffeinMark 3.0, Pendragon Software Corporation

New Load Balancing Method for VoD Service

Jinsul Kim, Kang Yong Lee and Sanghyun Park

Abstract As IPTV services advance, an increasing user demand and tremendous content volume of multimedia content cause some difficulties in network resource management. In this paper, we propose an improved load balancing algorithm for VoD service. Unlike existing algorithms, the proposed algorithm considers users' behaviors for VoD service, and determines the most efficient allocation of VoD requests by estimating expected server load and expected user waiting time. In order to evaluate our algorithm, we conduct a simulation of an IPTV network and verify the effectiveness of our system by comparing it with two baselines (Least Load and Nearest methods). As experimental results, the proposed system performs significantly better than the baselines. Consequently, the system can manage limited network resources efficiently and enhance QoE.

Keywords Resource management · IPTV · VoD server · Load balancing · User behaviors

J. Kim · S. Park

School of Electronics and Computer Engineering, Chonnam National University,
Gwangju 550-757, South Korea

e-mail: jsworld@jnu.ac.kr

S. Park

e-mail: sanghyun@ejnu.net

K. Y. Lee (✉)

Electronics and Telecommunication Research Institute, Gajeong-Dong,
Yuseong-Gu, Daejeon 305-700, Korea

e-mail: kanglee@etri.re.kr

1 Introduction

Due to the spread of IPTV services, public interest about the goodness of services is growing, and accordingly, service providers are trying to ensure important aspects (i.e., reliability, security, and quality) of the media and service network. However, since the number of IPTV subscribers has increased tremendously, a sharp traffic growth has occurred, which has caused some difficulties for keeping effective management of IPTV network resources, and consequently, the quality of service may degrade.

In this paper, in order to effectively manage network resources in IPTV services, we propose a load balancing method for Video-on-Demand (VoD) servicing server, utilizing usage behaviors of IPTV subscribers. While there have been many attempts to analyze usage behaviors of IPTV users or particularly VoD service users in the recent, few works have utilized the usage patterns for the network resource management. Based on usage information and characteristics of VoD service, the proposed method first generates a statistical model which estimates the end-watching time for VoD service users. Next, the system uses the potential workload and available bandwidth estimated by the statistical model so that workload in VoD server is effectively distributed. In addition, predicted user waiting time is considered to improve the balancing performance.

The rest of this paper is organized as follows. In [Sect. 2](#), we investigate previous work in server load balancing for VoD service and describe how it is limited to work in our task. We present the analysis of usage behaviors and patterns of VoD services in [Sect. 3](#). [Section 4](#) describes our load balancing system which exploits the analyzed usage patterns and features of VoD service. In [Sect. 5](#), we report the experimental results obtained by simulating VoD service network. Finally, we summarize the contributions of our research in [Sect. 6](#).

2 Related Works

There have been many previous studies related to VoD services. For example, content caching, admission control, and server load balancing [[1–3](#)] have been researched to provide effective VoD services, and among those important topics, the load balancing was one of the most intensively studied.

Especially in VoD service environments, service requests are not evenly generated but frequently skewed upon the time line and locations of users. For example, too many service requests can be delivered to a particular server, and due to the maximum capacity of requested services for the server, the requests are declined even if sufficient resources are available over the entire network. Thus, there are many users in a mass traffic network (e.g., IPTV service network), and load balancing is critical because effective balancing can reduce the probability of blocking service requests and help to preserve successful management of overall network resources.

For effective load balancing, most previous research has succeeded in replying to service requests by considering instant server status and available bandwidth, but usage behaviors and special characteristics of VoD service, which could improve the management performance, were largely ignored.

Recently, [1] proposed the algorithm to maximize server utility for effective distribution. They considered several features (e.g., transmission delay (i.e., latency), admission control, and popularity of serviced contents) to improve the server utility, thus effectively expanded the disc bandwidth.

In [2], network environments consist of a main multimedia server and the cluster of local proxy servers managed by a tracker, and to avoid overloading on the main server, the local proxy server group containing 6 adjacent media servers redistributes the traffic in accordance with validity and popularity of server contents. However, they ignored some important factors such as bandwidth and latency, and their system could not ensure Quality of Experience (QoE), e.g., delay of service waiting time occurs.

In order to reduce service waiting time, the system in [3] employed a server load balancer and they considered various components, e.g., the amount of discs, bandwidth, and contents popularity in the server.

3 Analysis of Features in VoD

According to [4, 5], VoD service contains two principle characteristics distinguished from ordinary IPTV services, e.g., multi-channel broadcasting and web-service, and there is useful information which can be utilized for our task. First, the termination time of VoD service is pre-informed. Since VoD content is already stored in a corresponding media server, its play time is fixed and thus serviced content is terminated on the fixed time unless users pause the services [6]. The other different characteristic is the terminating pattern of VoD service. In many actual cases, users have decided to continue VoD service after watching the beginning of requested video [7, 8], and we frequently observed that the service was terminated earlier by users [9].

Figure 1 shows the difference of two traffic logs, both of which contain an equal bandwidth, in three different cases: (i) two data streams start to flow at the same time, (ii) terminate at the same time with different start times, and (iii) last for the same period with different start and end times). The figure indicates that two traffics can cause potentially different effects on the server even if the data streams include an equal bandwidth.

Left-top of Fig. 1 denotes the case where the traffics begin identically but end at different times. At present, both traffics occupy 300 kbps bandwidth, but traffic A will be terminated after 10 min, and traffic B will end after 5 min, which means traffic A contains potentially more influences on the server than traffic B. Right-top of Fig. 1 is the case of different-starting and equal-terminating. In this case, due to the same termination time, both traffics seem to have the same effect to the server.

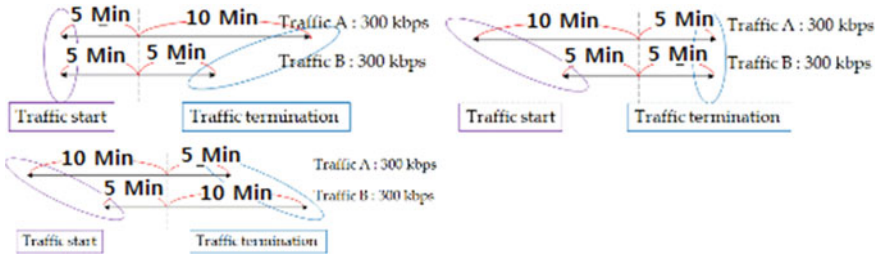


Fig. 1 Server influence by two same-bandwidth traffics in three different cases: same starting time, same terminating time, and same lasting time

However, beginning later would have higher probability to terminate, and so traffic A may include potentially more effects on the server. Bottom of Fig. 1 indicates the case where both traffics last for the same amount of time if they start and terminate differently. In this case, traffic starting and terminating times are different, and so obviously both traffics could differently influence on the server.

In this paper, we utilize the service terminating pattern and user behavior in VoD service and propose a metric which models currently available bandwidth and potential effects by service traffic, so that workload is effectively distributed to servers.

4 Load Balancing Method

In this section, we describe the method to distribute server load in a VoD service network. The load balancing system that we propose is expected to be effective in two different cases. First, if the bandwidth of a media server is available enough for requests, our system can allocate workload to the server. The other is the case that all media servers currently lack of available bandwidth for requests. In this case, workload is effectively distributed by considering the estimation of server waiting time. In the following sections, we describe how our system can handle those two cases in detail.

4.1 Case I: Sufficient of Available Bandwidth of Media Server

In our system, the main server receives VoD service requests of users and calculates the weight of an available bandwidth of each media server managed by the main server. Then, the requested services are dispatched to the media server engaged with the highest weight. Since each media server is implemented to report its status (e.g., available bandwidth and VoD session information) periodically, we can generally assume that the main server owns the status information of each managed media server.

The weight of an available bandwidth of each server is calculated as follows.

$$WBW = \alpha \times CBW + (1 - \alpha) \times PBW \quad (1)$$

where WAB indicates the weight of available bandwidth, CBW denotes the currently available bandwidth, PBW denotes the potentially available bandwidth, and α is a weight value.

In Eq. (1), CBW indicates the available bandwidth of a media server at the time that user requests are received, and PBW is a probabilistic value, the sum of estimated bandwidth obtained by terminating current running VoD sessions. Also, α is a bias to CBW spanned in $[0,1]$.

Moreover, as the CBW decreases, we intend to give more weights on the current available bandwidth (CBW) than the potential available bandwidth (PBW). Thus, the definition of α is given as:

$$\alpha = 1 - \frac{CBW}{OBW} \quad (2)$$

where OBW indicates the overall bandwidth of a media server in Eq. (2).

In addition, PBW is obtained as:

$$PBW = \sum_{i=1}^n \left\{ f\left(\frac{t_i}{T_i} \times 100\right) \times BW_i \right\} \quad (3)$$

In Eq. (3), n indicates the number of VoD sessions that a media server is servicing, the function, $f(\cdot)$ is the probabilistic density function that describes the distribution of terminating times for VoD sessions, t_i is the played time of the i -th VoD session, T_i is the total length of the i -th VoD session, and BW_i is the required bandwidth to service the i -th VoD session.

Based on $f(\cdot)$, we can estimate the termination time of each VoD session and apply the BW_i on this function to consider a potential bandwidth obtained by terminating VoD sessions. As a result, the main server computes the weight of an available bandwidth of every managed media server, and then assigns the requested services to the server containing the highest weight, which would effectively distribute the workload of VoD services in our expectation.

4.2 Case II: Lack of Available Bandwidth of Media Server

In this section, we describe how our system can work in the case that none of media servers contain sufficient available bandwidths. The main server traditionally assigns service requests to the media server expected to include the shortest estimated waiting time. In other words, requested services are dispatched to the media server where the shortest waiting time is predicted. In this paper, we define Expected Response Time (ERT), as the waiting time of users until requested services begin to run.

In order to estimate the ERT, we first compute the Expected Playing Time, EPT, defined as the expectation of remained playing time until terminating currently running sessions. We utilize the probabilistic density function of the VoD session termination time distribution to compute the EPT for the i -th VoD session, as follows.

$$EPT_i = \left\{ \int_{\frac{t_i}{T_i} \times 100}^{100} xf(x|x > \frac{t_i}{T_i} \times 100)dx \right\} \times \frac{T_i}{100} - t_i \tag{4}$$

In $f(\cdot)$ of Eq. (4), the total length of a VoD content is standardized and its maximum value is 100. Thus, the average playing time computed by using $f(\cdot)$ is also standardized with the same length. After converting the standardized time to a real average playing time and plugging the played time t_i in this, we can obtain the remained playing time of the i -th session. However, if the i -th session did not start to play on a corresponding media server, i.e., waited in the server waiting queue, we put the t_i as 0 and then estimate the EPT.

In addition, to compute the ERT, the main server calculates the Expected Waiting Time, EWT, for service requests which are waiting in the queue of a media server. The computation can proceed in the following order.

- (a) In n number of service requests waiting in the queue of a media server, the EWT of the first waiting request is the minimal of the expectation values of the playing times of currently running VoD sessions in the server.
- (b) The EWT of the i -th waiting session is calculated by comparing the sum of the EPTs and the EWT of its preceding waiting sessions.
- (c) The ERT of a media server is computed by the EWT of the $n + 1$ th service request.

To give better understanding, we show an example of the EWT computation in Table 1.

In Table 1, a media server is currently playing 5 VoD requests (sessions), i.e., S1, S2, S3, S4, and S5, and other 4 VoD services, i.e., Q1, Q2, Q3, and Q4, are waiting in the queue. In addition, the EPT value of each request is estimated by Eq. (4). The EWT computation is described as follows.

Table 1 Example of estimation

Running sessions		Waiting queue	
VoD request ID	EPT	VoD request ID	EPT
S ₁	6	Q ₁	2
S ₂	12	Q ₂	4
S ₃	7	Q ₃	10
S ₄	14	Q ₄	5
S ₅	9	-	-

- Q1, placed at the first position of the queue, can be processed if anyone of currently servicing sessions is terminated. So, the EWT of Q1 is the smallest EPT whose session is currently running, i.e., $EWTQ1 = EPTS1 = 6$.
- To calculate the EWTQ2 placed at the second position of the queue, we compare S3, the second smallest EPT, and Q1, waiting in just front of Q2, because Q2 can be processed after terminating Q1 and if the sum of EPTQ1 and EWTQ1 is smaller than EPTS3, i.e., $EPTQ2 + EWTQ2 = 8$, $EPTS3 = 7$, $EWTQ2 = 7$.
- Similar to the previous case, to estimate the EWTQ3, we need to compare S5, the third smallest EPT, and Q1&Q2. In mathematical notations, $EPTQ1 + EWTQ1 = 8$, $EPTQ2 + EWTQ2 = 11$, $EPTS5 = 9$, $EWTQ3 = 8$. That is, Q3 is not serviced right after ending S5, but after terminating S1 and processing Q1, Q3 will be processed.
- Since Q3 is processed after ending Q1, we should consider S5, and Q2&Q3, to estimate the EWTQ4, i.e., $EPTQ2 + EWTQ2 = 11$, $EPTQ3 + EWTQ3 = 18$, $EPTS5 = 9$, $EWTQ4 = 9$.

As shown above, we can obtain the EWT of every waiting request, and considering the obtained EWTs, we can estimate the ERT of a corresponding media server by postulating that new VoD requests can come to the server. In other words, the EWTQ5 (i.e., new VoD service request) needs to consider S2, Q2, Q3, and Q4. Besides, the sum of the EPTQ2 and EWTQ2 are minimized, and therefore the ERT of this server is 11.

The example we described depicts a way to estimate the ERT of a media server. Following this method, the main server can calculate the ERT of every media server and assign new VoD service requests to the media server whose ERT is the smallest.

5 Evaluation

In this section, we evaluate our system by simulating an IPTV service network. We analyze the experimental results.

5.1 Experimental Results

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

In this section, we verify the effectiveness of our load balancing method by comparing with two baselines: 5.1.1 Nearest and 5.1.2 Least Load methods.

5.1.1 Nearest Method

In this method, the VoD service request is assigned to the nearest server from the request position. Among the media servers within the maximum transmittable distance from the request location, the system searches the one whose available bandwidth is sufficient, and connects the request to the server. If no appropriate server is returned in the search, the nearest server is assigned.

5.1.2 Load Method

The Least Load method is the way to allocate the least workload server for VoD service requests. The system finds the servers whose available bandwidths are sufficient within the maximum transmission distance, and assigns the requests to the server whose bandwidth ratio (ratio of occupied bandwidth to the overall bandwidth) is the smallest. If none of servers contain enough available bandwidth, the system coercively allocates the media server that contains the least sum of playing times of waiting sessions.

5.1.3 Comparison Against Baselines

In the series of experiments, we measure the ERT of our system and two baseline systems over the changes of three environment parameters. In order to examine the effect of each parameter, two different parameters are fixed when another parameter changes where OBW indicates the overall bandwidth.

Figure 2 shows the ERT results of three methods where 9, 16, and 25 media servers are employed, 10 traffics are sent for a minute, and maximally 5 sessions can be waited in the queue. The x-axis indicates the number of media servers, and the y-axis denotes the ERT of each system. Overall, the ERT decreases as the number of media servers increases because the network capacity expands. However, the proposed method shows the least ERT over all cases, i.e., the most effective. In addition, the Least Load method performs better than Nearest method, and we guess the reason that allocating the least workload server can allow more quick reply to the requests than assigning the nearest server, i.e., server workload may have more effects on the ERT than server location.

In Fig. 3, the x-axis denotes how much traffic is generated per minute, and the y-axis indicates the ERTs of the three different methods. We change the amount of traffic generated to 10, 20, and 30 when the network contains 25 media servers whose length of waiting queue is 5. The ERT is proportional to the amount of traffic generated in general, but the increment from 20 to 30 in the x-axis values stagnates because new VoD requests are declined if the waiting queue is fully occupied. In accordance with the result in Fig. 2, our system shows the best performance, and significantly better performance is achieved if more traffic is generated.

Fig. 2 Expected response time against the number of media servers of three systems: proposed, nearest, and least load

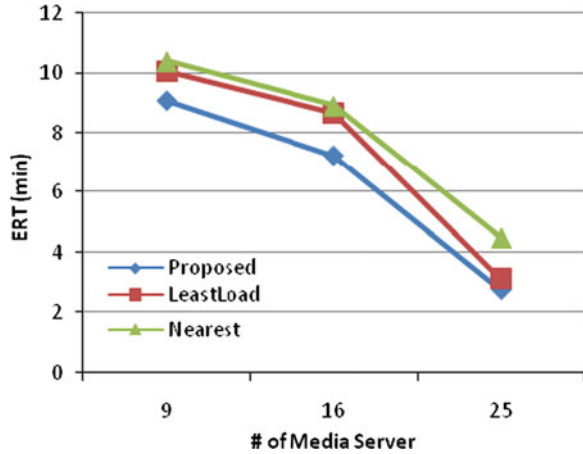
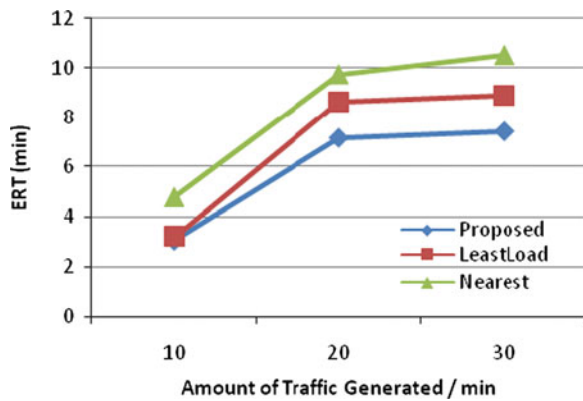


Fig. 3 ERT against the amount of traffics generated (per a minute) of three systems: proposed, nearest, and least load



6 Conclusions

In this paper, in order to maximize the utility of resources in an IPTV network, we proposed an effective load balancing algorithm which utilizes user usage patterns in VoD services. While previous methods considered only current server load and traffic status, our method estimates potential server load by utilizing features in VoD services, i.e., early termination pattern and usage behavior, which could distribute workload effectively. To improve the performance, our system statistically models termination times of VoD services, and based on this model, we estimate EBW and ERT to use network resources efficiently and minimize user waiting time. We measure user waiting time (i.e., ERT) which indicates the performance of a load balancing system, and examine the influences of three different network parameters on the performances: (i) the number of media servers, (ii) the amount of traffic generated, and (iii) the length of the waiting queue. As experimental results, the proposed system performs significantly better than the

baselines. The VoD server load balancing method we propose is more effective in a wider network such as an IPTV network, where many users access the network and various contents are serviced, and consequently, the system can manage limited network resources efficiently and enhance QoE.

Acknowledgments This research was financially supported by research program, Chonnam National University, Korea, 2012.

References

1. Huang YF, Fang CC (2004) Load balancing for clusters of VoD servers. *Inf Sci* 164(1–4): 113–138
2. Dakshayini M, Guruprasad H, Maheshappa H, Manjunath AS (2007) Load balancing in distributed VoD using local proxy server group. *International conference on computational intelligence and multimedia applications*, vol 4. pp 162–168
3. Sujatha D, Girish K, Rashmi B, Venugopal K, Patnaik L (2007) Load balancing in fault tolerant video server. *LNCS*, pp 306–315
4. Cha M, Rodriguez P, Moon S, Crowcroft J (2008) On next-generation telco-managed P2P TV architecture. *International workshop on peer-to-peer systems (IPTPS)*
5. Cha M, Rodriguez P, Crowcroft J, Moon S, Amatriain X (2008) Watching television over an IP network. *ACM SIGCOMM IMC*
6. Kusmierek E, Czyrnek M, Mazurek C, Stroinski M (2007) iTVP: large-scale content distribution for live and on-demand video services. *Multimedia computing and networking SPIE-IS&T electronic imaging, SPIE*, vol 6504
7. Yu H, Zheng D, Zhao B, Zheng W (2006) Understanding user behavior in large-scale video-on-demand systems. In: *Proceeding of Eurosys*
8. Kim J, Um T-W, Ryu W, Lee BS, Hahn M (2008) Heterogeneous networks & terminals-aware QoS/QoE-guaranteed mobile IPTV service. *IEEE Commun Mag* 46(5)
9. Jung Y, Park Y-m, Bae HJ, Lee BS, Kim J (2011) Employing collective intelligence for user driven service creation. *IEEE Commun Mag* 49(1):76–83
10. Leal RP, Cachinero JA, Martin EP (2011) New approach to inter-domain multicast protocols. *ETRI J* 33(3):355–365

Exploiting Mobility/Traffic Characteristics for PMIPv6-Based Distributed Mobility Management

Ki-Sik Kong

Abstract The centralized mobility management protocols such as mobile IPv6 (MIPv6) and proxy mobile IPv6 (PMIPv6) may result in significant amount of data and control traffic being pushed into the central mobility anchor of core network. Moreover, the use of such a central mobility anchor may be vulnerable to a single point of failure and degrade overall system performance. Therefore, in order to overcome the limitations of centralized IPv6 mobility management protocols, signal-driven PMIPv6 (S-PMIP) [3] was proposed. However, it cannot solve the problems of the potential bottleneck at local mobility anchor (LMA) and long handover delay. Therefore, FS-PMIP, which is our previous work in [4], was proposed to alleviate the problems of S-PMIP. By exploiting a user's movement locality, FS-PMIP effectively reduces both the access to LMA and the handover latency. However, even if FS-PMIP is an efficient PMIPv6-based distributed mobility management scheme, it still requires the access to LMA whenever the correspondent node (CN) sends the data packets to the mobile node (MN). Therefore, in this paper, we propose an Enhanced FS-PMIP (EFS-PMIP) to enhance FS-PMIP by exploiting a user's traffic locality as well as movement locality, which adopt both the pointer-forwarding concept and the working set concept in communications. The proposed EFS-PMIP is expected to have apparent potential to effectively reduce the access to LMA and distribute the binding and routing functionalities at mobile access gateways (MAGs).

Keywords PMIPv6 · S-PMIP · FS-PMIP · Distributed mobility management

K.-S. Kong (✉)

Department of Multimedia, Namseoul University, Cheonan, Republic of Korea
e-mail: kisik.kong@gmail.com

1 Introduction

The recent standardized IPv6 mobility management protocols such as Mobile IPv6 (MIPv6) and Proxy Mobile IPv6 (PMIPv6) [1] rely on a central mobility anchor (e.g., LMA in PMIPv6) that manage both data plane and control plane. However, such centralized mobility management architecture may result in significant amount of data and control traffic being pushed into the central mobility anchor of core network, which may cause serious bottleneck. In addition, the use of such a central mobility anchor may be vulnerable to a single point of failure and degrade overall system performance [2, 3]. In order to overcome the limitations of such centralized mobility management protocols, IETF has recently discussed the Distributed Mobility Management (DMM). In [3], as a partially distributed approach, a signal-driven PMIPv6 (S-PMIP) was proposed, which introduces the querying mechanism to find the mobile node's (MN's) location from the LMA when the data packets sent from the CN arrive at the CN's mobile access gateway (MAG). However, even if S-PMIP is an efficient distributed mobility approach, it still cannot solve the following problems: (i) the potential bottleneck problem at LMA that may be incurred by a huge number of the proxy binding update (PBU) messages, and (ii) long handover delay and packet losses due to the long distance between LMA and MAG.

Therefore, in order to alleviate such problems, a pointer forwarding-based S-PMIP for distributed mobility management (FS-PMIP), which exploits a mobile user's movement locality characteristics, has been proposed in our previous work [4]. In this work, the access to the LMA caused by the PBU messages by a mobile user's movement and the handover delay can be significantly minimized by registering with the previously-serviced MAG, not the LMA. However, even if this scheme effectively reduces the number of PBU messages caused by a mobile user's movement, it still reveals the problem that cannot reduce the number of the proxy binding query (PBQ) messages destined from CN-MAG to the LMA. Therefore, in this paper, we enhance our previous work in [4] by exploiting a mobile user's traffic locality as well as movement locality for more enhanced distributed mobility management.

2 Exploiting Mobility/Traffic Characteristics for PMIPv6-Based Distributed Mobility Management

2.1 Motivation

In [5], the authors made the key observation that while the potential set of sources for the MN may be large, the set of sources that a given MN communicates most frequently with is very small. Based on this observation and the concept of a working set in communications, we propose a new PMIPv6-based distributed

mobility management scheme that exploits effectively a user's traffic locality as well as movement locality in order to minimize the access to the LMA in terms of both location registration (i.e., the number of PBU messages) and packet delivery (i.e., the number of PBQ messages) procedures.

2.2 An Enhanced Pointer Forwarding-Based S-PMIP for Distributed Mobility Management (EFS-PMIP)

We propose an enhanced pointer forwarding-based S-PMIP, called EFS-PMIP, which adapts both the pointer-forwarding concept [6] and the working set concept [5] in communications to S-PMIP for more efficient distributed mobility management. In order to do this, we extend the MN's policy profile configurations stored in the AAA server when compared with that of PMIPv6. That is, the MN's current pointer forwarding length K_{cur} , the MN's predefined-pointer forwarding threshold K , the MN's patron MAG's proxy care-of-addresses (proxy-CoAs), and the MN's current MAG's proxy-CoA are assumed to be additionally included in the policy profile of AAA server (In this paper, we call the set of MAGs indicating strong traffic locality towards a certain MN as its "patron MAGs").

The illustrative example of EFS-PMIP is shown in Fig. 1. First of all, the location registration procedures in EFS-PMIP are as follows. When an MN1 attaches to an access network connected to the MAG1 (**Step 1**), the MAG1 sends a AAA query message including the MN1's identifier and the MN1's current MAG's (i.e., MAG1's) proxy-CoA to the AAA server (**Step 2**). After receiving the AAA query message, the AAA server performs the following three procedures:

- (i) **the comparison between the MN1's K_{cur} and K** : if K_{cur} is less than K , the value of K_{cur} is increased by 1 (Initially, the value of K_{cur} per MN at AAA server is configured with -1). In this case, the PBU/PBA message exchanges with the previously-visited MAG are performed. On the other hand, if K_{cur} is equal to K , the value of K_{cur} is reset to 0. In this case, the PBU/PBA message exchanges with the MN1's all patron MAGs as well as the LMA are performed.
- (ii) **retrieving either the MN1's previous MAG's proxy-CoA (in case of $K_{cur} < K$) or the MN1's patron MAGs' proxy-CoAs (in case of $K_{cur} = K$) and then storing the MN1's current MAG's proxy-CoA** (e.g., MAG1's proxy-CoA in Fig. 1) in the AAA server
- (iii) **normal authentication procedure** defined in PMIPv6

After that, the AAA server sends the AAA reply message including the MN1's previous MAG's proxy-CoA or the MN1's all patron MAGs' proxy-CoAs (**Step 3**). Note here that if the MN1 first enters the PMIPv6 domain, the previous MAG's proxy-CoA will be null. Otherwise, this address will be the MN1's previously-visited MAG's proxy-CoA. Then, the proxy binding update (PBU) and the proxy binding acknowledgement (PBA) messages are exchanged between MAG1 and MN1's all

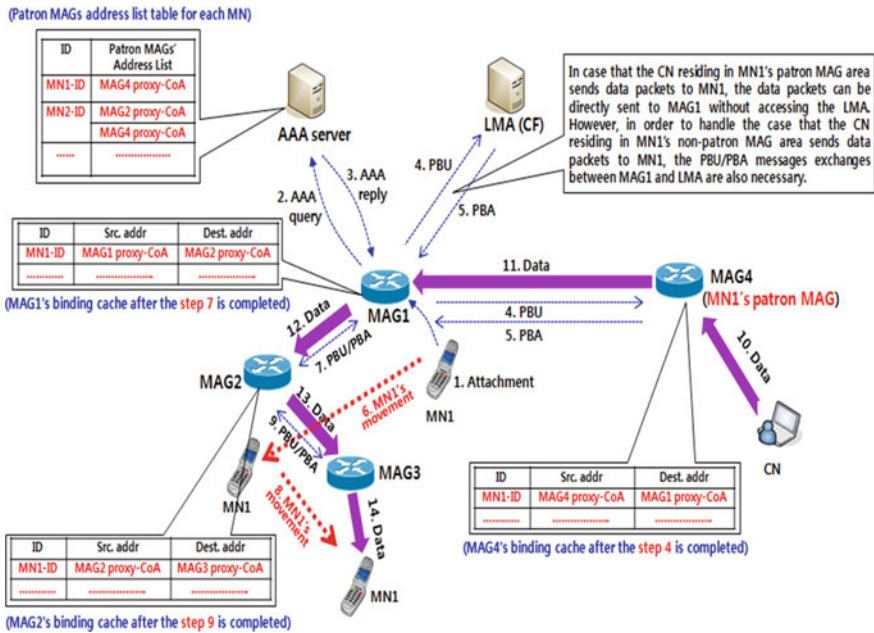


Fig. 1 The illustrative example of EFS-PMIP

patron MAGs (i.e., MAG4) as well as between MAG1 and LMA (Steps 4–5). Then, when the MN1 attaches to MAG2 coverage area (Step 6), after the AAA query/reply message exchanges between MAG2 and AAA server are performed just in the same manner as mentioned in Steps 2–3, the tunnel originating from the MAG1’s proxy-CoA is setup by PBU/PBA message exchanges between MAG1 and MAG2 (Step 7). Similarly, when the MN1 attaches to MAG3 coverage area (Step 8), the tunnel originating from the MAG2’s proxy-CoA is also setup by PBU/PBA message exchanges between MAG2 and MAG3 (Step 9). For simplicity, the AAA query/reply message exchanges occurred when the MN1 enters the MAG2 and MAG3 coverage areas are not drawn in Fig. 1, and these procedures follow the same manners in Steps 2–3 in Fig. 1.

Now, the packet delivery procedures are as follows. As shown in Fig. 1, the CN residing in the MN1’s patron MAG (i.e., MAG4) area sends data packets to the MN1 (Step 10). Then, MAG4 searches the binding cache in order to find if there is any binding cache entry for MN1 or not. In this example, MAG4 can find the binding cache entry for MN1 because MAG4 has been already informed of MN1’s anchor MAG (i.e., MAG1) (In this paper, we call the MN1’s MAG which has last been registered with the LMA as the MN1’s “anchor MAG”). Therefore, MAG4 sends the data packets directly to the MAG1 without the access to LMA after finding the MN1’s anchor MAG address (i.e., MAG1’s proxy-CoA) (Step 11). Then, as shown in Fig. 1, the data packets are delivered from MAG1 to MAG2,

and from MAG2 to MAG3 through the lookup processes of the MN1's binding caches at MAG1 and MAG2 (Steps 12–13). Finally, the MAG3 forwards the data packets to the MN1 (Step 14).

3 Discussions

In terms of *binding update* procedures, EFS-PMIP basically follows almost the same binding update procedures as in our previous work, FS-PMIP [4]. On the other hand, in terms of *packet delivery* procedures, EFS-PMIP doesn't require the access to LMA in case of the data packets sent from each MN's patron MAGs while FS-PMIP requires the access to LMA whenever the CN sends the data packets to the MN. Therefore, even if the performance evaluation on EFS-PMIP is currently being in progress and not completed yet, we can easily expect that EFS-PMIP outperforms FS-PMIP, S-PMIP and PMIPv6 in most situations. For performance evaluation, we will analyze the cost functions of EFS-PMIP, FS-PMIP, S-PMIP, and PMIPv6, respectively, and compare the performance of each scheme. In addition, as a future work, we will thoroughly investigate the impacts of the predefined-threshold K , the number of the MN's patron MAGs, and the MN's session-to-mobility (SMR) ratio, which may have significant effects on the performance of EFS-PMIP.

From the perspective of a MN, the working set of MAGs indicating strong traffic locality towards a certain MN can be selected as its patron MAGs. More specifically, in order to determine the patron MAGs for a particular MN, the MN's mobility and traffic patterns throughout the days or weeks may be observed over a long period of time. For example, the MN's LMA may measure the number of session arrivals and MAG area crossings in order to calculate its session-to-mobility ratio (SMR) at every MAG area that the MN resides, which implies that the larger the SMR of an MN is at specific MAG area, the more outgoing sessions for the MN can occur at that MAG area. In [5], the authors made the key observation that while the potential set of sources for the MN may be large, the set of sources that a given MN communicates most frequently with is "very small". Also, in [7, 8], a large number of actual packet data call traces in CDMA2000 cellular data network have been examined to characterize a user's mobility/traffic features. It was observed that (i) about 80 % of the mobile users visit fewer than 20 cells [7], and (ii) "a few cells" have a high level of packet call activity, while many cells have a relatively low level of activity [8]. Therefore, based on the above observation and the actual trace data in [5, 7] and [8], choosing only a few MAGs as each MN's patron MAGs would be sufficient. Otherwise, patron MAGs can be manually preconfigured in the MN's profile at AAA server according to the MN's willingness or the network service provider's policy to get a higher quality of service.

How to manage each MN's profile and how to determine its patron MAGs is an important research issue by itself [7]. Thus, our future research directions are also planned to conduct more in-depth study on these research issues.

4 Conclusions

Recently, the distributed mobility management technology is being actively discussed in IETF DMM working group, and it has much potential to overcome many limitations of centralized mobility management. In this paper, in order to support more enhanced PMIPv6-based distributed mobility management, we proposed EFS-PMIP that exploits a user's traffic locality as well as movement locality, which is inspired by the concept of working set in communications. By effectively distributing both binding functionality and routing functionality at the MAGs, the proposed EFS-PMIP is expected to significantly alleviate the unsolved drawbacks of FS-PMIP, S-PMIP, and PMIPv6.

Acknowledgments Funding for this paper was provided by Namseoul university.

References

1. Kong K-S et al (2008) Mobility management for All-IP mobile networks: Mobile IPv6 vs. Proxy Mobile IPv6. *IEEE Wirel Commun* 15(2):36–45
2. Yokota H, KDDI Lab, Seite P (2010) Use case scenarios for distributed mobility management. IETF internet-draft, draft-yokota-dmm-scenario-00.txt, 18 Oct 2010
3. Jung H et al (2011) Distributed mobility control in Proxy Mobile IPv6 networks. *IEICE Trans Commun F94-B(8):2216–2224*
4. Kong K-S A pointer forwarding-based signal-driven Proxy Mobile IPv6 for distributed mobility management. *Inf Int Interdiscip J* (to appear)
5. Rajagopalan S, Badrinath B (1995) An adaptive location management strategy for Mobile IP. In: *Proceeding of ACM MOBICOM'95*
6. Jain R et al (1995) An auxiliary user location strategy employing forwarding pointers to reduce network impacts of PCS. *Wirel Netw* 1(2):197–210
7. Zang H, Bolot J (2007) Mining call and mobility data to improve paging efficiency in cellular networks. In: *Proceeding of ACM MOBICOM'07*, pp 123–134
8. Williamson C, Halepovic E, Sun H, Wu Y (2005) Characterization of CDMA2000 cellular data network traffic. In: *Proceeding of IEEE LCN'05*, pp 712–719

A Multihoming-Based Vertical Handover Scheme

Hee-Dong Park and Kyung-Nam Park

Abstract This paper proposes a multihoming-based vertical handover scheme. Mobile nodes must have several radio interfaces to support vertical handover in heterogeneous wireless networks. In this paper, we consider each interface of a mobile node has its own protocol stack with physical, data link, and network layer. When a mobile node moves to a different type of access network, the proposed scheme can provide a mobile node with fast and seamless handover by performing layer-3 handover using its new interface while the other interface is still communicating in the old access network. This scheme uses a newly defined Proxy binding update to minimize handover delay and packet loss while a mobile node moves to a different type of access network. The proposed Proxy binding update is different from the Mobile IP binding update in that it includes home address (HoA) of the old interface instead of the new interface. The performance analysis shows that the proposed scheme can efficiently reduce vertical handover delay and packet loss.

Keywords Multihoming · Vertical handover · Proxy binding update · Seamless handover

H.-D. Park

Department of Information and Communication, Korea Nazarene University, 456,
Ssangyong-dong, Seobuk-gu, Cheonan-city, South Korea
e-mail: hdpark@kornu.ac.kr

K.-N. Park (✉)

Department of Multimedia, Korea Nazarene University, 456, Ssangyong-dong,
Seobuk-gu, Cheonan-city, South Korea
e-mail: knpark@kornu.ac.kr

1 Introduction

With the advent of diverse wireless networks and smart mobile devices, vertical handover is a key issue to support seamless communications while a mobile node moves to a different type of access network. Heterogeneous networks based on wireless local area networks (WLANs) and wireless wide area networks (WWANs) such as 3G/4G, beyond 3G/4G, and WiMAX/Wibro can combine their respective advantages on coverage and data rates. For example, WWANs such as UMTS 3G cellular networks have big coverage and low data rates, while WLANs have small coverage and high data rates. In such environment, multi-interface terminals should seamlessly switch from one access network to another in order to obtain improved performance and continuous wireless connection.

Many researches on vertical handover are being carried out and standardized in various standard organizations such as IETF, IEEE, ITU-T, 3GPP, etc. [1–6]. Especially in IETF, many IP-based mobile technologies such as Mobile IP, Fast Mobile IP, Proxy Mobile IP, etc. are considered to support vertical handover in the heterogeneous networks.

This paper proposes a multihoming-based seamless vertical handover scheme with newly defined Proxy binding update which is different from the original Binding update of Mobile IP.

2 Multihoming-Based Vertical Handover With Proxy Binding Update

2.1 Protocol Stack of a Mobile Node

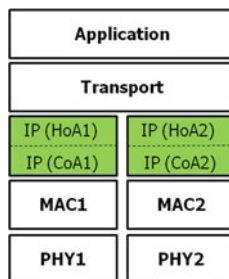
Mobile nodes must have several radio interfaces to support vertical handover in heterogeneous wireless networks. In this paper, we consider a mobile node with multiple interfaces corresponding to their access networks. Each interface has its own protocol stack with physical, data link, and network layer.

Figure 1 shows an example of protocol stack of a mobile node in the proposed scheme. The mobile node has dual interfaces which have dual PHYs, MACs, and

Table 1 Parameters for performance evaluation

Parameters	Definition	Value
D_{HO}	Total handover delay	N/A
τ	The interval of Router advertisements	1 s
D_{L2}	Layer 2 handover delay	100 ms
L_{wl}	Wireless link delay	2 ms
L_w	Wired link delay	0.5 ms
D_{DAD}	Times for duplicate address detection	500 ms
H_{AR-HA}	Hop-counts between AR and HA	–

Fig. 1 Protocol stack of a mobile node



IPs. Therefore, each interface has its own home address (HoA) and care-of address (CoA). We consider each interface of a mobile node can belong to its own access network. For example, an interface (INT-1) with WLAN has its own HoA1 and CoA1, and the other (INT-2) with 3G or 4G network has HoA2 and CoA2.

2.2 Handover Procedure

Figure 2 shows the vertical handover procedure of the proposed scheme. We consider a scenario where a mobile node moves from WLAN to 3G or 4G network. AR (Access router) in Fig. 2 can be replaced by a corresponding entity according to access networks. For example, SGSN (Serving GPRS support node) in UMTS performs the role of AR.

- (a) Phase 1: When a mobile node stays in the WLAN, it communicates with its home agent (HA) or correspondent node (CN) using INT-1. In Phase 1, the binding information of HA is HoA1 to CoA1. Therefore, HA intercepts the packets destined for the mobile node and tunnels them to CoA1.
- (b) Phase 2: As the mobile node moves to the duplicate area between WLAN and 3G/4G network, its MIH module triggers vertical handover. In Phase 2, INT-2 receives the prefix information from New_AR in the 3G/4G network. After the mobile node associates with the New_AR by creating a CoA (CoA2), it sends a Proxy binding update message to the HA. The Proxy binding update message contains CoA2 of INT-2 and HoA1 of INT-1 instead of HoA2 of INT-2. This makes the HA to be under the illusion that the INT-1 has moved to a new access network. The INT-1, however, actually continues to receive packets in the Old_AR's coverage area (WLAN), thus packet loss can be prevented. After receiving the Proxy binding update message, HA updates the binding information and sends Proxy binding update ACK message and data packets to the mobile node through 3G/4G network. In summary, the mobile node can send and receive data packets through INT-2 in 3G/4G networks while it may also receive in-transit packets through the INT-1 in WLAN.
- (c) Phase 3: After the completion of Proxy binding update, the mobile node can communicate with HA or CN through INT-2.

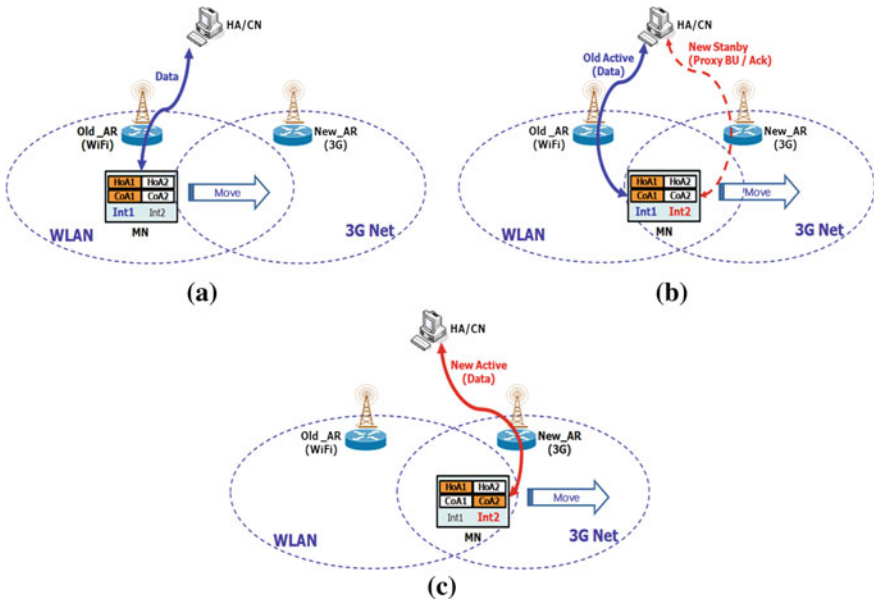


Fig. 2 Handover procedure of the proposed scheme: **a** Phase 1, **b** Phase 2, **c** Phase 3

In the proposed scheme, Proxy binding update and Proxy binding ACK messages are introduced to support seamless vertical handover. The formats of these messages, however, are the same as those of general Binding update and Binding update ACK messages in Mobile IPv6. Yet, the proposed Proxy binding update message is different from Proxy binding update message of Proxy Mobile IPv6.

3 Performance Evaluation

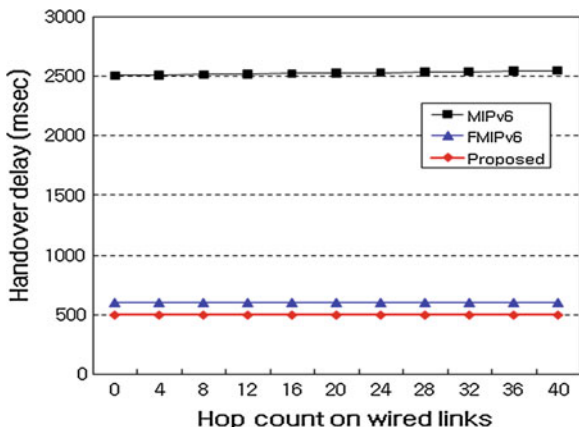
This section compares MIPv6, FMIPv6, and the proposed scheme by vertical handover delay and packet loss ratio [7].

3.1 Vertical Handover Delay

Vertical handover delay can be defined as the time duration between the start of layer 2 handover and reception of the first packet in a different access network after layer 3 binding update. In this paper, link delay is considered, but processing and transmission delay are not.

Handover delays of MIPv6, FMIPv6, and the proposed scheme can be represented as Eqs. (1), (2), and (3), respectively.

Fig. 3 Vertical handover delay



$$D_{HO(MIP)} = 2\tau + 4L_{wl} + 2L_w H_{AR-HA} + D_{DAD} \tag{1}$$

$$D_{HO(FMIP)} = D_{L2} + L_{wl} + D_{DAD} \tag{2}$$

$$D_{HO(Proposed)} = D_{DAD} \tag{3}$$

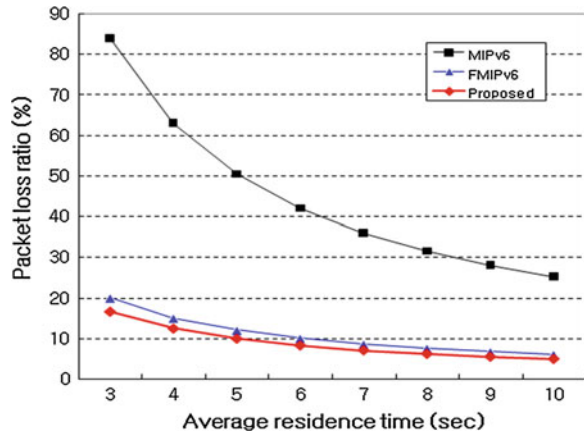
As shown in Eq. (3), handover delay of the proposed scheme is the lowest because the proposed handover is a type of ‘Make-before-break’ handover in network layer. Figure 3 compares their handover delays according to hop-counts between AR and HA. Handover delay of MIPv6 is proportional to wired link delay but handover delays of FMIPv6 and the proposed scheme are constant regardless of wired link delay.

3.2 Packet Loss Ratio

Packet loss ratio is defined as the ratio of the number of lost packets during a handover to the total numbers of transmission packets in a cell. This can be also expressed as:

$$\rho_{loss} = \frac{D_{HO}}{t_{cell}} \times 100 \quad [\%] \tag{4}$$

where t_{cell} is the average residence time in a cell. Figure 4 compares packet loss ratio during a vertical handover. As shown in Fig. 4, the packet loss ratio of the proposed handover scheme is the lowest because its handover delay is the lowest.

Fig. 4 Packet loss ratio

4 Conclusions

This paper proposes a multihoming-based vertical handover scheme with Proxy binding update. When a mobile node is located in the duplicate region between two different access networks, it can send and receive data packets using multiple interfaces simultaneously, which results in a seamless handover. The proposed scheme supports ‘Make-before-break’ handover in network layer with Proxy binding update messages.

Acknowledgments Corresponding author is Kyung-Nam Park.

References

1. Perkins C, Johnson D, Arkko J (2011) Mobility support in IPv6. IETF RFC 6275
2. Koodli R (2009) Mobile IPv6 fast handovers. IETF RFC 5568
3. Gundavelli S, Keung K, Devarapalli V, Chowdhury K, Patil B (2008) Proxy Mobile IPv6. IETF RFC 5213
4. IEEE 802.21 MIH WG D13 (2008) Media independent handover services
5. ITU-T Q.1706/Y.2801 (2006) Mobility management requirements for NGN
6. 3GPP TS 23.234 v7.6.0 (2007) 3GPP system to wireless local area network (WLAN) interworking; system description
7. Zhang X, Castellanos JG, Gapbell AT (2002) P-MIP: paging extensions for Mobile IP. ACM Mob Netw Appl 7(2):127–141

A Performance Prediction Model of Parallel DCT on Mobile Embedded Systems

Yeong-Kyu Lim and Cheong-Ghil Kim

Abstract The recent development on semiconductor process and design technologies enables multi-core processors to become a dominant market trend in mobile devices. The parallel programming enabled by multi-core CPU can provide a great opportunity to increase the processing performance. This paper explores a performance prediction model of parallel DCT on heterogeneous mobile systems by measuring power dissipation. For our simulation, we implemented the fast DCT algorithm on various computing environments and the simulation results show the feasibility of the proposed method to estimate the performance gain in terms of power consumption on heterogeneous embedded systems.

Keywords JPEG DCT · Mobile embedded systems · Multi-core · Parallel programming · Power dissipation · Heterogeneous computing

1 Introduction

As the wide spreading smart phones, the complexity of the smart phones software stack consisting of operating system, application framework, and applications are growing. Therefore, they must be matched by adequate performance increases of the underlying hardware. Single-core performance scaling is coming to an end

Y.-K. Lim
MC Division, LG Electronics, Seoul, Korea
e-mail: postrain@yonsei.ac.kr

C.-G. Kim (✉)
Department of Computer Science, Namseoul University, Cheonan, Korea
e-mail: cgkim@nsu.ac.kr

because fundamental laws of physics limit further performance gains from uniprocessor architectures [1]. Therefore, all future computing platforms will provide multiple cores. The good news for embedded systems is that multi-core processors are more energy-efficient, i.e., their FLOPS/Watt ratio is higher than uniprocessors [2–4].

Unfortunately, by now it has become a widely acknowledged fact that creating software for multi-core architectures is an extremely complex task that is hampered by the following key issues: (1) we are facing huge stacks of historically-grown legacy software, which has been developed over the course of decades with uniprocessors in mind. Retro-fitting this legacy software onto multi-core architectures is extremely complex and time-consuming, and it is to be decided on a case-by-case basis whether performance-critical parts should be parallelized or re-developed from scratch the multi-core way. At the same time, we are still waiting for the new generation of programmers to arrive, who deeply penetrate the wisdom of multi-core programming. (2) programming languages and tools are not ready yet. The IT industry is steering into a software crisis called the Parallel Programming Gap, which is caused by the growing divide between the capabilities of today's software tools and the complexity of multi-core architectures and applications. (3) hardware architectures are not fully matured yet. We know that future multi-core architectures will be heterogeneous, i.e., consisting of several different types of processors, but the final configuration is still under investigation.

This paper is structured as follows. [Section 2](#) introduces main parallel programming trends and [Sect. 3](#) conducts a feasibility study to determine the most advantageous mobile embedded architecture in terms of cost, power consumption, and performance. [Section 4](#) is devoted to our parallelization of DCT for multi-core CPUs and GPGPUs, and the integration of our parallel DCT with the JPEG library of Webkit. Finally, [Sect. 6](#) covers our conclusion.

2 Multi-Core Programming

This section introduces several task- and data-parallel solutions available today. Cilk is a multi-core programming model based on simple C extensions and a complex work-stealing scheduler. TBB is a collection of various C/C++ library functions and templates to create software that can run efficiently on multi-core HWs. TBL is Microsoft's counterpart to TBB for .NET framework. It is mainly for CLR-based languages such as C# or Visual Basic. GCD is Apple's solution to the multi-core programming problem. GCD is available as open source. GCD employs blocks, which encapsulate a portion of code that can be executed by the task dispatcher. Blocks are defined by the programmer, as extensions of C and C++. At run-time, blocks are entered into one of the dispatch queues of the dispatcher, where they will be distributed among the available processing cores. Blocks are supported by some branches of GCC, and by Apple's clang compiler framework [5].

OpenMP is a standardized programming model for shared-memory multiprocessing. It is very widely spread, and OpenMP is provided for Fortran, C, and C++. OpenCL was initially promoted by Apple. Meanwhile, it has been adopted by all major HW vendors. It is based on C, but it excludes support for several features, e.g., function pointers, recursion, variable length arrays and bitfields. Computation is encapsulated in kernels, execution happens on device context [5].

3 DCT

Discrete Cosine Transform (DCT) is adopted by various applications such as MP3 audio compression and JPEG image compression. The primitive operation of DCT is based on matrix computations in which parallel processing techniques must be considered for real time processing with large 2D data sets. This operation is characterized as data intensive tasks accompanied by heavy memory access; on the other hand, their computational complexities are relatively low. Thus, it naturally maps onto massively parallel architecture with distributed memory.

The 2-dimensional DCT algorithms are most computational intensive part of JPEG encoding. There are many fast 2D-DCT algorithms already studied. We implemented several algorithms and estimated its runtime on multi-core CPU and GPU environments. Experiments show that data parallelism can be fully exploited on CPU and GPU architecture. After that we integrated GPU empowered DCT with JPEG library. Data structure and procedure routines in JPEG library are quite optimized well on sequential processor because of performance issues. It computes DCT one row of blocks in image each step and it is not very good idea on multi-processor programming. Especially GPU has huge data transfer overhead so it is recommended to deliver all data to GPU [6].

4 Simulations

As for heterogeneous embedded architecture, it may consist of ARM A9 multi-core [7]. And we are considering possible configurations of CPUs/GPGUs for cost, power-consumption, and performance. To test feasibility, we try to predict possible power consumption and performance at high level approach because there is no openCL enabled embedded device. Our experimental environment is configured with three different systems: desktop PC, single-core embedded hardware, and multi-core embedded hardware. Details are summarized in Table 1 and Fig. 1 shows the power measuring environments in series. The simulation results are shown in Table 2. If we take same current and power usage of OpenCL benchmark, dual-core embedded hardware's energy consumption becomes 0.198 J, which is $776 \times$ more energy efficient.

Table 1 Hardware configuration

Desktop PC			
Hardware		Software	
CPU	Intel Xeon Quad Core 3.07 GHz	OS	Ubuntu Linux 10.10 Linux Kernel : 2.6.35
GPU	NVIDIA Quadro 60, 1 GHz, 96 Cuda cores Maximum power consumption : 40 W GPU Memory Specs: 1 GB DDR3, 128-bit, 25.6 GB/s		
RAM	4 GB	File system	EXT4
Bus speed	1.33 GHz	Compiler	GCC 4.x
Single-core embedded hardware			
Hardware		Software	
CPU	Cortex-A8 Single-Core	OS	Android 2.2
GPU	PowerVR SGX 540		
RAM	More than 128 MB	File system	NFS
Bus speed		Compiler	GCC 4.x
Multi-core embedded hardware			
Hardware		Software	
CPU	Cortex-A9 Dual-core	OS	Android 2.3
GPU	Mali 400		
RAM	More than 128 MB	File system	FAT/EXT4 (SD/MMC)
Bus speed		Compiler	GCC 4.x

**Fig. 1** Power measurement configuration

When we apply same current and power of OpenCL benchmark, we can get 0.198 J energy consumption and which is significantly large saving. Although Serial DCT and OpenCL DCT are ideal case for exploiting parallelism, we can guess heterogeneous architecture's possibility.

According to profile result, data transfer spends about $3 \times$ more time than kernel computation does. In desktop environment, communication between CPU and GPU is done via PCI channel which is quite slower than accessing main memory. In embedded system, CPU and GPU are usually on same ship and share memory so that we can mitigate data transfer overhead. We can do some experiments on this using AMD's APU because currently no embedded hardware supports openCL and any other kinds of general computing on GPU.

Also we can speed up current version of JPEG library on desktop hardware environment which has huge data transfer overhead. Current graphic hardware

Table 2 Simulations on various hardware

Benchmark	Idle	Serial DCT	OpenCL DCT	OpneCL benchmark
Desktop PC				
Current(A) $I = V/R$	2.2	4.2	4.4	6.2
Power(W)	26.2	48.6	48.64	69.22
Execution time(s)		5.625	0.0027	0.0074
Energy(J)		273.38	1.31	0.512
S5PV210 (Single-core)				
Current (A) ($I = V/R$)	0.42	0.6		0.68
Power(W)	2.01	2.76		3.02
Execution time(s)		59.356		0.071
Energy(J)		163.82		0.211
S5PV310				
Current (A) ($I = V/R$)	1.08	1.36		1.98
Power(W)	5.03	6.15		8.34
Execution time(S)		24.998		0.04
Energy(J)		153.67		0.33

support 2 simultaneous data transfer. When we apply pipeline data transfer, we can reduce data communication overhead to half. We pipeline data write, kernel computation and data read. Like our case, if kernel computation is smaller than data transfer, we can add more computation for free like quantization step in JPEG library. Also CPU computing resource can be exploited during GPU computation.

5 Conclusions

This paper explores a performance prediction model of parallel DCT on heterogeneous mobile systems by measuring power dissipation. For our simulation, we implemented the fast DCT algorithm on various computing environments. The evaluation results show the feasibility of the proposed method to estimate the performance gain in terms of power consumption on heterogeneous embedded architecture.

Acknowledgments This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research and Development Program 2012.

References

1. Suter H (2005) The free lunch is over: a fundamental turn towards concurrency in software. *Dr. Dobbs J* 30(3):1–7
2. Asanovic K, Bodik R, Catanzaro BC, Gebis JJ, Husbands P, Keutzer K, Patterson DA, Plishker WL, Shalf J, Williams SW, Yelick KA (2009) A view of the parallel computing landscape. *Commun ACM* 52(10):56–67

3. Rousseeuw PJ, Leroy AM (1987) Robust regression and outlier detection. Wiley, pp 134–150
4. Kornaros G (2010) Multi-core embedded systems, CRC Press, Boca roton
5. Kim CG, Choi YS (2012) A high performance parallel DCT with openCL on heterogeneous computing environment. Multimedia Tools Appl DOI [10.1007/s11042-012-1028-x](https://doi.org/10.1007/s11042-012-1028-x)
6. Hong JG, Wook JS, Kim CG, Burgstaller B (2011) Accelerating 2D DCT in multi-core and many-core environments. In: Proceedings of the 35th conference of Korea information processing society, May 2011
7. Leskela J, Nikula J, Salmela M (2009) OpenCL embedded profile prototype in mobile device. IEEE workshop on signal processing systems, pp 279–284, Oct 2009

Implementation of a Low Cost Home Energy Saver Based on OpenWrt

Se-Hwan Park, Hyun-Jun Shin, Myoung-Seo Kim
and Cheong-Ghil Kim

Abstract This paper introduces a system that can effectively save home energy by applying a small embedded system through remote control. We used a wireless router based on OpenWrt for the platform to develop an embedded system and a smart phone for the remote LED light control. The system was implemented by connecting a wireless router with an OS of OpenWrt installed and an interface board with an LED attached. The smart phone, which was the remote control device, was implemented by TCP/IP programming. The operation of the remote control system was verified by socket communication between the smart phone and the wireless router, and by USB communication between the wireless router and the interface board.

Keywords OpenWrt · Wireless router · Energy saving · Smart phone

S.-H. Park
ReSEAT Program, Korea Institute of Science and Technology Information,
Daejeon, Korea
e-mail: world00117@reseat.re.kr

H.-J. Shin · C.-G. Kim (✉)
Department of Computer Science, Namseoul University, Cheonan, Korea
e-mail: cgkim@nsu.ac.kr

H.-J. Shin
e-mail: fs_developer@naver.com

M.-S. Kim
Department of EECS, The Henry Samueli School of Engineering,
University of California, Irvine, CA, USA
e-mail: myoungseo.kim@uci.edu

1 Introduction

Nowadays, wireless networks are everywhere and even in home 802.11 wireless networks are popular, avoiding wiring costs and providing connectivity for all rooms. Therefore, we can find low cost wireless router easily and they can be utilized for low cost embedded Linux platform using OpenWrt which targeted at the Linksys WRTG54 initially, but now targets many embedded wireless devices including equipment from Asus, D-Link, NetGear, Soekris, Viewsonic, and Linksys [1]. Primarily, OpenWrt is an operating system used on embedded devices to route network traffic. The main components are the Linux kernel, uClibc and BusyBox. All components have been optimized for size, to be small enough to fit the limited storage and memory available in home routers [2].

In this paper, we take advantage of the wireless router with OpenWrt as the home energy saver with remote control feature. Over the years, as the number of electric and electronics devices increase dramatically at home and in buildings, it was difficult to manage the waste of energy due to the inefficient light control and illumination distribution [3]. In addition, it is not practical to rely on users to directly control the light switch to save energy. To this end, this paper introduces a remote LED control system which is based on OpenWrt. The system was implemented by porting a wireless router to OpenWrt and by connecting an interface board to an LED. Also, the system communicated with a smart phone via socket communication to control the LED.

This paper includes: introduction in [Sect. 1](#), system components and operation in [Sect. 2](#), experimental work and simulation in [Sect. 3](#), and the conclusion in [Sect. 4](#).

2 System Components and Operation

The system comprises of a smart phone to control a wireless router on OpenWrt, an interface board and LEDs, and a wireless router to communicate with the board. Here, the router is DIR-825 wireless router of DLink and acts as a CPU board to control and manage the system. The smart phone, iPhone4 of Apple, communicates with the wireless router and allows the user interface to control the LED remotely through the communication with the interface board installed in its USB. For the interface board, we use Arduino-UNO of Arduino Lab. We used DIR-825 wireless router of DLink and 5 ohm red LEDs. [Figure 1](#) illustrates the overall structure of the system with the functionality of controlling the brightness of LEDs using iPhone remotely.

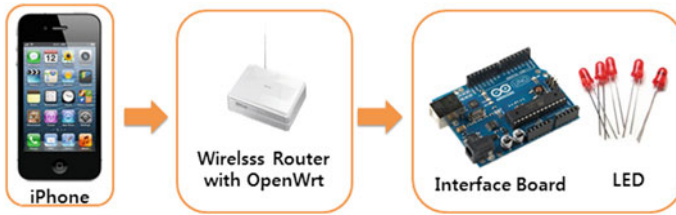


Fig. 1 System operating flow

3 Experimental Work and Simulation

First of all, we installed OpenWrt in the router and then KMOD-USB-ACM package. Next, we connected it to Arduino in order to prepare communication between the board and router. The user interface (UI) was applied to the smart phone and operated to enable socket communication through TCP/IP protocol by using programmed iOS SDK [2]. Then, we used the Library of C and Linux for socket communication by TCP/IP protocol [4], and set the Baud ratio of the server program for control data transfer by USB port. We selected a USB port and used a File Descriptor to send data by 1 byte each [5], so that it fits to the OpenWrt system by doing cross compile. The interface board of Arduino received data by each byte from the router through USB port. We programmed the interface board to deliver the control signal to the LED when all data was received [6]. Finally, we controlled the brightness of the LED by the MOSFET switch that depends on the control signal [7].

Figure 2 is the picture of the overall structure of the system. We created a breadboard with LED and then connected it to the main board with the USB port. We ran the OpenWrt-based software and then inserted command for starting as in

Fig. 2 Overall structure



```
BusyBox v1.15.3 (2011-11-24 00:44:20 CET) built-in shell (ash)
Enter 'help' for a list of built-in commands.

-----
|_| W I R E L E S S F R E E D O M
-----
Backfire (10.03.1, r29592) -----
* 1/3 shot Kahlua      In a shot glass, layer Kahlua
* 1/3 shot Bailey's   on the bottom, then Bailey's,
* 1/3 shot Vodka       then Vodka.
-----

root@OpenWrt:~# cd /shj
root@OpenWrt:/shj# ls
serial_Target
root@OpenWrt:/shj# ./serial_Target 50000
size : 3
ARDUINO Port /dev/ttyACM0 opened, waiting for board to boot up
Please Connect
Connected Client 192.168.1.124:60448
): 1 D : 50 S : 1
): 1 D : 0 S : 1
): 1 D : 50 S : 1
): 1 D : 0 S : 1
): 1 D : 50 S : 1
): 1 D : 50 S : 1
): 1 D : 58 S : 1
```

Fig. 3 Program start

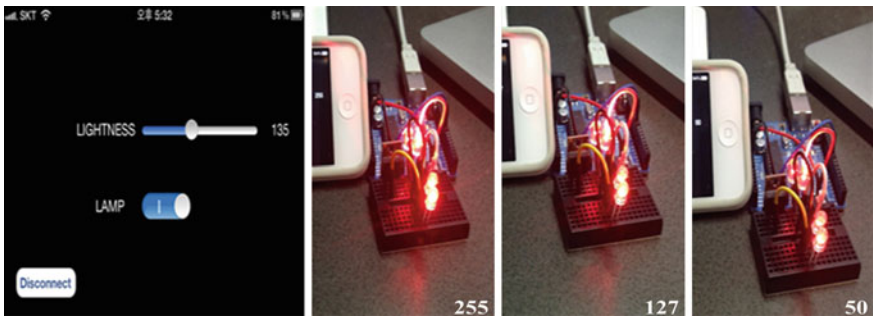


Fig. 4 Light control through the user interface

Fig. 3. As a result, the smart phone became ready to communicate with the wireless router.

We loaded the UI, and connected the wireless router and TCP socket communication from the smart phone as shown in Fig. 4. Then we controlled the

brightness of LEDs by data transfer. The brightness is controllable from light (255) to dark (50). Moreover, after changing the firewall settings of the router to allow external router access, not only local wireless router, but also control of 3G internet connection from external network became possible.

4 Conclusion

We implemented a system that can control LED by using an interface board which was connected to a wireless router ported with OpenWrt by a smart phone. This enabled socket communication and link with peripheral device using a real wireless router, and utilization of a wireless router as an embedded system. Another application of the results of this paper may include control of remote household automation system via internet. If such system is utilized, it is expected that power consumption will be reduced, since we can turn off the lights in our homes via the internet even when we are on the opposite side of the globe. Furthermore, this system can contribute to saving power by enabling individuals to control the brightness of unnecessarily bright LEDs in a more customized fashion.

Acknowledgments This study has been done by the 'ReSEAT program 2012' that is Korea Institute of Science and Technology Information (KISTI) and Ministry of education, science and technology (MEST) as science & technology promotion fund.

References

1. Murray J (2009) An inexpensive wireless IDS using kismet and openWRT. SANS Institute. Available at www.sans.org
2. OpenWrt. <http://en.wikipedia.org/wiki/OpenWrt>
3. Lee J (2011) IEEE1451 interface for smart grid. Hanyang University Graduate School, Aug 2011
4. Mark D, Nutting J, LaMarche J (2010) Beginning iPhone 5 development: exploring the iOS 5 SDK. Wikibooks, pp 259–366
5. Yoon S (2007) TCP/IP programming. Orangemedia, pp 110–112
6. UNO serial latency. <http://arduino.cc/forum/index.php?topic=96.0>
7. Arduino MOTOR/LED control. <http://arduino.cc/playground/MotorControlShieldV3/0>

Part VII
Multimedia and
Information Visualization

A Multimedia Authoring and Virtual Collaboration System Supporting Multi-Conferences for E-Learning

Yeongjoon Kim and Chuleui Hong

Abstract We are proposing the multimedia presentation authoring and virtual collaboration system that produces multimedia contents and enables multiple users who are dispersed in locally and timely to collaborate by the Internet. The proposed system consists of two parts—multimedia presentation authoring and recordable virtual collaboration tools. The authoring tool makes it possible to create and edit multimedia presentations that integrate diverse media types including images, video, sound, and texts for e-learning. Media objects are synchronized with the temporal and spatial information using SMIL defined by W3C. The collaboration tool categorizes users by the interested conference and any users can create new topics or join the existing topics by validating user's access right freely in a conference they belong. Users can participate in more than one topic simultaneously, so they can inquire and get a valuable knowledge on-line in one topic and participate in the other topic more informed and intelligent. The produced multimedia presentation may be provided to the users through the conference before the discussion begins. Users can use text along with associated symbols such as arrows and polygons over the presented images during the discussion. Users' opinions along with symbols are recorded to XML database.

Keywords Virtual collaboration • Multi-conferences • Multimedia authoring

Y. Kim · C. Hong (✉)

School of Software, Sangmyung University, 7, Hongji-dong, Chongro-gu,
Seoul, South Korea
e-mail: hongch@smu.ac.kr

Y. Kim

e-mail: yjkim@smu.ac.kr

1 Introduction

With the phenomenal growth in the Internet usage across the world, e-learning is expected to go as a powerful educational tool and virtual collaboration can be an efficient way for sharing information through the Internet in real time. SMIL (Synchronized Multimedia Integration Language) [1] is a XML based markup language by W3C recommendation to create multimedia presentations. SMIL integrates multimedia elements such as text, images, video and audio by synchronizing them. Bouyakoub et al. [2] have proposed a temporal authoring tool for SMIL document with incremental authoring based on H-SMIL-Net model. Téllez [3] has presented an IMS formats authoring tool for Docbook to produce SMIL compositions and e-learning contents. Pooshfam et al. [4] proposed a system for annotating images and videos in a collaborative way.

In this paper, we are proposing a multimedia authoring and virtual collaboration system that produces multimedia presentations for e-learning contents and enables multiple users to do collaborative work. The proposed system consists of two separate tools: one is a multimedia presentation authoring tool and the other is a virtual collaboration tool.

The authoring tool helps authors to make multimedia presentations by themselves and easy way and makes it possible temporal and spatial synchronization of integrated diverse media types using SMIL. The virtual collaboration tool categorizes users by expert area to make a conference. The uniqueness about the proposed collaboration system is the multimedia presentation authoring and the concept of layered architecture that is used to annotate digital images with various symbols such as arrows or polygons.

2 Multimedia Authoring System

The proposed multimedia presentation authoring system is composed of the following modules. The GUI of the system enables authors to use these modules without any knowledge for authoring e-learning contents. The system architecture is shown in Fig. 1.

Creating Module facilitates creating a new presentation. In order to initialize authoring a new presentation, it creates the presentation folder specified and named by an author, and load images and video or ready to record video. This module converts Power point slides to JPEG images and copy them from selected folder to the presentation folder. It specifies the layout that defines the regions of media objects.

Recording/Playing Module records the video and audio during a presentation and enables the author to view the video screen being currently recorded during the recording progress by using JMF. Existing video and audio can be played and showed by this module.

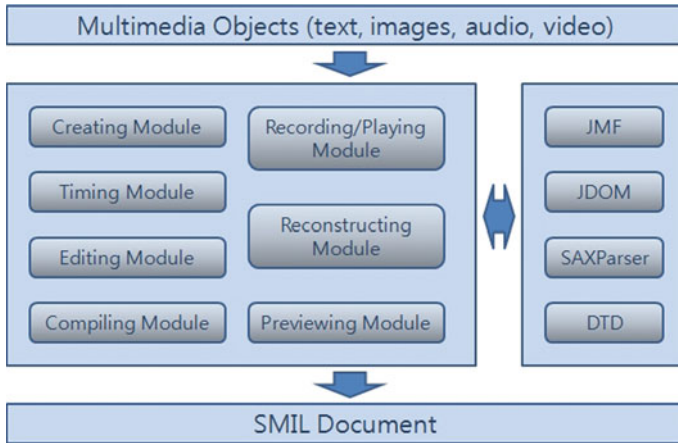


Fig. 1 System architecture of multimedia authoring tool

Timing Module enables to specify the time duration of a presentation with time stamp by author’s command. When an author starts and finishes authoring a presentation, it records the beginning and end time with time stamp for the video and audio and calculates the total time duration for the presentation. When author’s command is occurred for images, this module specifies the beginning time and time duration of each image, and uses that information to synchronize media objects.

Compiling Module corrects the temporal and spatial information about media objects from other modules and compiles it. This module provides the functionality enabling to generate a SMIL for synchronizing media objects and makes it possible to create a SMIL document without any knowledge of SMIL language for authors by hiding the source file of the SMIL document through the GUI. SAX-Parser and JDOM have been used for implementing this module.

Editing Module enables authors to edit existing presentations. This module allows an author to change the slide images in existing presentations to different slides. It offers very useful editing function that involves inserting images, text, audio, video and existing presentation and deleting slide with video or slide only or video only by using the cutting and reconnecting algorithm. When this module is activated, it loads SMIL document to be updated. Each editing operation by author’s command is reflected immediately to SMIL document. SAXParser and JDOM have been used for implementing this module.

Previewing Module provides the functionality to preview and check the temporal presentation during creation before editing or playing with a media player. It displays the image slide thumbnails in the temporal presentation through the navigation pane that helps the author to navigate the image objects and easy to move in the presentation.

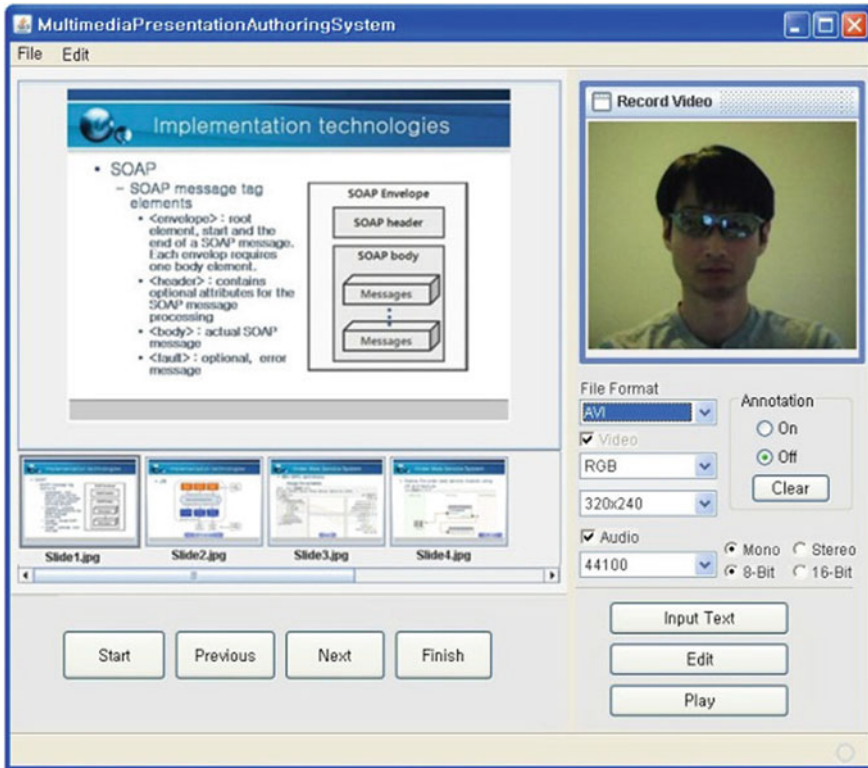


Fig. 2 The snapshot of multimedia presentation authoring tool

Reconstruction Module facilitates reusing existing multimedia presentations. It allows an author to copy, combine, and delete existing presentations with the synchronized SMIL documents.

We offer an authoring tool providing an easy to use authoring and editing environment with the graphic user interface to authors who want to produce multimedia e-learning contents that are available to be accessed by mobile users. It has been designed and implemented using Java with JMF, JDOM and SAXParser. We have used NetBeans IDE 7.0 development environment. The proposed system is composed of three parts: authoring part, editing part, reconstructing part. Figure 2 shows the authoring part of the multimedia authoring tool.

3 Virtual Collaboration System

The proposed system can offer multiple conferences and each conference can have multiple topics. A conference must be created by the system administrator upon the user's request. After the requested conference is created, the requestor becomes

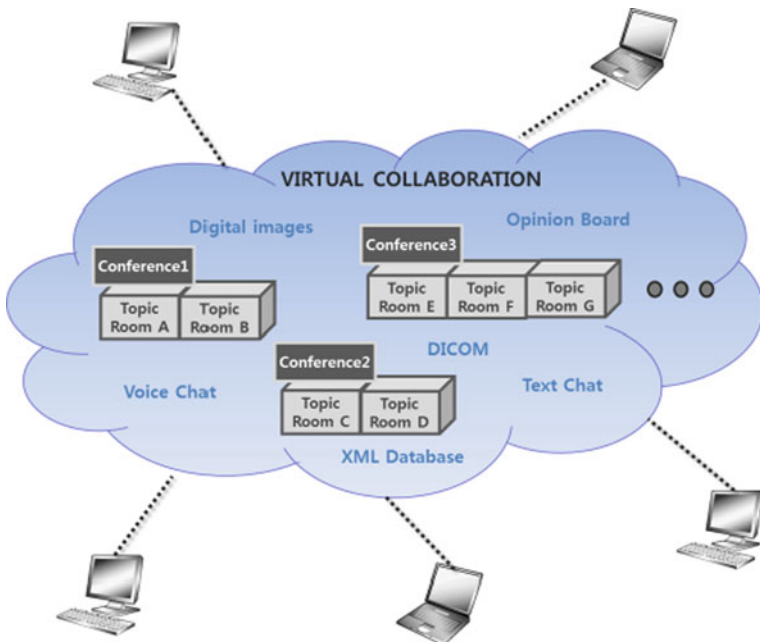


Fig. 3 Users accessing multi-topics in multi-conferences

the conference administrator. The conference administrator can have all the right to operate the conference. In Fig. 3, users can join several topics in different conferences simultaneously according to their access rights.

In this collaboration, the shared working space can be synchronous or asynchronous. In a synchronous mode, the viewpoints of all participants are synchronized. If any participant expresses opinion using a symbol and comment, the content of every participating client object is updated simultaneously. Therefore all participants can see the same view. The synchronous collaboration is typical in shared working spaces such as chatting, instant messenger or whiteboard. However, since the proposed collaboration can record all the synchronous discussions and related materials to the XML database, asynchronous collaboration is also possible by allowing legitimate users to log on the topic at their convenient time and leave new marks and comments which are recorded in a XML database too. Then, the other members can log on the topic and reference those opinions at any time.

Figure 4 shows the communication architecture between a server and clients. A user becomes a client when the user enters the topic room. First, a session manager in a client side tries to establish a socket communication with a connection manager in a server side. After socket connection, each working applet in a client side communicates with a topic manager independently using multi-thread operations. The connection and topic managers are implemented as Java applications and run on a server. The session manager is implemented as a Java applet running on a client's side.

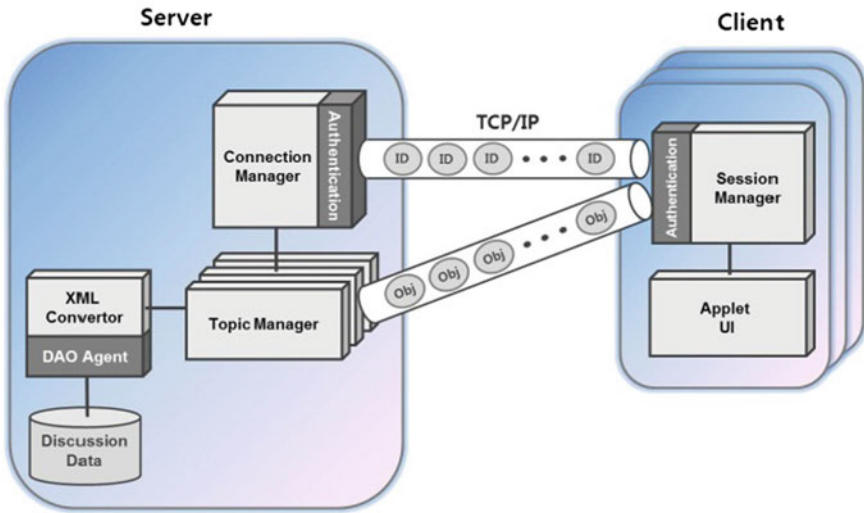


Fig. 4 The communication architecture between server and clients

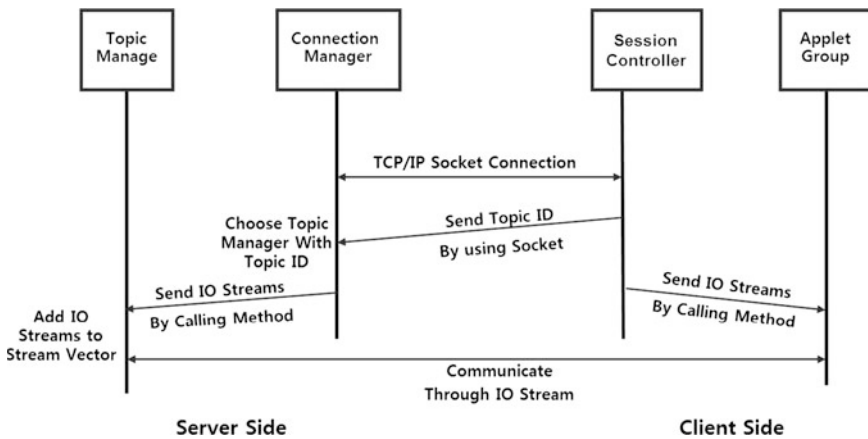


Fig. 5 The sequences of message communication in entering a topic

Figure 5 illustrates the sequences of message communication between a server and a client when a user enters an existing topic. When a user requests to enter a topic, the session manager tries to connect a server by a socket which establishes Input/output object streams. Then, it sends a requested topic ID (identification number) to the connection manager in a server side. The topic ID is unique over the collaboration system, so the connection manager assigns requested IO object streams to the available topic manager. The topic manager adds new IO streams to the existing stream vector for clients.

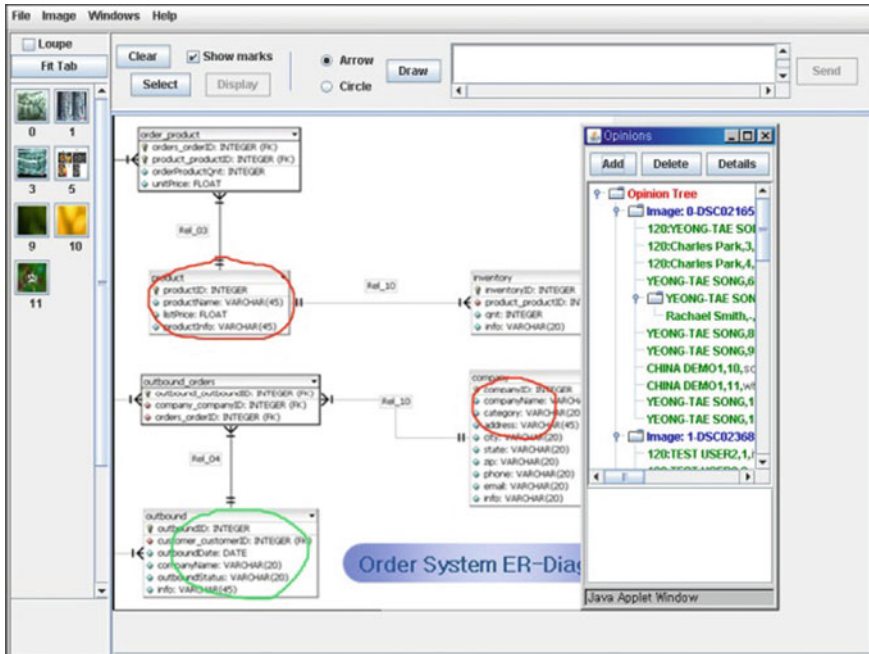


Fig. 6 Virtual collaboration tool

The proposed collaboration system is an Internet based distributed system that enables Internet-based recordable discussions over the images. In virtual collaboration, client objects display digital images, chat window, participant list and participants' opinions. When new opinion is submitted from a participant in a discussion, it can be represented by drawing a spatial symbol such as arrow or polygon on the image. On the server side, topic manager invokes XML converter to record each opinion and have the objects saved into an Oracle database so that they can be available for the client objects in searching and fast loading.

A discussion through the virtual collaboration tool is shown in the Fig. 6. The opinions from the participants will be associated with the symbol on the image.

4 Conclusions

We have proposed a multimedia authoring and virtual collaboration system that helps authors to produce multimedia presentations from diverse media objects and enables users to attend multiple topics in different conferences simultaneously.

The multimedia authoring system provides authors a tool to create multimedia e-learning contents easy to follow. The proposed system makes it possible to integrate diverse media types including text, images, audio and video, and synchronize

media objects by generating SMIL documents automatically. The produced multimedia presentation may be provided to the users through the conference before the discussion begins.

The virtual collaboration tool categorizes users by the interested conference and each user is restricted to access the system resources by the access control. It provides recordable discussions among users using spatial data associated with users' opinions and the images. Digital images and opinions are separated through the layered architecture. Users' opinions along with symbols are recorded to XML files and may be saved into web-based database for future context based intelligent search.

Our proposed system is expected to go as a powerful educational tool through the multimedia authoring and virtual collaboration. It can be an efficient way for learning valuable knowledge and sharing information through the Internet anytime and anywhere.

References

1. W3C Recommendation (2005) Synchronized multimedia integration language (SMIL 2.0). <http://www.w3.org/TR/2005/REC-SMIL2-20050107>
2. Bouyakoub S, Belkhir A (2011) SMIL BUILDER: an incremental authoring tool for SMIL documents. *ACM Trans Multimed Comput, Commun Appl* 7(1): 2:1–2:30
3. Téllez A (2010) E-learning authoring with Docbook and Smil. In: 10th IEEE international conference on advanced learning technologies, Sousse, Tunisia
4. Pooshfam H, Rajeswari M, Ramachandram D (2009) A web-based framework for resource sharing and collaborative medical image annotation. In: Proceedings of the 7th international conference on advances in mobile computing and multimedia, Kuala Lumpur, Malaysia

An Immersive Ski Game Based on a Simulator

Gil Ho Song, Won-Hyung Park, Eun-Jung Lim, Goo Cheol Jeong and Sang-Youn Kim

Abstract This paper addresses a real-time immersive ski game that allows a game player not only to naturally interact with graphic environment but also to sense motion-haptic feedback according to his/her interaction. The proposed system measures a player's motion and then re-creates motion-haptic feedback according to the measured input. To re-create the motion-haptic information in response to a player's interaction with virtual objects, we develop a motion generation platform which consists of a control part, a slope generation part and a horizontal acceleration part. To measure a player's motion input, inertia measurement units (IMUs) are used. We conduct an experiment in order to verify the effectiveness and the feasibility of the proposed game. The experiment clearly shows that the proposed system creates realistic sensation as if a user enjoys a real ski on the snow.

Keywords Haptics · Motion · Immersive sensation · Virtual environment

G. H. Song · W.-H. Park · E.-J. Lim · G. C. Jeong · S.-Y. Kim (✉)
Interaction Lab., Advanced Technology Research Center, Korea Tech, 1600,
Chungjeolno, Byeongchunmyun, Cheonan, ChungNam 330-708, Korea
e-mail: sykim@kut.ac.kr

G. H. Song
e-mail: ime05skh@koreatech.ac.kr

W.-H. Park
e-mail: ipo1001@koreatech.ac.kr

E.-J. Lim
e-mail: naravk1004@koreatech.ac.kr

G. C. Jeong
e-mail: jeong@koreatech.ac.kr

1 Introduction

As the market for leisure sports grew, skiing becomes popular sports for winter season. According to the Korea Ski-Resort Business Association [1], the number of skiers which is tallied 2 million in 1998 ballooned to 6.7 million by 2012 and the seven ski resorts were added in Korea during recent 4 years. Although skiing has become very popular recently, only winter is the best season for skiing in most country. The reason is that the skiing needs special conditions, for example, snow, slope, cold weather, and etc. Recently, as the demands for enjoying skiing at anytime and anywhere are higher, a lot of systems which make users experience a skiing indirectly were studied. Furthermore, these systems have been applied to the research work for game applications.

Due to the advancement of graphic hardware and rendering technology, ski games which started at 2D graphics [2] have been expanded to the game with 3D or stereo graphics. 49GamesTM developed a skiing game based on 3D graphics and the game was constructed from the first person point of view [3]. GameTwistTM was developed a ski game, which is name “Ski Challenge”, on mobile platform [4]. MicrosoftTM was developed a skiing game based on stereo vision technology to represent the ski slope and the avatar in three-dimensional structure [5]. Since a user provides command input with buttons or keypads in those systems, it is not easy to naturally interact with skiing environment.

The natural interaction system, where a user intuitively manipulates game contents with his/her gesture or motion input, can be a solution. Therefore, natural interaction technology based on motion sensing has been incorporated into game systems. NintendoTM showed a ski game (Vancouver Olympic 2010) based on the Nintendo Wii, where user’s motion input is captured by an accelerometer and infrared camera [6, 7]. Solina and Batagelj designed a virtual skiing game system where user’s motion was grasped by an infrared stereo camera [8]. Sugaman and Eichler developed a motion sensing platform which captures a user’s motion input using two pressure sensors. In their system, they attached each pressure sensor beneath the sole of each foot and computed the difference between the pressure values in order to compute a user’s movement. Although, a user naturally interacts with game contents using above systems, the systems are still limited to provide immersion to users. The completely revolutionizing way the users immersively interact with game contents is to provide motion and haptic feedback to them simultaneously.

To express the dynamic motion of skiing, ski systems based on a simulator were studied. Watanabe presents a system which makes user’s feet accelerate horizontally [10]. Déborah Nourrit-Lu-cas et al. mentioned that an arched-shaped ski simulator provides the higher sense of the reality than a flat-shaped simulator [11]. Skytec Interactive developed a huge ski simulator, which creates lateral movement of skies [12]. Even though, these ski simulator systems increase the sense of the reality, these systems have a problem to be solved before accepting industry. The problem is that it is not easy to control each ski plate independently. The united ski

plates may interrupt user’s interaction on feet and it can also spoil the movement feedback which creates separately controlled force. In this paper, we propose a ski simulator which not only creates horizontal motion of ski plates but also controls two ski plates independently.

2 An Immersive Ski Game

In this section, we introduce a ski game based on a simulator which can provide a visual effect, synchronized interaction, horizontal acceleration and the inclination of a ski slope to users. Figure 1a shows a sketch of the proposed simulator. A user can change his/her position horizontally by tilting a ski plate. Two motors were attached at both end sides of the system for creating horizontal acceleration and two pneumatic cylinders which place under each of the ski were used for generating a gradient of a ski slope in virtual environment. Figure 1b shows a structure of the proposed simulator consisting of a motion generation platform and a main controller. DC motors and pneumatic cylinders are used to control the developed a motion generation platform. The user’s intention is captured by the degree of tilting the plate and then transmitted to a main controller. The main controller creates a virtual environment and visualizes user’s sight. Furthermore the main controller computes motion-haptic information and transmits the information to a microprocessor. The transmitted information investigate by the microprocessor and used for moving the ski plate to the left or the right through the actuators.

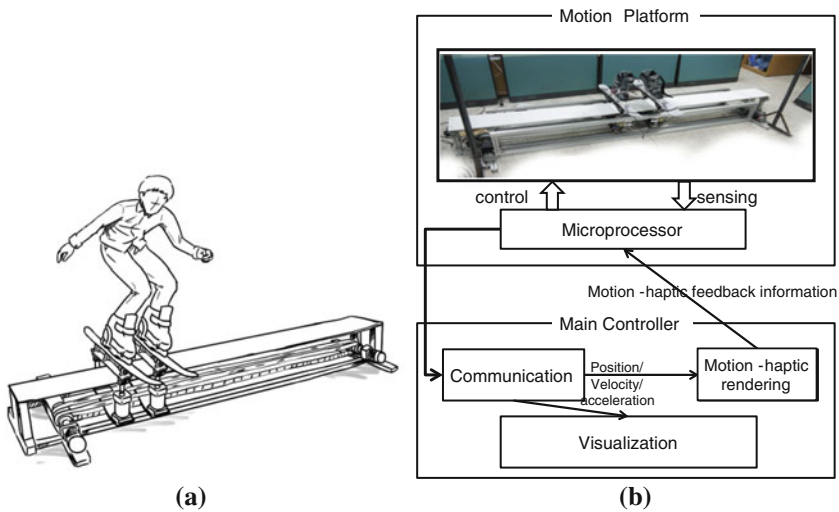


Fig. 1 The developed ski game. **a** A sketch model of the developed game. **b** A structure of the developed game platform

2.1 Main Controller

The main controller consists of a visualization component, a motion-haptic rendering component, and a communication component. We constructed a visualization component in order to convey the game state to a user with visual information. The graphic environment was constructed by UDK (Unreal[®] Development Kit) which is the free version of the award-winning Unreal[®] Engine 3. Virtual objects in the graphic environment were modeled with 3D Max. The visualization component was constructed with Unreal Script to make easier connection with UDK.

Figure 2a shows a virtual model for visualization, and Fig. 2b shows rendered image. The communication component, which is built by Visual C++, receives a user's command from the microprocessor. The microprocessor embedded in the motion input part, via wireless communication and conveys the user's command to the motion-haptic rendering component. Finally, the motion-haptic rendering component analyzes the command and delivers it to the microprocessor.

2.2 Motion Generation Platform

Figure 3 shows the motion generation platform which is controlled by four microprocessors. The motion generation platform consists a motion input part, a slope generation part, and a horizontal acceleration generation part. A motion input part is controlled by two microprocessors which are attached beneath each ski plate. These two microprocessors measure tilted angle of a ski plate. The third microprocessor controls pneumatic cylinders and speed controllers. Cylinders are lift up and down the ski plate's point of head or tail. Through this motion, the motion generation platform can simulate an angle of inclination of a ski slope. The last microprocessor controls two DC motors in order to move the plate from side to side.

Figure 4 shows the system component for the developed simulator. In the control part, two IMU (inertia measurement unit) which are attached to the ski plates, measure the two tilting angles. The measured values are compensated by a Kalman filter. The tilt motion was produced by six bearings and a shaft attaching

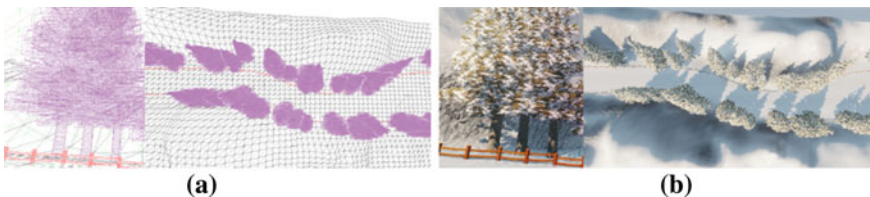


Fig. 2 Block diagram of multi-modal chatter model of a high speed machining center

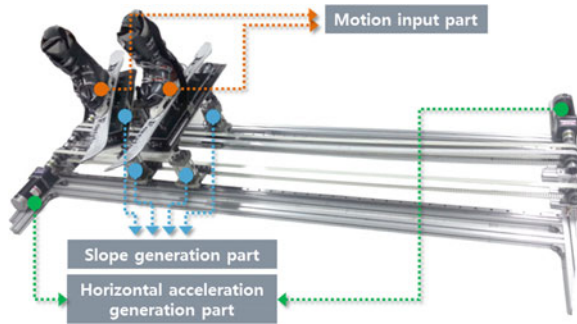


Fig. 3 The motion generation platform

beneath each ski plate as shown in Fig. 4. The slope generation part consisting of pneumatic cylinders, speed controllers and magnetic sensors re-creates an angle of inclination of a ski slope. The speed of each pneumatic cylinder is controlled by speed controllers and magnetic sensors in order to express not only the state of snow but also the amount of slope's inclination. The horizontal acceleration part makes a user to move right or left with a certain velocity or acceleration. The horizontal acceleration part consists of two DC motors, encoders, timing belts and pulleys as shown in Fig. 4.

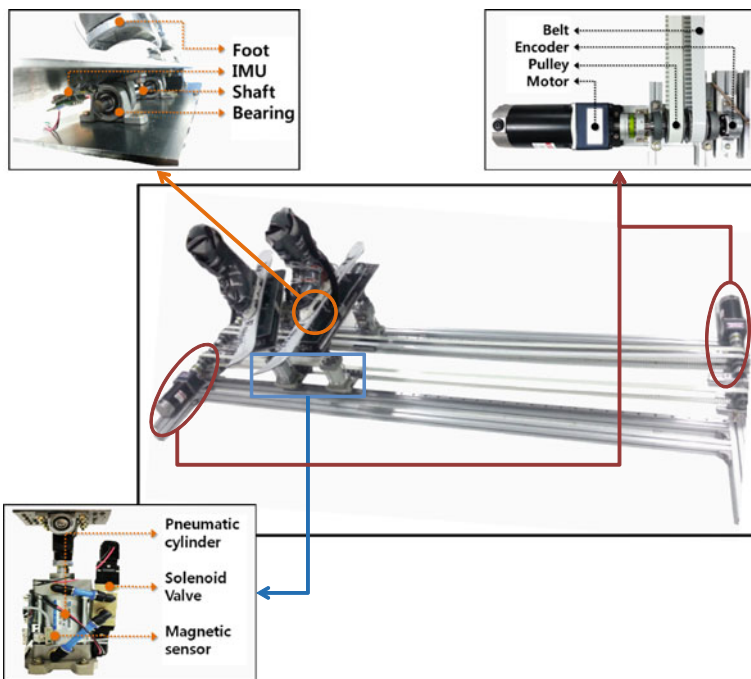


Fig. 4 The system components of a simulator system

$$D = \frac{D_{\max} \cdot \text{Diff}}{2(D_{\max} - \text{Cal}_v)} \quad (1)$$

$$a = \frac{D}{(T_p - T_o) + (T_o) \cdot r_{\min} A_{\max} \omega \cdot \omega^2} \quad (2)$$

where,

D_{\max} maximum distance between user's feet,

Diff difference between two encoders value,

Cal_v initial distance between two feet,

T_p tilting angle of the left ski plate,

T_o tilting angle of the right plate,

A_{\max} maximum tilting angle,

r_{\min} a radius of rotation for ski plates,

ω angular velocity of an avatar

The two DC motors are connected with a timing belt and rotated a role to move the ski plates horizontally. In order to create horizontal acceleration and amount of ski slope's inclination, first we measure the tilted angle of the ski plate and two forces from force sensors attached to ski plates. After that we computed the difference between two forces and then predicted user's motion. According to the predicted motion, we computed a horizontal acceleration of the ski plates. Finally, we computed the horizontal acceleration (a) through (1) and (2).

3 Experiments

We conducted experiments in order to investigate whether the developed system provides immersive sensation to a user as if he/she enjoys a real ski. Fig. 5 shows the still shot where a subject enjoys the game with the developed simulator. In the experiments, ten subjects who did not have experience the immersive simulator joined. They enjoyed the ski game without the motion-haptic feedback. After enjoying the game without motion haptic-feedback, they experienced the proposed game with the motion haptic-feedback and then they filled in and returned questionnaire. To evaluate the developed game, the five questions listed below were presented to each subject. The last question represents the degree of subjects' satisfaction on the developed game with motion-haptic feedback.

- Questionnaire

(a) Did you sense a horizontal acceleration?

(b) Did you distinguish the variance of ski's speed.

(c) Did you feel the inclination of ski slope?

(d) How would you rate our system in its ability to provide realistic feeling to the subject compared to the previous game?

Fig. 5 The still shot of the developed motion-haptic game



In order to investigate the degree of improvement of the game with the developed motion feedback platform ($game_{wm}$) compared with the game without the developed platform ($game_{wom}$) in the range of zero to ten, respectively: Ten score indicates total satisfaction of a subject, while zero represents a subject's complete dissatisfaction. The scores received from the subjects for the $game_{wm}$ were 9, 9, 10, 10, 10, 8, 8, 6, 9, and 9, whereas the scores received from the subjects for the $game_{wom}$ were 6, 4, 5, 5, 7, 6, 4, 5, 6, and 5. We evaluated the degree of the improvement of $game_{wm}$ with (3). The maximum and the minimum improvement rate of $game_{wm}$, compared with $game_{wom}$, have 125 and 20 %, respectively. For more reasonable evaluation of the $game_{wm}$, we excluded the scores received from two subjects who rated two extreme scores. Figure 6a shows the result for the developed system. The horizontal axis of Fig. 6a shows the subject number and the vertical axis shows two scores received from the subjects for the $game_{wm}$ and $game_{wom}$, respectively. From the results, the developed system is shown to be satisfactory for the interactive and immersive game. In addition, we showed that our system adds zest to an interactive game.

$$Rate_{im} = \frac{(game_{wm} - game_{wom})_{wom}}{game_{wom}} \cdot 100 \% \tag{3}$$

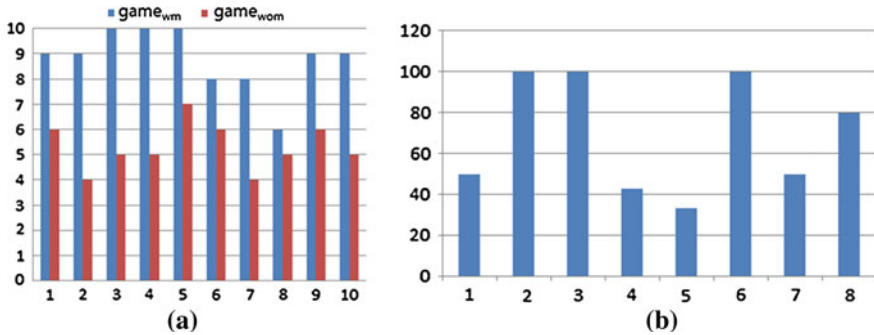


Fig. 6 The result for the developed system. **a** The scores received from the subjects for the $game_{wm}$ and $game_{wom}$. **b** The improvement rate

4 Conclusion

In this paper, we developed a ski game system based on a simulator that provides various motion-haptic feedbacks to users. The developed game stimulates human's somatosensory system according to user's interaction with graphic environment. The vertical translation motion and the horizontal acceleration motion of the actuators were computed on the basis of the graphic environment and the user's command input. In order to show the feasibility of the developed game, we conducted an experiment for two games ($game_{wo}$ and $game_{wom}$), respectively. The developed game with the motion feedback platform makes users sense a ski's acceleration and the state of a ski slope during the game.

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 (grant number : 2012-0004550). This research was also supported by Dual Use Technology Program through Dual Use Program Cooperation Center funded by the Ministry of National Defense (Development of Tactile Display Device for Virtual Reality-based Flight Simulator 2012-DU-EE-03).

References

1. Korea Ski resort Business Association (2012) Website, <http://www.skiresort.or.kr>
2. DEFIANT Development, Ski Safari (2012) Website, <http://www.defiantdev.com>
3. 49Games (2006) Bode Miller Alpine Skiing, Website, <http://www.49games.com>
4. Gametwist (2012) Ski Challenge 2012. Website, <http://www.gametwist.com>
5. Microsoft (2010) ESPN Winter X Games. Website, <http://www.microsoft.com>
6. Nintendo (2006) Wii Remote. Website, <http://www.nintendo.com>
7. Sega Vancouver (2010) Website, <http://www.sega.com>
8. Solina F, Batagelj B, Glamocanin S (2008) Virtual skiing as an Art installation. In: International symposium ELMAR-2008

9. Sugaman H, Weisel Eichler A (2009) Use of the Wii Fit system for the treatment of balance problems in the elderly: A feasibility study. In: Virtual rehabilitation international conference, pp 111–116
10. Watanabe T, Inoue H, Sugimori Y, Sugawara M (2001) SKI SIMULATOR. Patent, US 6,270,430 B1
11. Nourrit-Lucas D, Zelic G, Deschamps T, Hilpron M, Delignieres D (2012) Once you have learned how to ride a bicycle. Archives, hal-00683491, version 1
12. Aleshin V, Klimenko S, Astakhov J, Bobkov A, Borodina M, Volegov D, Kazansky I, Novgorodtsev D, Frolov P (2009) 3D scenes simulation, animation, and synchronization in training systems with force back-coupling. In: 19 International conference graphicon-2009, pp 166–170

A Method of Combining Gaussian Mixture Model and K-Means for Automatic Audio Segmentation of Popular Music

Ing-Jr Ding

Abstract In this study, a hybrid scheme that combines Gaussian mixture model (GMM) and the k-means approach, called GMM-kmeans, is proposed for automatic audio segmentation (AAS) of popular music. Generally, the structure of a popular music is composed of verse, chorus and non-repetitive (such as intro, bridge and outro) segments. The combined GMM-kmeans scheme including mainly two developed algorithms, GMMMAAS and SFS, will efficiently divide a song into these three parts. In GMM-kmeans, the GMM classifier is to recognize the vocal segments and then calculate the section boundary between them and non-repetitive sections first. The song with vocal segments extracted by GMM, containing only the remaining verse and chorus sections, is then analyzed by the k-means clustering algorithm where the verse section is further discriminated from the chorus section. In classification of verse and chorus by k-means, the developed switching frame search (SFS) algorithm with the devise of verse group-of-frames (Verse-GoF) and Chorus-GoF will accurately estimate the separation boundary of verse and chorus sections. Experimental results obtained from a musical data set of numerous Chinese popular songs show the superiority of both proposed GMMMAAS and SFS.

Keywords: Automatic audio segmentation • Gaussian mixture model • K-means • GMMMAAS • SFS

I.-J. Ding (✉)

Department of Electrical Engineering, National Formosa University,
No.64, Wunhua Rd, Huwei Township, Yunlin County 632, Taiwan ROC
e-mail: ingjr@nfu.edu.tw

1 Introduction

Automatic audio segmentation (AAS) has been a popular technique in the recent years. AAS could be frequently seen in many multimedia software applications to be an intelligent multimedia tool. In fact, AAS could be as a branch in the field of speech/audio processing. Research pertaining to speech/audio information processing encompasses myriad branches including encoding/decoding, identification/verification and analysis/synthesis [1]. Although AAS that belongs to the category of analysis/synthesis is not so matured as audio codecs, speech recognition and audio synthesis techniques in the aspect of commercial products, more and more studies focused on AAS have been seen recently [2–4].

Most works on AAS aims to the feature-based approaches. In [2], a method of segmenting musical audio into structural sections based on a hierarchical labelling of spectral features is presented. Such the approach is typically a feature-based classification mechanism with constrained clustering. In the work of [3], Lukashevich proposes an approach of applied normalization, which will enable the comparison of the automatic evaluation results, obtained for songs with a different amount of states. In addition, Peiszer develops a two-phase algorithm for boundary and structure detection where the complete annotation of all song parts both with sequential-unaware approaches and an approach that takes temporal information into account is paid much attention [4].

Rare AAS studies use statistically modeling methods. Although feature-based AAS methods may achieve certain degree of classification accuracy, the main weakness of those is less flexible. Feature-based AAS lacks the capability of environmental adaption and system learning. In contrast, model-based AAS approaches could continuously improve the recognition accuracy of the system by batch or incremental learning of the increased training data. This work focuses on the model-based AAS category of methods, especially on the processing of combining multiple classification models/clustering methods. Gaussian mixture model (GMM) [5] and the k-means clustering method [6] are combined for AAS applications. For these principal problems of AAS, detection of the vocal scene and classification of verse and chorus, this paper proposes GMMAAS and switching frame search (SFS) algorithms to effectively solve, respectively, which will be detailed in the following sections.

2 Automatic Audio Segmentation

In general, automatic audio segmentation could be performed in two categories of methods, feature-based and model-based. In this paper, the model-based approach is adopted. The framework of an AAS system that uses the model-based technique is associated with established audio models, frequently-seen GMM for example, where the input popular music from the database is segmented into the frame

sequence, and from which audio features are extracted to evaluate the characteristic of this music via pattern clustering (the famous k-means for example) operations cooperated with GMM audio model calculations. The segmentation task is finally completed and verse, chorus and non-repetitive (intro, bridge and outro) sections can then be recognized.

Mathematically, a GMM is a weighted sum of M Gaussians, denoted as [5].

$$\lambda = \{w_i, \mu_i, \Sigma_i\}, i = 1, 2, \dots, M, \sum_{i=1}^M w_i = 1, \tag{1}$$

where w_i is the weight, μ_i is the mean and Σ_i is the covariance. In this study, two parameter sets, λ_1 and λ_2 (two GMM models, that is), for representing the music characteristics of vocal and non-vocal (non-repetitive) categories are determined, respectively.

In the operation of testing phase, the estimated likelihood score that an audio frame x_i , belongs to a GMM of class $Z \in \{\lambda_1, \lambda_2\}$ is then

$$L(x_i|Z) = \sum_{j=1}^M w_j \cdot \frac{1}{(2\pi)^{D/2} \cdot |\Sigma_j|^{1/2}} \cdot \exp\left\{-\frac{1}{2}(x_i - \mu_j)^T (\Sigma_j)^{-1} (x_i - \mu_j)\right\}. \tag{2}$$

Although lots of clustering methods performs successfully in clustering applications, the k-means approach remains a more preferred choice due its simplicity and effectiveness in large-scaled pattern clustering. The k-means objective function provides a GMM-like classification model for data clustering (audio frames clustering in this work). K-means belongs to the category of “hard clustering” of data [7].

Given a set of audio frames x_1, x_2, \dots, x_n for clustering, the k-means objective function tries to find clusters m_1, m_2, \dots, m_k to achieve the optimization operation to minimize the objective function as follows [6]:

$$\min_{\{m_c\}_{c=1}^k} \sum_{c=1}^k \sum_{x \in m_c} \|x - \mu_c\|_2^2, \tag{3}$$

where

$$\mu_c = \frac{1}{|m_c|} \sum_{x \in m_c} x. \tag{4}$$

The k-means method to fulfill optimization of Eq. (3) contains three main steps: (1) Initialization of clusters: the initialization step is to carry out a hard clustering of audio frames along with the cluster mean of these clusters; (2) Reassignment of audio frames: the reassignment step is to reassign audio frames to clusters and recalculate the means of clusters. The squared Euclidean distance is computed from each audio frame to each cluster mean, and the minimum distance is then acquired by calculating $m^*(i) = \arg \min_c \|x_i - \mu_c\|_2^2$. $m^*(i)$ is an index that will indicate the cluster where the frame x_i is assigned; (3) Update of the centroid:

the step is to update μ_c for each cluster by recalculate the means of clusters. Note that the value of k is set to 2, indicating two categories of audio frames for clustering, verse- and chorus-type frames.

3 The Proposed Combined GMM-Kmeans Method for AAS

This section presents the developed combined GMM-Kmeans approach for automatic audio segmentation. In general, a Chinese popular song contains mainly five parts, which are intro, first vocal segment, bridge, second vocal segment and outro. The GMM classifier will be firstly used to recognize two vocal and three non-vocal (non-repetitive) segments. Each of these two recognized vocal segments will then be detected its verse and chorus parts by the k-means clustering method.

3.1 GMM for Vocal Parts AAS (GMMAAS)

GMM technique is famous for its excellent performance on recognition of acoustic patterns, speech recognition and speaker recognition, for example. Following the line of thoughts, GMM is employed to detect the acoustically vocal parts of a song in this work. In the GMM calculations for audio segmentation here, the song chosen from a database for segmentation is fed into two GMM models, vocal and non-vocal classifiers, and a decision scheme of decision window (DW) is designed to consider all classification results of these two classifiers.

Figure 1 depicts a stream of predefined-length decision windows, each of which covers the same number of audio frames and is thus of the same time span. The process of vocal segment detection on a song would feed the data stream of audio frames into GMM classifiers by which successive analysis on a pre-determined DW is conducted and then the decision as to whether a starting point or an ending point of the vocal segment being detected over the associated time span, is made.

Then the evaluation of the audio data segment X containing n frames ($X = \{x_i | i = 1, 2, \dots, n\}$) within an DW is done by comparing X with both the vocal GMM model and the non-vocal GMM model, i.e. λ_1 and λ_2 . For the overall audio frames X , the likelihood score between each frame and GMM models is locally computed and then accumulated as log-likelihood ratio (LLR):

$$\begin{aligned} LLR(X) &= \log \frac{L(X|\lambda_1)}{L(X|\lambda_2)} = \log(L(X|\lambda_1)) - \log(L(X|\lambda_2)) \\ &= \sum_{i=1}^n \log L(x_i|\lambda_1) - \log L(x_i|\lambda_2), \end{aligned} \quad (5)$$

where $L(x_i|\lambda_1)$ and $L(x_i|\lambda_2)$ are given by Eq. (2), representing the likelihood of λ_1 and λ_2 model classification, respectively, for frame x_i .

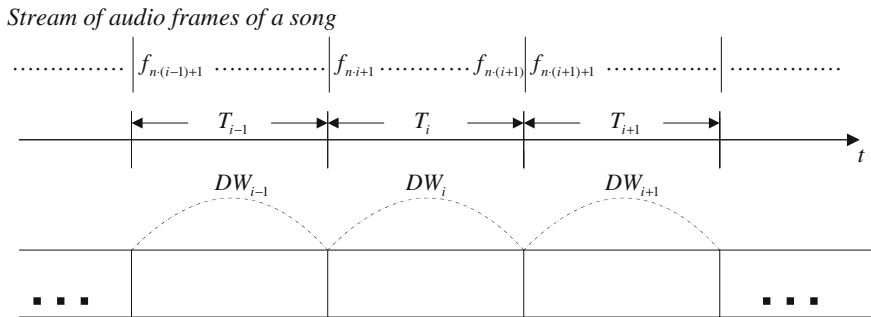


Fig. 1 DW with predefined-length, each covering the same number of audio frames, n , over the time span

The rationale behind Eq. (5) is that at the beginning stage covering n frames, say, of a decision window, if the class inclination of the frames has clearly exhibited, one term in Eq. (5) will be substantially greater than the other. As a consequence, a salient LLR value that is relatively large or relatively small is acquired, indicating that a vocal or non-vocal audio fragment is appearing respectively. If the class of the n frames can not be resolved, possibly containing both vocal and non-vocal audio frames, both terms in Eq. (5) would be competitive and lead to a non-prominent LLR implying the occurrence of the junction between vocal and non-vocal audio segments (intro and the first vocal segment, the first vocal segment and bridge, bridge and the second vocal segment, or the second vocal segment and outro). Figure 2 shows the developed GMMAAS algorithm to utilize GMM to carry out the vocal/non-vocal automatic audio segmentation procedure.

Note that values of *Threshold1*, *Threshold2*, *Threshold3* and *Threshold4* in the GMMAAS algorithm are determined in a trivial procedure. In this study, *Threshold1* would be a large value to check if a vocal fragment appears. A small value would be set to *Threshold2* for indicating if a non-vocal fragment occurs. The values of *Threshold3* and *Threshold4* would have the tendency to approach to zero for reflecting a junction between vocal and non-repetitive segments.

3.2 K-means for Verse and Chorus Detection

As mentioned, the vocal segments of a song may be distinguished by the proposed GMMAAS approach. In vast majority of popular Chinese songs, the vocal segment contains a verse and a chorus sections. The popular k-means clustering method will be used to separate these two categories of sections in this section. Since each of verse and chorus sections is successive in essence, a simple and direct way to detect verse\chorus sections is to search the connection position of these two sections. Such the connection points of verse and chorus segments are defined as

```

GMMAAS algorithm for detecting vocal segments
Initialize LLR values of each frame and all accumulated frames X in a
decision window to be zero:
LLR_Counter = LLR(X) = 0;
for each decision window DWm in the duration of a song
  for each frame i in a DWm
    for each model s (s = 1 to 2) /* likelihood calculations of 2 GMM */
      Derive  $L(x_i | \lambda_s) = \sum_{j=1}^M w_j \cdot b_j(x_i)$ ;
    end for
    LLR_Counter +=  $\log L(x_i | \lambda_1) - \log L(x_i | \lambda_2)$ ;
  end for
  LLR(X) = LLR_Counter;
  if (LLR(X) > Threshold1) then
    X is labeled as the vocal fragment;
  else if (LLR(X) < Threshold2) then
    X is labeled as the non-vocal (non-repetitive) fragment;
  else if (Threshold3 < LLR(X) < Threshold4) then
    X is labeled as the junction between vocal and non-vocal segments;
  else
    X is an undetermined fragment and neglected;
  end if
end for
Estimate two vocal (the first and the second) and three non-repetitive
segments (intro, bridge and outro) according to all X with labels;
    
```

Fig. 2 The developed GMMAAS algorithm

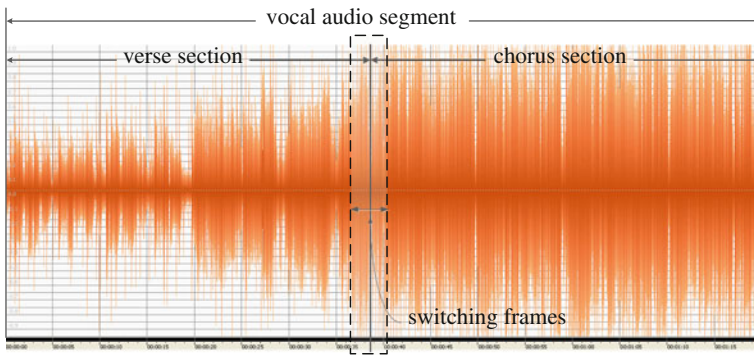


Fig. 3 Switching frames located between the verse and the chorus sections

switching frames (SF), denoting the vague region that will switch the verse to the chorus frames, in this paper. Figure 3 depicts the relationships of the first continuous verse section, followed switching frames and the last continuous chorus section. A switching frame search (SFS) algorithm is developed herein to accurately determine the appearance location of the switching frame. The proposed SFS is an iterative procedure, which is showed in Fig. 4.

```

SFS algorithm for determining the switching frame
Initialize numbers of verse-GoF and chorus-GoF to be zero:
no. of verse-GoF = no. of chorus-GoF = 0;
for each vocal segment  $s$  derived from GMMAS algorithm
  for each decision region  $t$  in the vocal segment  $s$ 
    for each GoF( $u$ ) in the decision region  $t$ 
      Determine the label of GoF( $u$ ) as verse (verse-GoF) or chorus
      (chorus-GoF) by k-means;
    end
    Calculate numbers of verse-GoF and chorus-GoF;
    if (no. of verse-GoF  $\geq$  no. of chorus-GoF)
      Label decision region  $t$  as the verse fragment;
    else if (no. of verse-GoF < no. of chorus-GoF)
      Label decision region  $t$  as the chorus fragment;
      switching region = decision region  $t$ ;
      break;
    end if
  end
end
Derive the switching frame in the switching region by trivial processing;

```

Fig. 4 The proposed switching frame search (SFS) algorithm using k-means

Note that in SFS algorithm, GoF(u) denoting the u -th group of frames is designed as an unit with extremely short time duration. The k-means clustering method is used to divide the group of frames into two sides, verse and chorus categories. The classification tendency (verse-GoF or chorus-GoF) of these frames contained in the same group is then evaluated by comparing the number of verse-labeled frames and the number chorus-labeled frames. The decision region, analog to the above-mentioned decision window, is a pre-determined time interval that contains numerous GoFs. The decision region will provide a decision of verse or chorus according to numbers of verse-GoF and chorus-GoF covered in the region. In case a chorus-labeled decision region is found, the decision region denoting the switching region must contain the desired switching frame. The switching frame is finally able to be acquired by trivial processing where the switching region is slightly shorten, and then numbers of verse-GoF and chorus-GoF are re-estimated again and again. Figure 5 illustrates the switching frame between verse and chorus successive sections of a vocal audio segment acquired by the SFS approach.

4 Experiments and Results

The automatic audio segmentation experiments includes detections of the vocal segments by the proposed GMMAS method and classification of the verse and the chores sections in a vocal segment using the developed SFS algorithm with the support of k-means. Hundreds of popular Chinese songs were collected for establishing the database to evaluate the performance of presented approaches.

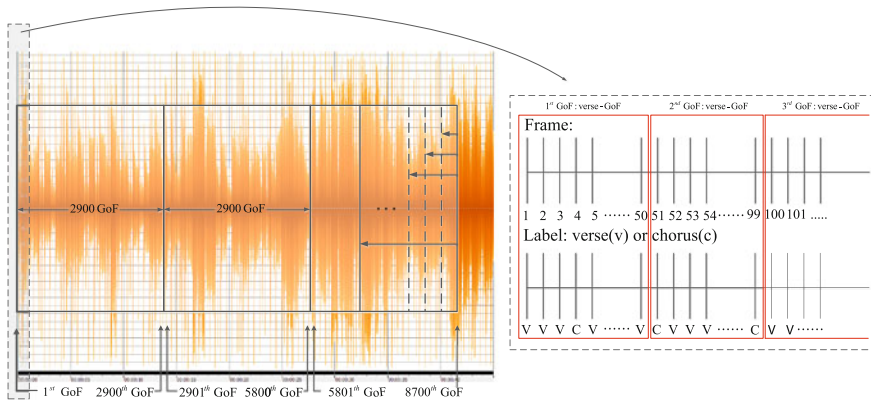


Fig. 5 SFS algorithm to find the switching frame in a vocal audio segment

Table 1 Recognition accuracy of proposed GMMAAS in music vocal detection

GMM mixtures: 3	Patterns for classification		Recognition rates (%)
	Correct no.	Wrong no.	
Vocal	185	22	89.37
Non-vocal	29	10	74.36
Vocal	188	19	90.82
Non-vocal	29	10	74.36
Vocal	175	32	84.54
Non-vocal	33	6	84.62

The analysis frames were 20-ms wide without frame overlaps. For each frame, a 10-dimensional feature vector was extracted. The feature vector for each frame was a 10-dimensional cepstral vector.

In the phase of GMM vocal segment extractions, two GMM models, vocal and non-vocal classifiers, were built up in advance using the prepared database. The classification accuracy of developed GMMAAS was then further evaluated by the likelihood calculations between two trained GMM models. Table 1 shows the recognition performance of GMMAAS. Note that the number of GMM mixtures is set to 3, 4 and 5, each of which has different classification accuracy. The setting of 4 GMM mixtures achieves the highest recognition rate of 90.82 %, which is an acceptable performance for audio segmentation applications.

In the phase of verse/chorus classification based on k-means, the presented SFS algorithm with k-means is to evaluate the performance using the same music database as the GMM vocal/non-vocal classification experiments. Table 2 shows the competitive performance of developed SFS. Observed from Table 2, the 2nd song and 4th song have the lowest error rate, which is 0 %. The first song performs the highest error rate, which is just 2.5 % and will be acceptable. The average error

Table 2 Recognition accuracy of developed SFS in music verse/chorus detection

No.	Popular Chinese songs (wav files)	Switching time of verse and chorus by SFS (s)	Real switching time of verse and chorus (s)	Classification error rates (%)
1	“陳勢安-天后”	49.5	47	2.5
2	“鄧福如-為填詞”	36	36	0
3	“神木與瞳-守護者”	41	40	1
4	“張惠妹-我最親愛的”	39	39	0
5	“蕭敬騰-只能想念你”	45	46	1
6	“A-Lin-偽裝”	38	37	1
20	“周杰倫-我不配”	53	54	1
Avg.				1.1

rate of these 20 Chinese popular songs achieves the excellent performance of 1.1 %, which shows that the proposed SFS algorithm for finding the switching frame of verse and chorus will be extremely efficient and effective.

5 Conclusions

This paper proposes two schemes for automatic audio segmentation applications, GMMAAS and SFS methods. The GMMAAS approach is to use Gaussian mixture model to classify vocal and non-vocal segments of a song. The design of GMMAAS mainly calculates the log-likelihood ratio of two classes of GMM in a decision window. The SFS algorithm with the support of k-means clustering could detect the switching frame between verse and chorus of a vocal segment, and therefore verse and chorus will be classified effectively. Experimental results demonstrated that both developed GMMAAS and SFS achieved competitive and acceptable recognition accuracy.

Acknowledgments This research is partially supported by the National Science Council (NSC) in Taiwan under grant NSC 101-2221-E-150-084.

References

1. WutiwWATCHAI C, FURUI S (2007) Thai speech processing technology: a review. *Speech Commun* 49:8–27
2. LEVY M, SANDLER M (2008) Structural segmentation of musical audio by constrained clustering. *IEEE Trans Audio, Speech, Lang Process* 16:318–326
3. LUKASHEVICH H (2008) Towards quantitative measures of evaluating song segmentation. In: *Proceedings of international conference on music information retrieval*, pp 375–380
4. PEISZER E, LIDY T, RAUBER A (2008) Automatic audio segmentation: segment boundary and structure detection in popular music. In: *Proceedings of international workshop on learning the semantics of audio signals*

5. Reynolds DA, Rose RC (1995) Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans Speech Audio Process* 3:72–83
6. Hartigan JA, Wong MA (1979) A K-means clustering algorithm. *Appl Stat* 28:100–108
7. Bezdek JC, Ehrlich R, Full W (1984) FCM: the fuzzy c-means clustering algorithm. *Comput Geosci* 10:191–203

Human Action Classification and Unusual Action Recognition Algorithm for Intelligent Surveillance System

Nae Jung Kwak and Teuk-Seob Song

Abstract This paper suggests an algorithm to classify human actions for the intelligent surveillance system. In order to classify actions, the proposed method calculates the difference image between input images and modeled background, uses motion information histogram and traces of center points of objects. Human actions are categorized into four types: they are the most frequently three actions people take walking, sitting, standing up and unusual action like as sudden falling down. We examine the proposed method on eight people with a sequence captured by using a web camera and the result shows that the proposed method classifies human actions well and recognitions of the unusual action.

Keywords Silhouette · Object tracking · Auto detection · Surveillance

1 Introduction

The surveillance system has attracted attention as a core factor for evolving the ubiquitous environment and has widely been utilized in public spaces, the regions which require high level of security or have high possibility of crime such as hospitals, banks and parking lots. A growing concern over terrorist attacks has been paid attention to the surveillance system, in particular, intelligent surveillance

N. J. Kwak · T.-S. Song (✉)

Department of Computer Engineering, Mokwon University, Mokwongil 21,
Daejeon, South Korea
e-mail: teukseob@mokwon.ac.kr

N. J. Kwak

e-mail: knj0125@hanmail.net

one which recognizes human actions and sounds an alarm after determining the context of a situation. Therefore, there have been studying intelligent surveillance systems [1–3].

The studies include methods to recognize actions by observing continuous human actions [4], by using predefined scenario models [5, 6] and by utilizing a probabilistic model [3, 7]. Existing techniques for recognizing human actions use such information as location, traces and velocity of objects. This information is used as basic data by analyzing images from cameras, extracting objects and combining features of extracted objects and events. Thus, the methods that extract objects of interest from input images and recognize their features and actions by using video cameras have continuously developed based on image recognition [8–10].

This paper suggests an algorithm effectively applied to the intelligent surveillance system and classifies human actions into four types: walking, sitting and standing up and unusual action such as faint or sudden falling down. Conventional methods of recognizing human actions and estimating postures have to learn features of extracted human actions and then recognize or estimate postures from input images. That means those methods require a lot of learning data and complex learning algorithms. This paper suggests a method to recognize objects' actions by using motion information histogram (MIH) and traces of center points of objects. Thus, the proposed method does not require a lot of learning data or learning algorithms but efficiently recognize human actions and detect unusual ones.

Section 2 describes the way to extract objects and to classify actions along with alarm service for unusual actions. In Sect. 3, we evaluate the performance of the proposed method and in Sect. 4, provide conclusions.

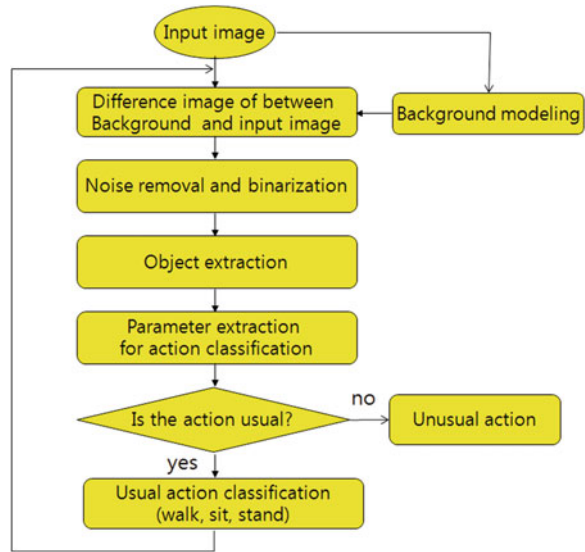
2 Action Classification by Using Motion Information Histogram and Traces of Center Points of Objects

This paper separates objects from background by calculating the difference image between modeled background and the input image from a single camera, extracts objects based on background information and uses histogram of the extracted objects as well as traces of center points to propose an algorithm classifying actions into walking, sitting, standing up and unusual action. Figure 1 shows the flowchart of the proposed algorithm.

2.1 Background Modeling and Object Extraction

This paper models background by using the method proposed in [11]. In this paper, we first calculate the difference image between the modeled background and input frames. The difference image is an image separating objects from the background. If an input image has a pixel value between high and low thresholds on three

Fig. 1 Flowchart of the proposed method



planes, it is “Background”, otherwise, it is “Object.” The following is the equation to separate objects from background.

$$B(x,y) = \begin{cases} 0, & Th_l \leq I(x,y) \leq Th_h \\ 255, & otherwise \end{cases} \quad (1)$$

Here, x,y indicates location of the input image. In the background image, “Object” is white (255) and “Background” is black (0). By combining the results of three planes, the resulting image is produced. The resulting image has various noise and areas which are not included in the object. Thus, we reduce areas and noises by using morphological filters to get background-object separation image, $B(x,y)$.

2.2 Parameter for Classifying Object’s Actions

The proposed method uses motion information histogram (MIH) and traces of center points of objects to classify human actions into “walking”, “sitting”, “standing up”, and unusual actions such as “falling down” and “dropping down”. Figure 2 is a histogram of extracting objects from the image and projecting the object area onto the x-axis. Object histogram of Fig. 2 shows values calculated by projecting object area on the x-axis and so it does not provide valid parameters to determine human actions. This paper calculates MIH by using motion information of objects and determines human actions by defining features of each action of MIH.

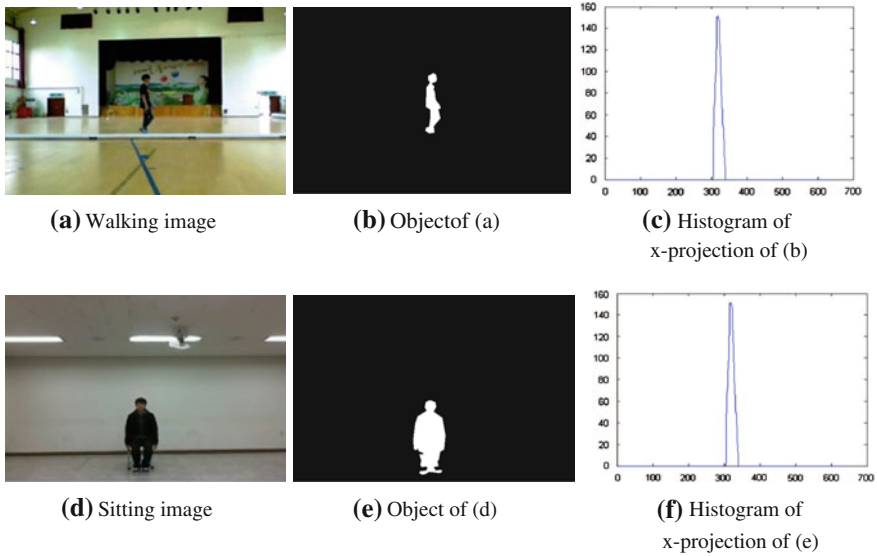


Fig. 2 Histogram for each action: **a** walking image, **b** object of **a**, **c** histogram of x-projection of **b**, **d** sitting image, **e** object of **d**, **f** histogram of x-projection of **e**

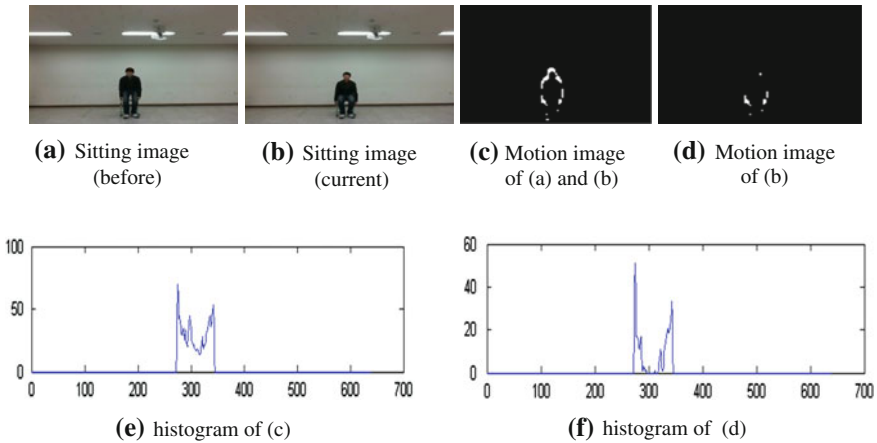






Fig. 3 Motion Information Histogram of sitting action: **a** sitting image (before), **b** sitting image (current), **c** motion image of **a** and **b**, **d** motion image of **b**, **e** histogram of **c**, **f** histogram of **d**

Figure 3 shows MIH of walking and sitting action. Figure 3e, f is the MIH and demonstrates well features of sitting action. This paper chooses MIH of objects on the current frame of Fig. 3f as characteristics parameters. Equation (2) is the equation of proposed MIH.

$$MIH = (FB \cdot FC) - FB \tag{2}$$

Table 1 MIH of 4 action

Action	Walk	Sit	Stand-up	Unusual
MIH				

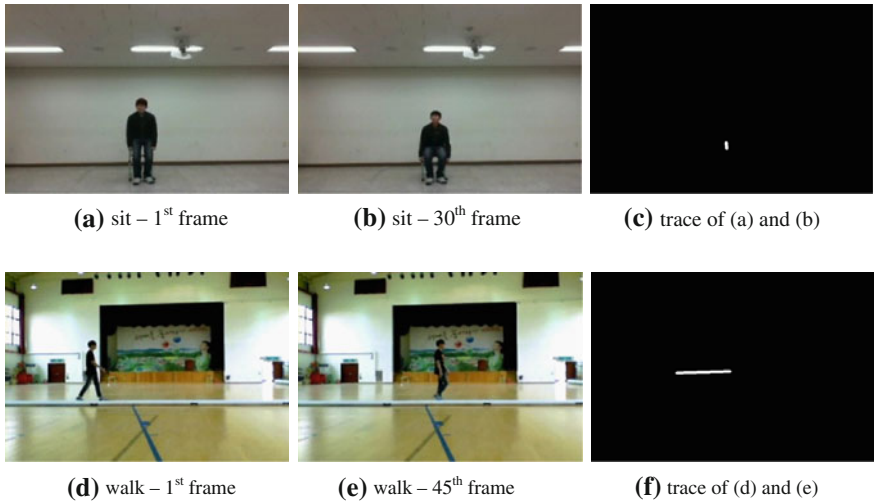


Fig. 4 Trace of object’s center point for each action **a** sit—1st frame, **b** sit—30th frame, **c** trace of **a** and **b**, **d** walk—1st frame, **e** walk—45th frame, **f** trace of **d** and **e**

Here, FB indicates previous frame, FC, the current frame and ‘.’, exclusive-or. MIH is a value of objects’ motion on the only current frame.

Table 1 defines MIH parameters for classification according to 4 actions. We classify actions by comparing MIH patterns of Table 1 with that of the input image. To determine similarity, we use EMD [12].

Furthermore, a change in movement and motion of extracted objects accompanies a change in center points of extracted objects. Figure 4 is the traces of changed center points of objects for “walking” and “sitting” and shows that each

Table 2 Center point’s change for action classification

Action	Center point	
	X	Y
Walking	±	*
Sitting	*	+
Standing	*	-
Unusual action	±	+

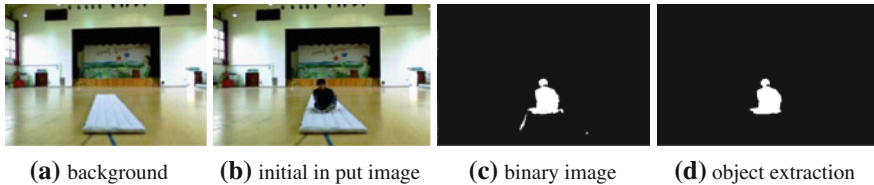


Fig. 5 Each result by the proposed method: **a** background, **b** initial input image, **c** binary image, **d** object extraction

action has different trace of center points. The proposed method tracks center points and uses their variation as an action classification parameter.

Table 2 shows the values (+ and -) to express the change of the center points by tracking center points to classify objects' actions. "*" indicates "irrelevant."

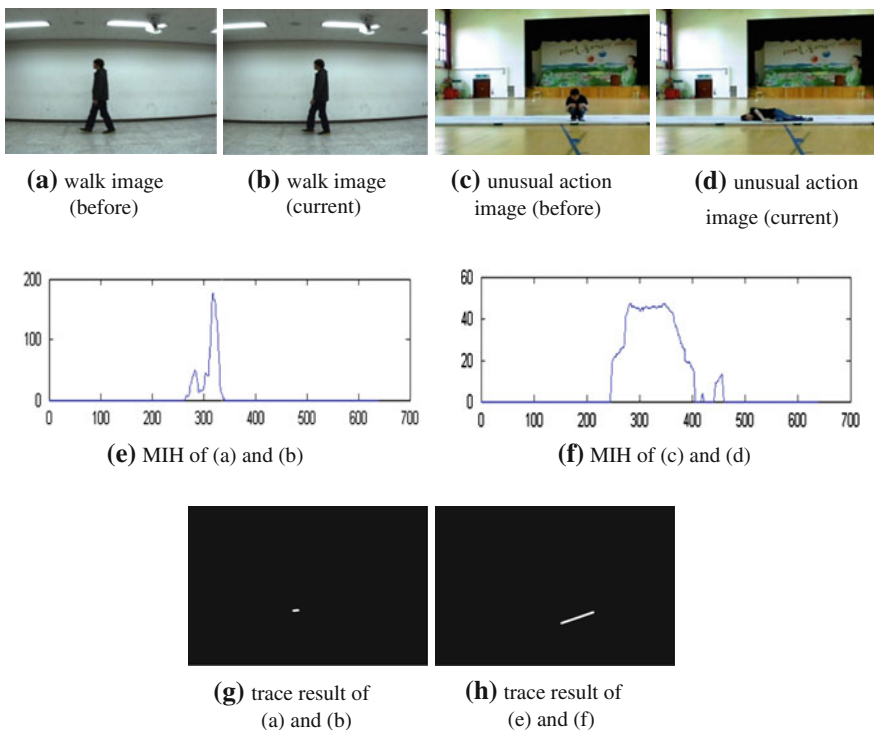


Fig. 6 MIH and trace result of walk action and unusual action **a** walk image (before), **b** walk image (current), **c** unusual action image (before), **d** unusual action image (current), **e** MIH of **a** and **b**, **f** MIH of **c** and **d**, **g** trace result of **a** and **b**, **h** trace result of **e** and **f**

Table 3 The result of human action recognition

Action	1	2	3	4	5	6	7	8	Recognition ratio (%)
Walking	300	300	300	300	300	300	300	300	100
Sitting	300	300	300	300	300	300	300	300	100
Standing	300	300	300	300	300	300	300	300	100
Unusual action	-	-	-	-	-	-	-	-	100

3 Experiment and Result Analysis

In order to analyze performance of the proposed method, we process background and input images in real time by using a camera. We conduct an experiment on eight people. Each person is asked to perform four actions: sitting on chair, standing up and unusual actions such as falling down and faint. Moreover, we create two different indoor settings. We use Intel cpu 2.0 GHz, RAM, Visual Studio 2008 and OpenCV 2.1 for the experiment. The resolution 1G of the input image is 640×480 24 bits at 15 frames. Figure 5 shows test images.

Figure 5 shows results of each stage by the proposed method. (c) is the binary image to separate foreground from background of inputted image and (d) is the extracted object of (c).

Figure 6 shows results of MIH and center point tracking of two frames. (a) and (b) are input frame of walking action while (c) and (d) are that of unusual action. (e) is MIH of (a) and (b) while (f) is MIH of (c) and (d). (g) and (h) are trace results of walking and unusual action.

Table 3 shows results of human action recognition obtained by using the proposed method. It shows the number of recognized frames and mean recognition rates on each 300 frames. Since unusual actions suddenly take place, we do not constrain the number of frames.

4 Conclusion

This paper proposes an algorithm which extracts objects by using the difference image between background and input image captured by a single camera and classifies human actions by using object histogram and motion information of the extracted objects. In order to effectively extract objects, background image is modeled with multiple background frames. We classify human actions into four types, walking, sitting and standing up and unusual action such as falling down and dropping down. To classify objects' actions, we use MIH of objects and the traces of center points of objects' motion.

We examine the proposed method on frames from a web camera and the experiment shows perfect performance in recognizing the four types of actions. The proposed method does not require complex learning or algorithms and can

have various applications such as surveillance cameras or u-Health in the ubiquitous environment.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0002852).

References

1. Ayers D, Shah M (1998) Monitoring human behavior in an office environment. In: Computer society workshop on interpretation of visual motion
2. Robertson NM, Reid ID (2006) A general method for human activity in video. *Comput Vis Image Underst* 104:232–248
3. Peursum P, Bui HH, Svetha V, Geoff W (2005) Robust recognition and segmentation of Human actions using HMMs with missing observation *EURASIP J Appl Signal Process* 13:2110–2126
4. Kellokumpu V, Pietikäinen M, Heikkilä (2005) Human activity recognition using sequences of postures. In: IAPR conference on machine vision applications, pp 570–573
5. Rota N, Thonnat M (2000) Video sequence interpretation for visual surveillance. In: 3rd IEEE international workshop on visual surveillance, VS'00, pp 59–67
6. Vu VT, Bremond F, Thonnat M (2003) Automatic video interpretation: a recognition algorithm for temporal scenarios based on pre-compiled scenario models. In: Third international conference on computer vision systems (ICVS 2003), Lecture notes in computer science (LNCS), vol 2626. Springer, Berlin, Graz, Austria. pp 44–53
7. Ayers D, Chellappa R (2000) Scenario recognition from video using a hierarchy of dynamic belief networks. In: ICPR, pp 1835–1838
8. Tao J, Tan YP (2004) A probabilistic reasoning approach to closed-room people monitoring. *IEEE ISCAS* 2:185–188
9. Utsumi A, Yang H, Ohya J (2000) Adaptive human motion tracking using non-synchronous multiple viewpoint observations. In: Proceedings of IEEE 15th international conference on pattern recognition, vol 4, pp 607–610
10. de Campos TE, Murray DW (2006) Regression- based hand pose estimation from multiple cameras. *CVPR* 1:782–789
11. Kim K, Chalidabhongse TH, Harwood D, Davis L (2005) Real-time foreground-background segmentation using codebook model. *Real-time Imaging* 11:167–256
12. Rubner Y, Tomasi C, Guibas LJ (1998) A metric for distributions with applications to image databases. In: IEEE international conference on computer vision, pp 59–66, Jan 1998

Design of Configurable Pin Control Block for Multimedia System-on-a-Chip

Myoung-Seo Kim and Jean-Luc Gaudiot

Abstract The complexity of generic pin control blocks of multimedia system-on-a-chip (SoC) which implements input/output (I/O) paths for off-chip communication has been increased significantly. Accordingly, the possibility of making human errors in designing this block has been magnified as a new controversy. Generic pin control blocks possess several productivity issues since special registers for an additional function and multi-I/O paths are usually fixed at relatively late stages of design activities. Also, generic pin control blocks may have different types of design according to the designer. This feature results in various human errors when we use the traditional RTL description. Thus, this paper presents an approach to reduce human errors based on design automation. In our case study, we succeeded in auto-generating a configurable pin control block in a multimedia SoC platform which has more than 300 generic pins including whether it is an input and output and 900 PAD pins. Ultimately, we reduced the amount of manual description for generating configurable pin control block by 97 %.

Keywords Configurable pin control block · Design automation · Multimedia system-on-a-chip

M.-S. Kim (✉) · J.-L. Gaudiot
Department of EECS, The Henry Samueli School of Engineering, University of California,
Irvine, CA, USA
e-mail: myoungseo.kim@uci.edu

J.-L. Gaudiot
e-mail: gaudiot@uci.edu

1 Introduction

As the design productivity cannot follow the rapid advance of fabrication technologies, development of a new design methodology to improve design productivity has become necessary. In response to this need, the platform integration methodology has proposed the automation of platform integration and verification process [1, 2]. Particularly, commercial solutions of design automation for signal multiplexer and PAD block have been presented [3, 4]. However, several architectural parameters of such commercial solutions; such as the maximum number of I/O pins [5], PAD control signals [6], and various address regions; make it difficult to reuse generic pin control blocks. As a result, RTL designers rely on manual design work which is time-consuming and error-prone in a limited timeframe [7].

In this paper, we propose an automated design scheme of configurable pin control blocks. The key of this automated design scheme is that it processes a formalized text through parser which is programmed by script language. By using this approach, designers can automatically generate RTL blocks to reduce human errors and design time, while maintaining consistency among the generated outputs.

The remainder of this paper is organized as follows. We first briefly review related work in [Sect. 2](#). [Section 3](#) introduces the structure of generic pin control blocks. Then, detailed structure and functions of the proposed method are introduced in [Sect. 4](#). [Section 5](#) shows an experimental result of the proposed method, followed by the conclusion of this paper in [Sect. 6](#).

2 Related Work

Nowadays, complex SoCs include more peripheral interfaces in the core that can be accessed at one time. This requires a complex scheme having a flexible muxing strategy that allows pins to configure. As a result, peripheral interface designers spend much time designing complex I/O subsystems and then adapting them to inevitable specification changes during the course of a design.

IP-XACT (IEEE-1685) standard which was originally progressed by SPIRIT consortium [8] defines a way for describing and handling multi-sourced IP components. In addition, it enables automated design integration and configuration within multi-vendor design flows. The IP-XACT uses the form of an XML schema to define an IP meta-data description. This meta-data description represents an IP in several aspects such as hardware model, signals, bus interfaces, memory map, address space, and model views. This standard can also be used to describe a hierarchical subsystem resulting from an assembly of subcomponents.

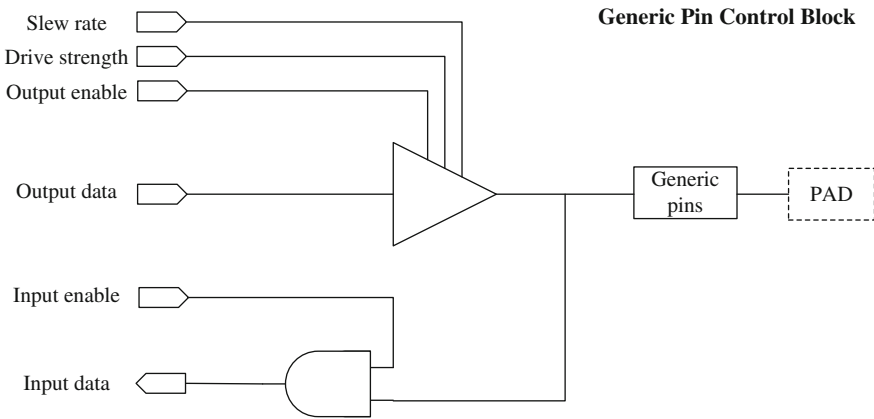


Fig. 1 Overall architecture of generic pin control block

3 Structure of Generic Pin Control Block

The overall architecture of generic pin control block is shown in Fig. 1. This architecture has multi-functional port pins organized in several groups. Each port can be easily configured by software to meet various system configuration and design requirements.

These multi-functional pins need to be properly configured before their use. If a multiplexed pin is not used as a dedicated functional pin, this pin can be configured as a generic input or output port.

When a pin is configured as an input port by output enable clear and input enable set, external signals are written on the input data register. Input enable always sets in a 65 nm PAD scheme due to one of the AND logic inputs. Hence, the other AND input logic becomes the prime determinant of the input data value. The state of the current input can be read from the input data register. Then, in case the pin is configured as an output port, the value in the output data register will drive the generic pin. Some pins need a special mode such as low-active input, bi-directional and low-active bi-directional mode. If an appropriate pin of input mode is not selected, the default value will set to 0. However, in case of the low-active input, the default value will set to 1. The pin of bi-directional mode has an output enable signal which is directly connected to the output enable signal of generic pin control blocks. As for the low-active bi-directional mode, the inverted signal of the output enable is connected to the output enable signal of general pin control blocks.

Table 1 Functions and parameters

Type	Function	Parameter
ARM I/O multiplexer	Data I/O through bus interface	Bus interface type Initial value Address region Register control
IP Blocks I/O multiplexer	Programmable I/O multiplexer for IP blocks	Control signal name I/O mode Module name
PAD Control	Pull-up/down Direction control Drive strength control	PAD control register

4 Specification with Formatted Description

In this section, we define the types of special registers according to functional requirements of configurable pin control blocks and propose a formalized text approach to unify various architectures for configurable pin control blocks.

4.1 Formatted Description

Table 1 shows the summary of functionalities and structural parameters that are essential to implement configurable pin control blocks. Our goal is to develop a formalized description approach to express all the functions and parameters described in Table 1.

4.2 Specific Functional Requirement

Except for the basic I/O function of generic pin control blocks, the proposed block also has additional functions that can efficiently control an interface of multimedia SoC platform. This module consists of specific registers for supporting additional functions as shown in Table 2.

4.2.1 Port Configuration Register

Table 3 displays the structure of PCON. Most pins are multiplexed and have up to 15 different functions. The function of each pin can be configured by 4 bits of PCON#. After reset, all PCON# have '0' value in ROM boot mode. Thus, all pins are controlled by PCON# work as input mode of normal generic pin.

Table 2 Composition of specific register group

Register	Address	R/W	Description	Initial Value
PCON#	0 × 000000#0	R/W	Configure the pins of port#	0 × 0000_0000
PDAT#	0 × 000000#4	R/W	Data register for port#	0 × 0000_00XX
PCTL#	0 × 000000#8	R/W	Control the pins of port#	0 × 0000_0000
PPUR#	0 × 000000#C	R/W	Pull enable register for port#	0 × 0000_0000
PPDR#	0 × 000000(# + 1)0	R/W	Pull selection register for port#	0 × 0000_0000

Table 3 Composition of PCON

Port#	Register component	Bits definition
Port#_0....7	PCON#[3:0] PCON#[31:28]	0000 = Function0 (Input) 0010 = Function2 (Output) 0100 = Function3 (IP's) 1111 = Function15 (IP's)

Table 4 Composition of PDAT

Port#	Register component	Bits definition
P#	PDAT#	{24'h0000_00, P#[7:0]}

4.2.2 Port Data Register

When ports are configured as output ports, data can be written on the corresponding bit of PDAT# as shown in Table 4. On the contrary, when ports are configured as input ports, the data can be read from the corresponding bit of PDAT#.

4.2.3 Port Control Register

For the port specified by PCTL# as described in Table 5, 1 in Set[15:8] will set the corresponding generic pin to 1, while 0 will cause no action. 1 in Clear[7:0] will set the corresponding generic pin to 0, while 0 will cause no action. If Set and Clear are both 1 at pin, then the value of the pin is toggled. This operation is performed only when the corresponding pins are designated as Output in PCON#.

4.2.4 Port Pull Up/Down Register

Pull up/down register as shown in Table 6 controls the pull up/down resistor enable/disable of each port group. When the corresponding bit is 1, the pull up/down resistor of the pin is enabled. When the corresponding bit is 0, the pull up/

Table 5 Composition of PCTL

Port#	Register component	Bits definition
P#	PCTL#	{16'h0000, Set[15:8], Clear[7:0]}

Table 6 Composition of PPUR/PPDR

Port#	Register component	Bits definition
P#	PPUR#/PPDR#	Pull up/down resistor control register {24'h0000_00, P#[7:0]} 0: Pull up/down resistor is disabled 1: Pull up/down resistor is enabled

Table 7 Composition of FSEL

Type	Register component	Bits definition
FSEL	PortNumber[20:16] PinNumber[10:8] Function[3:0]	Select function for a specific pin of a specific port

down resistor is disabled. It is prohibited to set the pull up register and pull down register to enable at the same time.

4.2.5 Function Selection Register

For the pin and port specified by PortNumber[20:16] and PinNumber[10:8] as described in Table 7, writing a number to Function[3:0] will copy the number to an appropriate place of PCONn, and immediately change the function of specified generic pin.

Writing 1110 and 1111 at Function[3:0] will cause 2 separate actions. For example, Writing 1110 at Function[3:0] will set the appropriate bits of PCONn to 0001, and will set the appropriate bits of PDATn[0] to 0. Writing 1111 will set the appropriate bits of PCONn to 0001, and will set the appropriate bits of PDATn[0] to 1.

5 Experiment Results

To demonstrate the feasibility of description capability for configurable pin control blocks with our proposed formalized text, we used a multimedia SoC platform that was in some Samsung projects based on 65 nm technology. Also, we chose python script language to make our proposed RTL blocks auto-generator. To evaluate the efficiency of our proposed design approach, we applied our automated design

Table 8 Composition of multimedia SoC platform

Item platform	MUX type	PAD control	Power control	Data control
Multimedia	8:1	3	No	Byte

Fig. 2 Quantitative analysis in a multimedia SoC platform (Bytes)

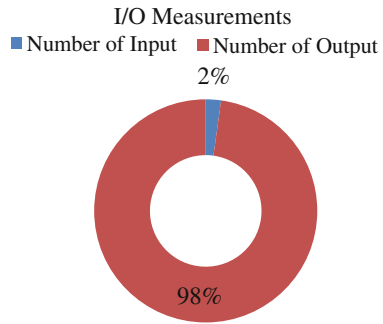
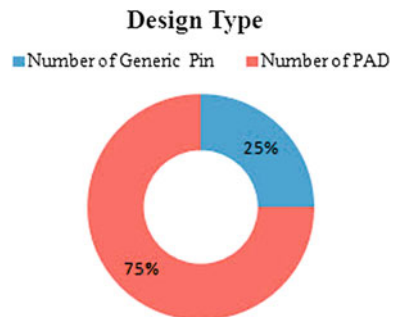


Fig. 3 Design volume in a multimedia SoC platform



scheme to a typical multimedia SoC platform. We generated an architecture model that has hardware characteristics as summarized in Table 8.

We measured the efficiency ratio of traditional RTL design time to RTL auto-generating time by our proposed RTL blocks auto-generator.

As a result, the amount of manual description for generating the configurable pin control block was reduced by 97 %, compared to the traditional RTL description. This automatic configurable pin control block generator is applicable to the different platform with simply changing the input formalized description text format. As shown in Figs. 2 and 3, we measured the size of specification input and the size of generated output to check the ratio in efficiency with 49 %.

Moreover, our proposed approach is able to generate RTL blocks automatically and the the possibility of human error can be reduced.

6 Conclusions

In this paper, we proposed an automated design scheme for configurable pin control blocks based on the process of a formalized through parser. By this formalized text, designers can specify an arbitrary configurable pin control block with minimal amount of description effort. In addition, designers can rapidly obtain RTL blocks by such an automated procedure. This can ultimately reduce design time and human errors significantly.

Experimental results show that it is possible to describe various configurable pin control architecture with the proposed description format. Lastly, the amount of manual description for designing configurable pin control blocks was reduced by 97 %, compared to the traditional RTL description.

References

1. Gajski D et al (2000) Essential issues for IP reuse. In: Proceeding of ASP-DAC, Jan 2000
2. Cho K et al (2008) Reusable platform design methodology for SoC integration and verification. In: Proceeding of ISOCC, Nov 2008
3. Gensys IO. <http://www.atrenta.com/solutions/gensys-family/gensys-io.htm>
4. Duolog Technology. <http://www.duolog.com/products/spinner/>
5. Vijayaraghaven N et al (2006) Novel architecture for on-chip AC characterization of I/Os. Proceeding of IEEE international test conference, Oct 2006
6. Zhang J et al (2004) Design and implementation of general purpose interface controller. Microelectron Comp, Oct 2004
7. Kruijtzter W et al (2008) Industrial IP integration flow based on IP-XACT standards. In: Proceedings of design, automation and test in Europe, March 2008
8. The SPIRIT consortium (2006) SPIRIT 1.2 specification. www.spiritconsortium.org. Accessed April 2006

Depression and Fatigue Analysis Using a Mental-Physical Model

Xue-Wei Tian, Zhen-Xing Zhang, Sang-Hong Lee, Hee-Jin Yoon
and Joon S. Lim

Abstract Recent research has indicated a significant association between depression and fatigue. To analyze depression and fatigue, an experiment was conducted that provided the subjects with affective content to induce a variety of emotions and heart rate variability (HRV). This paper presents a mental–physical model that describes the relationship between depression and fatigue by using a neuro-fuzzy network with a weighted fuzzy membership function using two time-domain and four frequency-domain features of HRV. HRV data were collected from 24 patients. At the end of the experiment, we determined the relationship between depression and fatigue with the mental–physical model, and our analysis results had an accuracy of 95.8 %.

Keywords Depression · Fatigue · Mental-physical · Electroencephalography (EEG) · Heart rate variability (HRV) · Takagi–Sugeno defuzzification

X.-W. Tian · S.-H. Lee · H.-J. Yoon · J. S. Lim (✉)
IT College, Gachon University, San 65 Bokjeong-dong, Sujeong-gu, Seongnam,
Gyeonggi-do, South Korea
e-mail: jslim@gachon.ac.kr

X.-W. Tian
e-mail: tianxuemaog@gmail.com

S.-H. Lee
e-mail: shleedosa@gmail.comshleedosa@gachon.ac.kr

H.-J. Yoon
e-mail: yhj68@hanmail.net

Z.-X. Zhang
School of Information and Electric Engineering, Ludong University, YanTai, China
e-mail: billzhenxing@gmail.com

1 Introduction

Depression is a common mood disorder. It is characterized by impairment of mood regulation and a loss of interest in enjoyable activities. However, owing to the fluctuating nature of the disorder, it is difficult to both diagnose and treat [1, 2]. In recent years, P600 components [3] and P300 variations [4] have been used to distinguish depression from normal controls based on event-related potential signals. However, disadvantages of electroencephalography (EEG) include the poor signal-to-noise ratio and complicated multi-channel EEG electrode. Hence, sophisticated EEG signal analysis consumes a large quantity of available CPU resources, which makes the application inappropriate for mobile and embedded applications.

Fatigue is measured and evaluated using EEG signals, which are analyzed while video display terminal tasks are carried out [1, 4]. Recently, a number of studies have indicated a significant association between fatigue and the cardiac autonomic nervous system (ANS) based on analyses of the heart rate variability (HRV) [5]. To properly evaluate and detect fatigue, the issue of measurement accuracy must be addressed. It is difficult to detect the symptoms of fatigue using the EEG signal of the subtle brain waveform, which is not sharp enough for the amplitude to be read correctly [4]. Statistical algorithms are used in many studies to analyze and evaluate fatigue based on HRV [5]. HRV are collected during the head-up tilt (HUT) test, which generally exceeds 30 min in length [5]. During the HUT test, the subjects lie on a tilt table, which is tilted in a head-up position at different levels. This HRV collection method is inconvenient as HRV data cannot be collected outside a hospital.

A depression–fatigue analysis method that uses a mental–physical model based on a neuro-fuzzy network was designed in this study. The subjects' HRV data were obtained during a 13-min affective-content video used to induce a variety of HRV. The process to gather HRV data used six subjects with fatigue and 14 control subjects. Six features were extracted from the time and frequency domain of the original HRV and used as input features of the neuro-fuzzy network. The results showed that the analysis accuracy was 95.8 %.

2 Materials and Method

2.1 Materials

In this study, 24 subjects were assessed via the Zung self-rating depression scale (SDS) and fatigue severity scale (FSS). The assessment results showed 11 subjects with no fatigue or depression, three subjects with fatigue but no depression, nine subjects with both depression and fatigue, and only one subject with depression but no fatigue. All subjects were 21–27 year old male computer science majors, and all were healthy volunteers without any history of heart disease or neurological or psychiatric illness. They did not exercise heavily 4 h before the experiment.

The SDS is a 20-item self-reporting questionnaire. Each item is scored on a Likert scale ranging from 1 to 4 (a little of the time, some of the time, a good part of the time, most of the time) [6]. The SDS score is the total value from all 20 items, and it ranges from 20 to 80. The scores fall into four common characteristics of depression: normal range (20–49), mildly depressed (50–59), moderately depressed (60–69), and severely depressed (≥ 70) [8].

The FSS, developed by Krupp, is based on a nine-item questionnaire assessing the reported fatigue. This scale shows the sensitivity and reliability of the fatigue assessment [1]. Each item is scored on a scale of 1–7, with 1 indicating strong disagreement and seven indicating agreement. The FSS score is the total value from all nine items, and scores of 36 or higher are indicators of fatigue [1].

To induce a variety of emotions and ANS values in the subjects, this study used the MAC stimulus, which can evoke various emotions such as happiness, joy, pain, stress, irritability, and fear. Each subject who underwent the MAC test ate some soft marshmallows and super-sour candies, drank sweet juice, and blew a balloon [1]. The MAC stimulus scenario is summarized in Table 1. The MAC test lasted for approximately 800 s (about 13 min). This included a 10 s transition time between the affective contents.

2.2 Depression and Fatigue Analysis by Mental–Physical Model

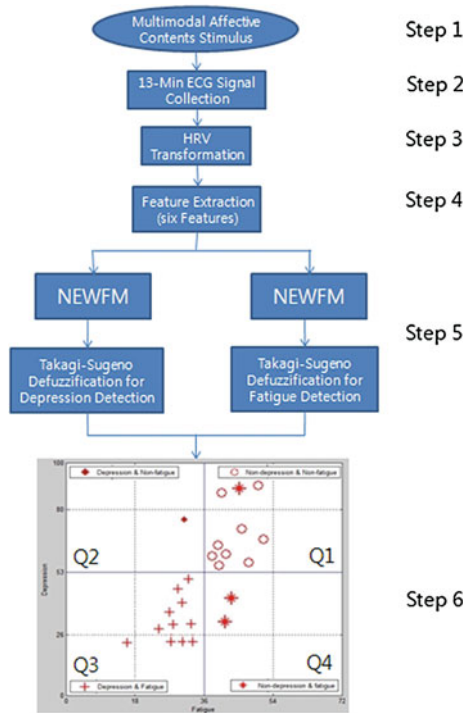
2.2.1 Pearson Correlation Analysis

First, we used the questionnaires on depression and fatigue to perform a Pearson correlation analysis. The Pearson correlation of depression and fatigue was 0.452, and the p value was 0.026, which is less than 0.05. Based on the above analysis, there is a relationship between depression and fatigue. Thus, we built a mental–physical model to determine the relationship between depression and fatigue.

2.2.2 Depression and Fatigue Analysis Method

An analysis model was used to prove the relationship between depression and fatigue based on a neuro-fuzzy network with a weighted fuzzy membership function (NEWFM) [7]. NEWFM is a supervised classification neuro-fuzzy system that uses the bounded sum of weighted fuzzy membership functions [7]. Figure 1 shows the structure of the depression and fatigue analysis method, which has six steps. Step 1 involved watching a 13-min affective-content video that can induce a variety of ANS and HRV values [1]. Each subject watched the affective-content video wearing a wireless Holter monitor. In step 2, EEG signals were collected for 13 min for further analysis. Step 3 was the HRV transformation process derived from 13-min ECG signals. HRV was measured through the variations in the beat-to-beat interval of ECG. In step 4, six features as NEWFM input features obtained via time–frequency domain extraction from the original HRV were examined [1].

Fig. 1 Depression and fatigue analysis method flowchart



In step 5, Takagi–Sugeno defuzzification values were obtained by using NEWFM. In step 6, the Takagi–Sugeno defuzzification values were used to determine the distribution for the depression and fatigue analysis model (i.e., mental–physical model) and the relationship between depression and fatigue. The mental–physical model has four quadrants: Q1–Q4. Q1 is for subjects with depression and fatigue in the same time; they are represented by circles. Q2 is for subjects with depression but no fatigue; they are represented by diamonds. Q3 is for subjects with no depression or fatigue; they are represented by plus signs. Q4 is for subjects with fatigue but no depression; they are represented by snowflakes.

3 Experimental Results

Figure 1 shows the results of the depression and fatigue analysis based on NEWFM using the mental–physical model after Step 6. There were 24 subjects in this experiment: 11 with no depression or fatigue, three with fatigue but no depression, nine with depression and fatigue, and one with depression but no fatigue. There was only one error in our experiment: a subject with fatigue but no depression was placed in quadrant Q1, which was the quadrant for subjects with depression and fatigue. The depression and fatigue analysis accuracy was 95.8 %.

4 Concluding Remarks

Using the Takagi–Sugeno defuzzification values, we determined depression and fatigue with the mental–physical model. The proposed system achieved 95.8 % overall analysis accuracy with the mental–physical model. This model visualizes the relationship between depression and fatigue. The relationship is that a subject with depression will also experience fatigue at the same time.

An earlier version of this paper was present at the [1].

Acknowledgments This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the Convergence-ITRC (Convergence Information Technology Research Center) support program (NIPA-2012-H0401-12-1001) supervised by the NIPA (National IT Industry Promotion Agency).

References

1. Zhang ZX, Tian XW, Lim JS (2012) Neuro-fuzzy network-based depression diagnosis algorithm using optimal features of HRV. *J Korea Contents Acad* 12(2):1–9
2. Edgerton JE, Campbell RE (1994) *American psychiatric glossary*, 7th edn. American Psychiatric Press, Washington DC
3. American Psychiatric Association (2000) *Diagnostic and statistical manual of mental disorders*, 4th edn. Text Revision: DSM-IV-TR, American Psychiatric Publishing, Washington
4. Uetake A, Murata A (2000) Assessment of mental fatigue during VDT task using event related potential (P300). In: *Proceedings of the 2000 IEEE international workshop on robot and human interactive communication*, pp 235–240
5. Atsuo M, Atsushi U, Yosuke T (2005) Evaluation of mental fatigue using feature parameter extracted from event related potential. *Int J Ind Ergon* 35:761–770
6. Zung WW (1965) A self-rating depression scale. *Arch Gen Psychiatry* 12:63–70
7. Lim JS (2009) Finding features for real-time premature ventricular contraction detection using a fuzzy neural network system. *IEEE Trans Neural Netw* 20(3):522–527

A Study of Non-photorealistic Rendering Method Using Orientation Sensor in Mobile Devices

Sungtae Kim, Minseok Kang, Jiyeon Kim, Hongil Kim, Gukboh Kim and Jongjin Jung

Abstract As mobile devices have been rapidly spread up throughout the world, various services have been provided through mobile devices recently. Many services of them have been developed by using built-in sensors mounted on the mobile devices. Non-photorealistic rendering techniques have been also applied to mobile devices for the purpose of user friendliness or attractions. This paper proposes non-photorealistic rendering method using built-in orientation sensor of mobile devices. The method makes changes of non-photorealistic rendering effects in real time by adjustment of vector fields according as coordinate values from orientation sensor get changed on a mobile device. The proposed method renders pen sketch for a given photo image with direction and strength of line strokes which are applied by vector fields. The method executes rendering based on enhanced LIC filter to make pen sketch on a mobile device. Finally, this paper proves efficiency through the experimental results and shows usefulness of the proposed method.

Keywords Non-photorealistic rendering · Pencil drawing filtering · Orientation sensor · Enhanced LIC · Vector field · Mobile device

S. Kim · M. Kang · J. Kim · H. Kim · G. Kim · J. Jung (✉)
Computer Engineering, Daejin University, Seondan-dong, Pocheon-si,
Gyeonggi-do, Korea
e-mail: jjjung@daejin.ac.kr

S. Kim
e-mail: kstlove123@daejin.ac.kr

J. Kim
e-mail: jini_69@naver.com

1 Introduction

Recently, many kinds of apps have been developed according as mobile devices such as smart phones, tablets and e-book terminals have been rapidly spread up throughout the world. The graphics are essential part for user convenience and for attraction in those apps. 2D/3D photorealistic rendering techniques of graphics have been mainly used in User Interface (UI), game, e-book and education, etc. but graphics has emerged apps using non-photorealistic rendering techniques to transform my photos in recent years. These techniques to the existing PC-based techniques, applied to the mobile devices environment. At this point, most researchers have focused on the rendering technologies optimized for low-end mobile devices even though it required a high level of processing power. In this paper, we propose a meaningful non-photorealistic technique using built-in orientation sensor considering computational optimization for low-level processing power of mobile devices. That is, we create new non-photorealistic rendering effects which are applied to the photos taken with camera embedded in mobile devices, or the images that are stored in the devices. The built-in orientation sensor of mobile device has a range of coordinate values on three axes. When the sensor operates by user's device shaking, it produces vectors corresponding to the coordinate values. We apply the vectors to parameters of pencil drawing filter which give change of line stroke. The previous researches for using orientation sensor have been some progressed in objects location detection [1] or virtual reality games [2].

This paper applies vectors from the orientation sensor to pencil drawing filter of non-photorealistic rendering technology. We implement automatic pencil drawing filter based on the algorithm of enhanced LIC pencil filter. Specifically, the proposed method focuses on the modified version of enhanced LIC pencil filter to optimize rendering time in low-end mobile device environment.

2 Related Works

2.1 Pencil Drawing Filter Techniques

Traditional computer graphics to create photorealistic rendering results have been applied in various field, in contrast non-photorealistic rendering techniques to render objects as a way to express a variety of unrealistic shapes have recently attracting attention. Non-photorealistic rendering is a method that converts 2D input image to painter image, pencil drawing image, cartoon image, water color image or ink painting image by image processing. In this paper, we focus on the pencil drawing filtering because the vectors of coordinate values from the oriental sensor are matched to line strokes well. The pencil drawing filtering uses tones and

textures to express contrast of object. Typical methods of pencil drawing filtering are hatching and LIC pencil filter.

Hatching is a shading effect processing to draw out lifetime line compactly. Vary the length and thickness of line, due to the difference in concentration of several tones of atmosphere. A method to express pencil hatching effect using blurring and sharpening was proposed in [3]. Line Integral Convolution (LIC) is a texture using vector field visualization technique [4]. Using a two-dimensional vector field and a white noise image, the directionality of the vector field generated LIC image which was stained through a low-pass filter in [5].

2.2 Non-photorealistic Rendering on Mobile Devices

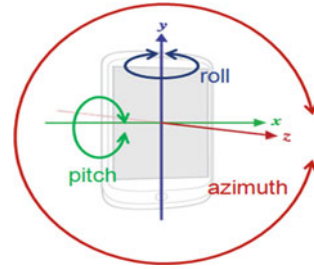
As mobile devices have been rapidly spread up throughout the world, many researchers have interested in the implementation of good quality graphics in the mobile environment. Preprocessing techniques to reduce the computational complexity in mobile environments has been studied in [6]. To reduce the computational complexity, they used silhouette edge file as part of the rendering process and shading pretreatment techniques. The researches for applying orientation sensor to graphics have been some progressed in objects location detection [1], or virtual reality games [2]. These researches are to create realistic effect. They can't be processed on desktop environment but are only possible to be carried out on mobile devices with orientation sensor.

2.3 Orientation Sensor-based Coordination System in Android Platform Devices

The movement of the mobile device can be detected by acceleration sensor and geomagnetic sensor in Android platform. We call them orientation sensor which implicates coordinate, direction and speed. Physical orientation sensor does not exist but is just values that are returned by internal calculation between the acceleration sensor and the geomagnetic sensor. The orientation sensor makes three values which consist of pitch, roll and azimuth. The coordinate axes of orientation sensor are defined as shown in Fig. 1.

In this paper, we use pitch values and roll values to adjust the direction of stroke in the pencil drawing filtering. We also use azimuth values to represent strength of stroke. The pitch rotates around the X-axis and has range of angles between -180 and 180 . The roll rotates around the Y-axis and has range of angles between -90 and 90 . The azimuth rotates around the Z-axis and has range of angles between 0 and 360 .

Fig. 1 Coordinate system of orientation sensor



3 Pencil Drawing Filter Using Orientation Sensor

In this paper, we propose pencil drawing filtering technique using orientation sensor. The proposed technique is based on the enhanced LIC technique to implement pencil drawing filtering [7].

3.1 Algorithm for the Proposed Filtering Effect

The enhanced LIC has been designed to complement the problems caused by the splitting image of the LIC pencil filter. LIC pencil filter uses image segmentation and texture direction detection for defining outlines and stroke directions. As the result, the method can be failed to generate LIC image if the result of image segmentation is not consistent with the source image structure because the quality of the result is dependent on the image segmentation. Therefore, enhanced LIC method creates intensity layers instead of image segmentation according to level of contrast in source image. The method applies line stroke to vector fields for each layer and overlaps all layers.

In this paper, we create a new pencil drawing filter effect using orientation sensor based on enhanced LIC filter. The algorithm of the proposed filtering effect is shown in Fig. 2.

In Fig. 2, the proposed method extracts outlines from the source image and then separates three levels of layers according to distribution of contrast in source image. The method generates vector fields for each layer. At this point, the values of orientation sensor affect to the vector fields. At first, both pitch and roll values are related to the direction of stroke in the pencil drawing filtering. To change the direction of stroke, we use pitch values and roll values, which are generated from movement of the mobile device. At second, the azimuth values are related to the strength of the stroke. To represent length of stroke, we use azimuth values from the sensor. Finally, the proposed method applies line stroke to vector fields for each layer and overlaps all layers. We explain the detailed process of our proposed method in the following sections.

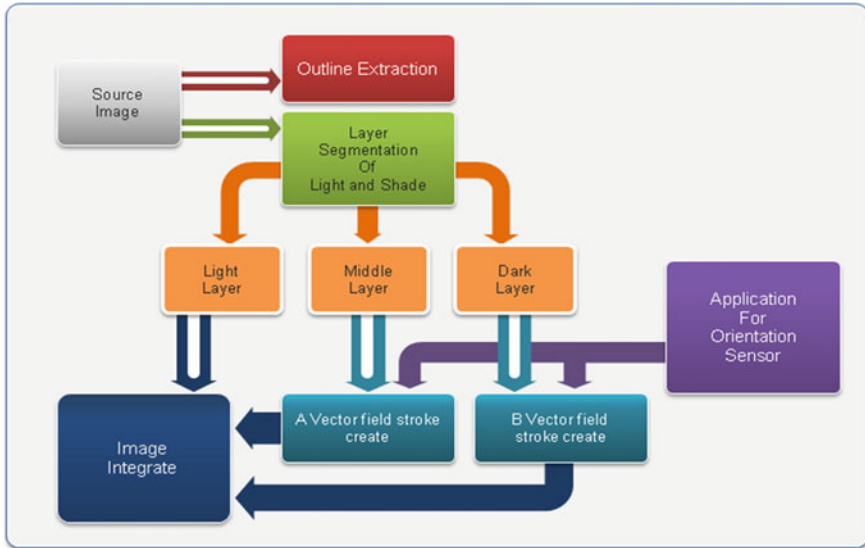


Fig. 2 Algorithm for the proposed filtering effect

3.2 Outline Extraction

Outline extraction is a basic procedure in non-photorealistic rendering. We extract outlines with Sobel mask in source image. Sobel mask is a general method for extracting outline. We compared Sobel mask with Kenny mask and employed Sobel mask in terms of quality. Figure 3 shows the result of outline extraction with Sobel mask.

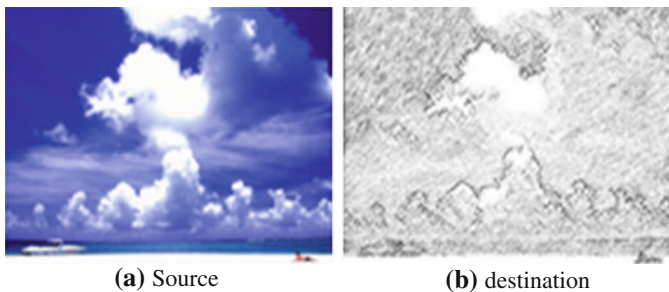


Fig. 3 Outline extraction (a) source (b) destination

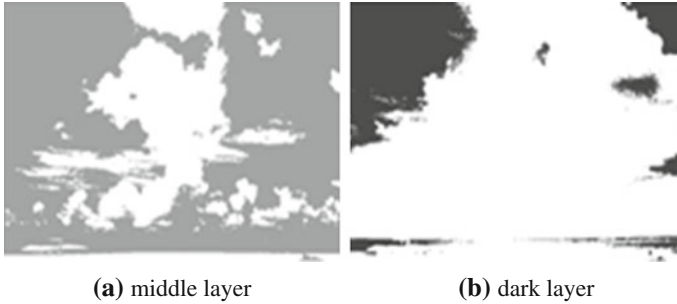


Fig. 4 Separation of layers by contrast distribution (a) middle layer (b) dark layer

3.3 Layer Separation by Contrast Distribution

The proposed method separates three levels of layers according to distribution of contrast in the source image. We referred to them as light layer, middle layer and dark layer. At this point, light is very important issue in separating layers according to distribution of contrast, because light affects difference of contrast. The levels of layers can be separated properly if light influence is minimized by homogenizing contrast in the source image. Figure 4 shows the result of layer separation by contrast distribution.

We divide into three parts with frequency of brightness for an image as the bright area corresponding to the light layer with bright color values than 1/3 of whole area, the dark area corresponding to the dark layer that has color values darker than 2/3 point and the rest of the medium corresponding to the middle layer.

3.4 Vector Field Change and Stroke Strength Control by Orientation Sensor

LIC pencil filter method suggests image distortion which uses vector fields to white noise in order to sketch a feeling. This method makes hatching effect for source image similar to actual pencil drawing. In this paper, we use pitch values and roll values to change direction of the strokes and use azimuth values to control strengths of stroke from the orientation sensor in mobile device.

To change direction of the stroke, the proposed method generates vector fields through transformation from the coordinate values to vectors for each pitch and roll. As the result, vector fields are made by transformed vectors. Then, the method applies these vector fields to middle layer added white noise and to dark layer added white noise respectively. So, hatching image can be made by these processes. We can take three types of values, pitch values for X-axis, roll values for Y-axis, azimuth values for Z-axis from orientation sensor. These values mean angle variation. So, the

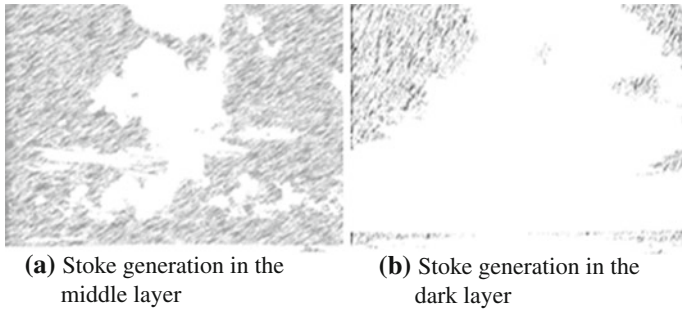


Fig. 5 Stroke effect by vector field per layer (a) Stoke generation in the middle layer (b) Stoke generation in the dark layer

proposed method transforms angle variation to scalar and calculates slope of normal vector. Then it makes A vector for middle layer and B vector for dark layer. Finally, the proposed method generates vector fields of A and B by repetition of these operations. The proposed method generates line strokes by applying white noise to source image according to the direction of vector fields after vector fields is formed by the above operation. The proposed method does not distort image with white noise according to the direction of vector fields after generating white noise first but produces white noise lines according to the direction of vector fields. This is to reduce computation. We focus on computational optimization in low-end mobile devices.

In Fig. 5, strokes of (a) is different to strokes of (b), because the direction of strokes is dependent to the direction of the vector fields. The azimuth values are corresponding to scalar of vectors to control strength of stroke. The scalar of vector determines length of white noise line to distort the image. If the azimuth value is larger, the stroke is longer because the azimuth value affects length of vector which makes white noise. Therefore, we can represent various strengths of strokes with azimuth values.

3.5 Integration of Layer-specific Strokes

We complete light layer which contains length of strokes, middle layer which is hatched by A vector fields and dark layer which is hatched by B vector fields. Then both A vector fields and B vector fields can be changed according to orientation sensor in mobile devices. The length of strokes can be also changed according to orientation sensor. Therefore, movement of orientation sensor will change feeling and sharpness of the sketch.

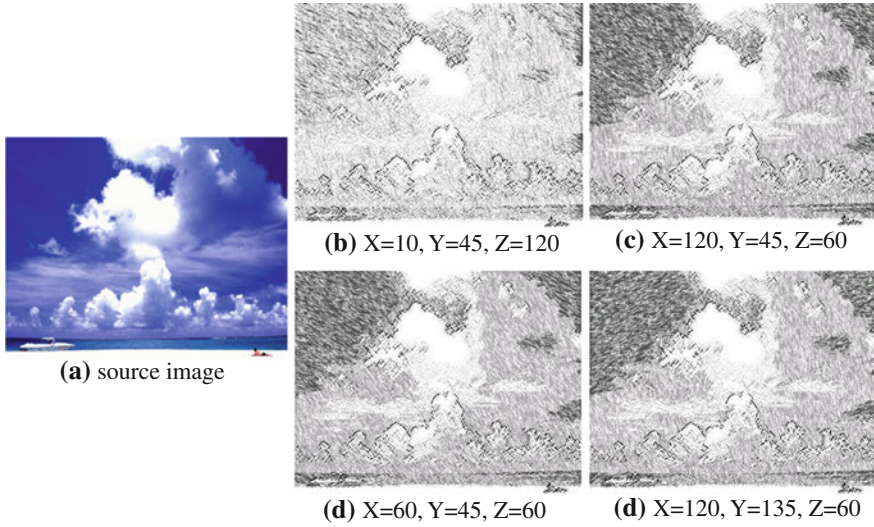


Fig. 6 Rendered results according to the changes in orientation sensor (a) source image (b) $X = 10, Y = 45, Z = 120$ (c) $X = 120, Y = 45, Z = 60$ (d) $X = 60, Y = 45, Z = 60$ (e) $X = 120, Y = 135, Z = 60$

4 Experimentation and Evaluation

In this paper, we implemented the proposed method in form of Android app. We used OpenCV 2.0 on Galaxy 2 1.2 GHz dual-core processor and Android Gingerbread 2.6 platform. We examined efficiency of the proposed method and evaluated usefulness of our developed app. Figure 6 shows the rendered results according to changes in orientation sensor.

We can see that the stroke of image appears differently according to the combination of X, Y, Z values is changed by applying value of the orientation sensor. X value of the change represents a direction change of the stroke in the middle layer. Y value of the change represents a direction change of the stroke in the dark layer. Finally, Z value of the change represents a length change of the stroke.

5 Conclusion

In this paper, we studied non-photorealistic rendering method using orientation sensor mounted on a mobile devices to apply pencil drawing filtering effect to photo image. In the proposed method, vector fields were generated by pitch values and roll values from movement of orientation sensor and accordingly the direction of the strokes was created differently. The proposed method did not rendered for

an image only one time but made different rendering results in real time whenever movement of the orientation sensor were occurred. However, the proposed method not yet optimized for the mobile device environment.

We are currently studying non-photorealistic rendering optimization techniques adapted to the low-end mobile device environment. In addition, we will study other non-photorealistic rendering methods using orientation sensor.

Acknowledgments This work was supported by the Daejin University Research Grants in 2012.

References

1. Kim E-G, Yeom M-R, Kim J-H (2011) Development of android-based application for measure a space coordinate and an area using of orientation sensor. *J Korean Assoc Inf Educ* 439–447
2. Yong H-j, Back J-w, Jang T-J (2006) Stereo vision based virtual reality game using position/orientation sensors and vibrotactile device. In: *Proceedings of the conference on Korean human computer interaction*, pp 88–93
3. Ma J-y, Yong H-s, Park J, Yoon K (2005) Pencil hatching effect using sharpening and blurring spatial filter. *Korea Comp Graphics Soc* 11(1):8–12
4. Cabral B, Leecom C (1993) Imaging vector fields using line integral convolution. In: *Proceedings of SIGGRAPH93 conference*, pp 263–270
5. Mao X, Nagasaka Y, Imamiya A (2001) Automatic generation of pencil drawing from 2D images using line integral convolution. In: *Proceedings of 7th international conference on computer aided design and computer graphics CAD/GRAPHICS2001*, pp 240–248
6. Jeon J, Jang H, Choy Y-C (2011) Processing techniques for non-photorealistic contents rendering in mobile. *J Korea Contents Assoc* 11(3):119–129
7. Mao X, Nagasaka Y, Imamiya A (2004) Enhanced LIC pencil filter. In: *Proceedings of the international conference on computer graphics, imaging and visualization, 2004*
8. Gooch A, Gooch B (2001) *Non-photorealistic rendering*. AK Peters Ltd, Wellesley

A BCI Contents Development Method Based Templates

Yunsick Sung, Kyungeun Cho and Kyhyun Um

Abstract With the drop in the price of Brain-Computer Interface (BCI) devices, which have been widely used in the medical sector, the development of serious BCI-based games has been accelerated. However, developers involved with the development of these games have found it difficult to acquire brain- and brain-wave-related knowledge. This paper defines templates that are necessary for developing BCI contents and proposes a method for developing BCI contents; the definition of the knowledge is based on templates. We present an example of a BCI-based game that has been developed using the proposed method. Since technical knowledge required for the development of serious games has been defined on the basis of templates, a developer can concentrate on the task of developing games.

Keywords Brain-computer interface · BCI contents · BCI game · Brainwave

1 Introduction

With the recent availability of low-price measurement devices with a brain-computer interface, BCI contents have drawn considerable attention [1]. In particular, BCI technology that has been widely used in the medical sector has been

Y. Sung

Department of Game Engineering, Graduate School, Dongguk University,
26, Pil-dong 3-ga, Jung-gu Seoul, South Korea
e-mail: cke@dongguk.edu

K. Cho (✉) · K. Um

Multimedia Engineering, Dongguk University, 26, Pil-dong 3-ga,
Jung-gu Seoul, South Korea
e-mail: cke@dongguk.edu

used in games, which has in turn accelerated the development of serious games for patients [2, 3]. In patients with dementia, for example, the amplitudes of alpha and beta waves are lower than those in the case of a normal person, while the amplitudes of theta and delta waves are higher [4]. Even though dementia is incurable, a game that increases the amplitudes of alpha and beta waves and lowers those of the theta and delta waves has been developed to prevent the disease and slow down its progress. A difficulty that is faced in the development of a serious game using a BCI is that the BCI-content developer needs to plan and develop the contents after acquiring extensive knowledge on the brain and brainwaves. For the easy development of serious BCI-based games, brain- and brainwave-related contents should be properly defined by an expert in advance.

This paper proposes templates that define knowledge for developing brain- or brainwave-related contents and proposes a method for the development of BCI contents. The proposed method does not require a developer to obtain knowledge on brain or brainwaves; this knowledge can be provided by an expert through templates. If the proposed method is adopted, it will be more convenient to modulate functions that are necessary for the development of BCI contents, which would in turn make it easy to develop a multifunctional serious game.

This paper is structured as follows: Sect. 2 describes past researches on the use of brainwaves for the development of BCI contents. Section 3 introduces a framework for the development of BCI contents using templates. In Sect. 4, a case study on the proposed framework is presented. In Sect. 5, the results of this study are summarized.

2 Related Work

Generally, brainwaves measured with a BCI can be used to control game characters or the progress of a game on the basis of a critical point. In this section, past researches that have used brainwaves for controlling the game character and papers pertaining to BCI contents are discussed.

There has been a study on a wheelchair control method for the handicapped [5]; this study is one of the studies on brainwave-based hardware control. In the study, brainwaves are used to determine the destination of a wheelchair. On the other hand, an example of a study on brainwave-based software control is [6], which is a study on sound control. In addition, there have been studies for ALS, ischemic stroke patients, and epileptic patients, on the use of the Internet for aiding the handicapped [7].

Past studies on frameworks for the development of BCI-based contents are as follows. First, to find a solution to the problem caused by the difference between the environment in which measuring devices are developed and that in which BCI contents are developed, a study was performed to determine a framework in which brainwaves can be transmitted in XML format through UDP [8]. In some cases, COM technology is used for the compatibility of various brainwave-measuring

devices [9]. In [9], a method to trigger events using brainwaves and a modulation method for signal handling are presented. There has also been a paper in which a brainwave handling process and authoring tools for the development of BCI contents have been presented [10]. In fact, these studies have helped in enhancing the efficiency of the methods used for developing BCI contents by making it easy to handle diverse types of brainwaves. However, it is also necessary to examine how to effectively describe and use an expert's knowledge of the brain and brainwaves.

This study proposes a method to describe the knowledge using templates and simplifies a developer's task, enabling him/her to concentrate on making the game more interesting.

3 Template Framework

In order to efficiently develop BCI contents, a developer requires a systematic brainwave-handling method. This paper proposes a BCI template framework in which brainwave-processing functions are modulated using templates. The paper further discusses the development of BCI contents using the modulated functions.

In a BCI template framework, BCI contents are developed in four stages, as shown in Fig. 1. First, by using a template authoring tool, an expert on brain/brainwaves defines templates that can be used for developing BCI contents. A template is a basic functional unit that can be used to handle brainwaves. It can be of two types: one is used to control brainwave-measuring devices and the other is used to control the progress of a game. The former defines a function to fetch brainwaves measured by a brainwave-measuring device, while the latter offers an interface to use brainwaves in a game.

A content authoring tool offers a function to create a game using a template defined as a template authoring tool. Using the latter tool, a BCI content developer prepares more than one template and sets the attributes that are necessary for executing each template. For example, brainwaves are measured by using the template for brainwave measuring devices, while the character or progress of a game is controlled by using the template for game progress.

Once template setting is done using the authoring tool, it is necessary to generate a template code that can be integrated with the BCI content code. The template defined as a template authoring tool is generated using the value set in the content authoring tool.

The BCI content code is used to create contents; for this, the template for measurement of brainwaves and the template for game progress are used. The template authoring tool and content authoring tool are configured as shown in Fig. 2. The former consists of a template edit tool, template creation tool, and template management tool. The template edit tool offers a user interface that can be used to define a template. The defined template is created using the template creation tool. When the template for brainwave measuring devices is used, it is

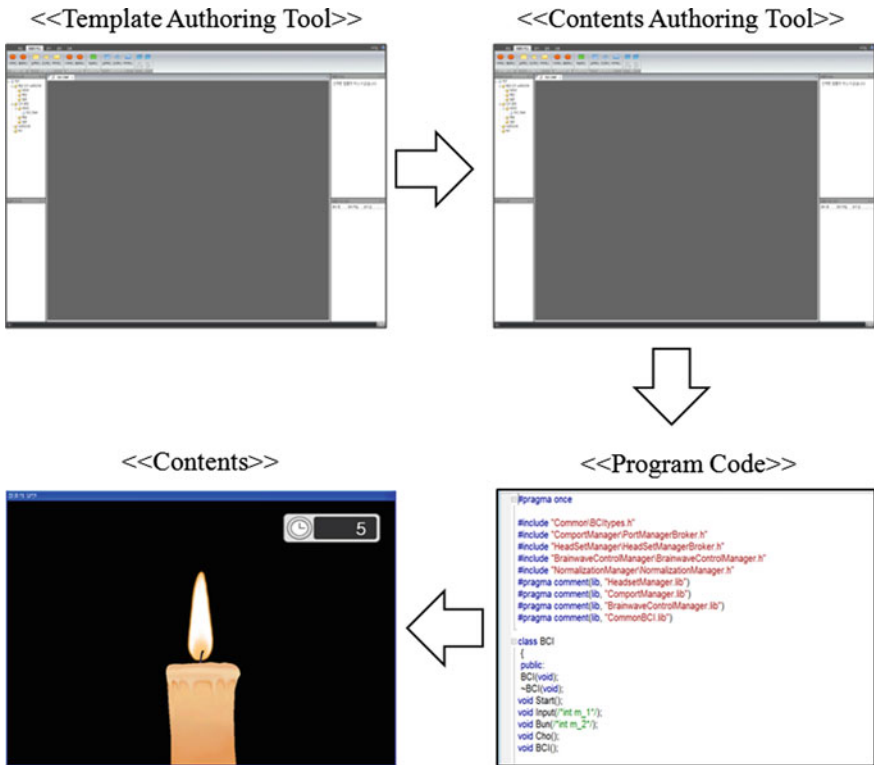


Fig. 1 Four stages in the BCI template framework. First, a template is defined as a template authoring tool. Then, the template attribute is set by content authoring tool. The template codes generated are integrated with BCI content codes

interlocked to measure brainwaves through the BCI module. Lastly, templates are stored in the database using the template management tool. The content authoring tool consists of a template management tool, template setting tool, template integration tool, and code generator tool. The template management tool is the same as the template management tool of the template authoring tool. It fetches templates from the database. The template setting tool offers a user interface that sets templates that have been determined to be appropriate for the development of BCI contents. It also provides a function to define a new template using more than one template. Lastly, the code generator tool generates codes from templates.

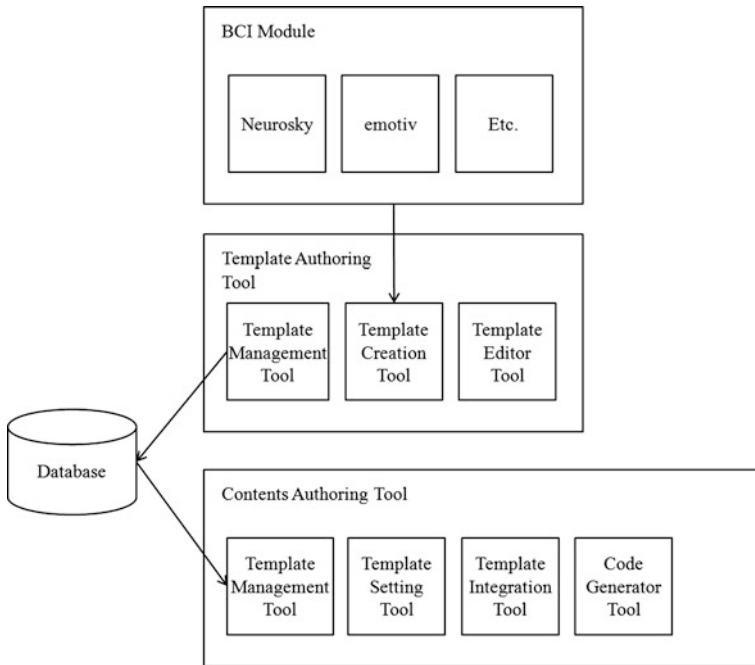


Fig. 2 BCI template framework. It consists of a template authoring tool and content authoring tool. The former involves the use of a BCI module; it consists of three tools and defines templates. In contrast, the content authoring tool consists of four tools and sets the templates required for developing BCI contents

4 Experiment

In this section, an example of a game developed using the proposed method is presented (Figs. 3, 4).

5 Conclusion

This paper proposes a method for developing a serious BCI-based game using templates after acquiring considerable knowledge on the brain/brainwaves. The efficient development of a game is possible by clearly stipulating the roles of the expert and the developer. In addition, the components and functions of the template authoring tool and content authoring tool have been stated. A case study on the development of BCI contents for a game titled “Master of Concentration” has been presented. A method to define templates using the template authoring tool has

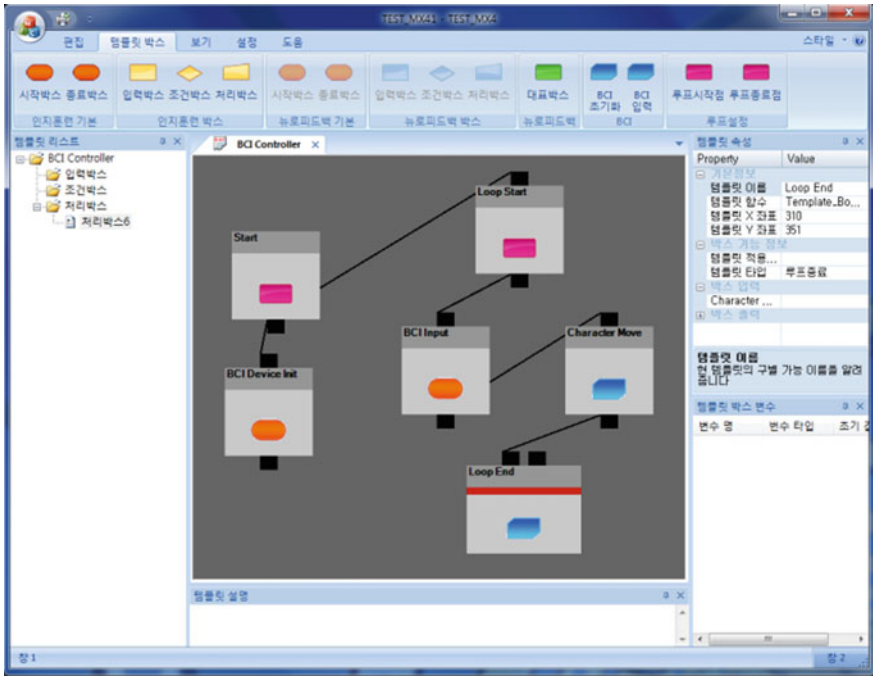


Fig. 3 In the template authoring process, the progress of the game is described by using a flow chart and the values that are necessary in each stage of progress is set in the property window. To describe the progress of the game, specific modules are defined after connecting modules in lines

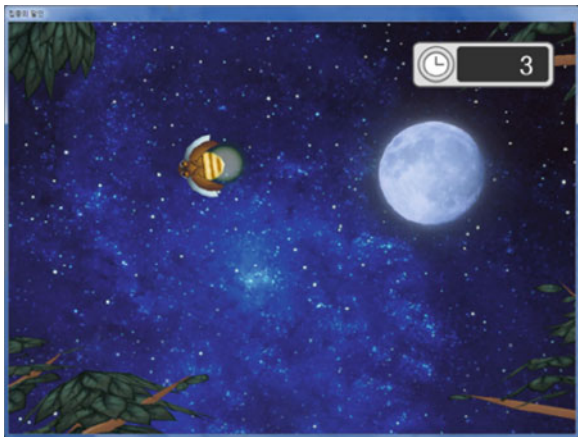


Fig. 4 It is a game whose attributes are generated using template authoring tool and set using the content authoring tool. It consists of three mini games. In the firefly game, a bug becomes brighter when the concentration of the player increases, and the bug becomes darker with a decrease in concentration. If a specific status continues for a while, the game ends. To distract the player, different types of distractions such as the sound of shooting stars or flying objects are generated

been explained, and the developed contents have been presented. Further studies should be conducted to develop a method for defining templates in a more systematic manner. Such a method will help in developing diverse serious BCI-based games.

Acknowledgments This paper is presented by summarizing part of the research, “A Development of Serious Game Technology using Brain Computer Interface” that was supported by Korea Creative Content Agency (KOCCA) in the Culture Technology(CT) Research & Development Program 2009.

References

1. Neurosky. www.neurosky.com
2. Lubar JO, Lubar JF (1984) Electroencephalographic biofeedback of SMR and beta for treatment of attention deficit disorders in a clinical setting. *Biofeedback Self Regul* 9(1):1–23
3. Nijholt A, Oude BD, Reuderink B (2009) Turning shortcomings into challenges: brain-computer interfaces for games. *Entertain Comput* 1(2):85–94
4. Jeong J (2004) EEG dynamics in patients with Alzheimer’s disease. *Clin Neurophysiol* 115(7):1490–1505
5. Iturrate I, Antelis JM, Kübler A, Minguez J (2009) A non-invasive brain-actuated wheelchair based on a P300 neurophysiological protocol and automated navigation. *IEEE Trans Rob* 25(3):614–627
6. Miranda ER, Soucaret V (2008) Mix-it-yourself with a brain-computer music interface. In: *Proceeding of 7th ICDVRAT with art abilitation*
7. Bensch M et al (2007) Nessi: an eeg-controlled web browser for severely paralyzed patients. *Computat Intell Neurosci* 2007:1–5
8. Venthur B, Blankertz B (2008) A platform-independent open-source feedback framework for BCI systems, In: *Proceedings of the 4th International Brain-Computer Interface Workshop and Training Course 2008*, pp 385–389
9. Jiang D, Yin J (2009) Research of auxiliary game platform based on BCI technology. In: *Proceedings of the Asia-Pacific conference on information processing, Vol 1*. pp 424–428
10. Renard Y et al (2010) OpenViBE: an open-source software platform to design, test, and use brain-computer interfaces in real and virtual environment. *Presence Teleoperations and Virtual Environments* 19(1):35–53

Pen-Ink Rendering for Traditional Building Images

Dokyung Shin and Eunyoung Ahn

Abstract This investigation proposes a non-photo realistic rendering to reflect enough to emphasize the delicate pen touch of Korea traditional building images. The earlier studies for pen-ink sketch technologies produce good results for the general images but they should not make successful results for the images of traditional architectures. It because that the image comprises a large proportion of repetitive patterns of exquisite lines and curves in its roof and latticed door. For featuring out the distinguishable characteristics of the traditional building images, the important thing is to decide the direction and length of strokes. We suggest a method to extract stroke's information using Tensor Subspace Analysis (TSA) technique and to draw pen-ink strokes according to the information. Therefore, the proposed method generates pen-ink sketches reflecting the texture property for the traditional building images and emphasizing the characteristics of light and shade for each area.

Keywords Non-photo Realistic Rendering (NPR) • Pen and ink sketch
Pen touch • Stroke

D. Shin

Department of Computer Science and Engineering, Hanyang University,
1271 Sa 3-Dong Sangnok-gu, Aansan, South Korea
e-mail: dkshin@cse.hanyang.ac.kr

E. Ahn (✉)

Department of Information Communication Engineering, Hanbat National University,
San 16-1, Deokmyeong-dong, Yuseong-gu, Daejeon, South Korea
e-mail: aey@hanbat.ac.kr

1 Introduction

The purpose of Non-photo Realistic Rendering (NPR) [1, 2] is to create images that could produce the very characteristics and realistic feelings of hand-drawings by human in order to produce the feelings of human-friendly drawing arts. The NPR technology started to draw attentions from the 1990s, and it is now taking a major part in computer graphics, and many NPR researches have tried to devise art media expression methods. The NPR technology is also increasingly applied to animations, advertisements, games, medical areas and architectures.

In the NPR technology, determination of stroke directions is essential because the final image to describe various types of drawing art is produced based on the strokes. In two-dimensional images, uniform direction filters or gradient-based direction filters are used to determine stroke direction. The uniform direction filter applies direction filters of the same direction. Therefore, it cannot reflect the overall property of the image and produces unnatural output. Furthermore, when gradient-based direction filter is used, stroke directions are created mainly on edge areas where the difference between gradients is large, and then it is impossible to create strokes in the area without the surrounding outlines.

In this paper, with the artist Youngtaek Kim's Korean style pen and ink sketch technique as a motif, we present a new method to redescribe Korean traditional architecture images into pen and ink sketch using computer graphic techniques. Therefore, in order to express the image with pen and ink sketch techniques from an input image, methods to determine the directions of pen touches and describe the beauty of blank areas are very important. A new pen and ink sketch technique is presented to compensate problems in existing pen and ink sketch generation techniques where pen touch directions of all area are described with just a single direction with which the characteristics of each area could not be expressed enough.

2 Proposed Method

2.1 White Noise Generation Technique

In this research, gray level of an input image is extracted by creating white noise from the input image in order to match the tone of input image and the tone of pen and ink sketch output. White noise is generated as a probability value of pixel set's white pixel value proportional to pixel's brightness value corresponding to the input image pixel. I_{input} means pixel's brightness value in the input image, and P means the real number value of input pixel created with random function. Pixel's I_{noise} value corresponding to the image is determined as in the following the Eq. (1). I_{max} is the biggest index value in an input image and the value is 255 in general. k is a coefficient to adjust the overall output tone, and the larger the value

of k is, the more the noise is created. An area with white noise is a bright tone area with reflection of light, and it is depicted as a blank area where either light stroke or no stroke is created over it. The area outside the white noise means an area with relatively dark tone, and it is classified as an area described with strong strokes.

$$I_{noise} = \begin{cases} I_{max} & \text{if } P \leq T \\ 0 & \text{otherwise} \end{cases} \quad P \in [0.0, 1.0] \quad (1)$$

$$T = k \left(\frac{1 - I_{input}}{I_{max}} \right) \quad k \in (0.0, 1.0)$$

2.2 Method to Generate Stroke with Tensor

In general, pen and ink sketch basically consists of repeated creation of lines. It maintains relatively regular line distances and strengths, and information for light and shade, texture and morphology are specified by using batching technique. Hatching technique is one of the expression techniques for printings or drawings, and it is used to depict light and shade of an object with fine and thin parallel lines or cross lines.

In this paper, Tensor Subspace Analysis (TSA) technique [3–5] is used to extract directional property information in an image, and the information determines direction and length of strokes. TSA describes an image with second degree tensor, and the spatial correlation between column vectors and row vectors of image matrix is characterized by TSA. Also, by learning low level tensor space, it creates intrinsic geometric structure of tensor space. In a diffusion tensor image, if a specific property vector within a specific array has the same probability value for several directions, it is called isotropic diffusion. If it indicates a specific uniform direction, it is called anisotropic diffusion. Anisotropic diffusion tensor can be displayed in ellipsoid diagram, and the ellipsoid figure of diffusion tensor is determined by eigenvalues on the three axes (eigenvalue: λ_1, λ_2) which forms the principal coordinate axes of the ellipsoid and eigenvector (v_1, v_2).

3 The Experimental Result

In this thesis, in order to depict input images of traditional architecture with pen and ink sketching techniques, white noises are generated and principal coordination vectors over tensor space are extracted to determine direction and length of pen touches. Figure 1a is the input image, Fig. 1b is output image after applying Photoshop sketching technique, and Fig. 1c is the output images after creating strokes with the proposed method.

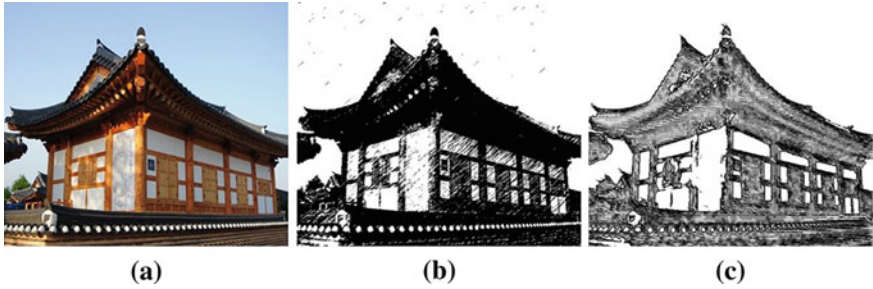


Fig. 1 Pen and ink sketch output for traditional architecture image. **a** Input image. **b** Output image after photoshop sketching effect. **c** Proposed result

The method proposed in this paper creates white noise, which makes it possible to describe light and shade of an area created by lights. Furthermore, It determines whether strokes should be created or not with the texture characteristics. Strokes are created for the area with strong texture, but strokes are not created or left as blank for the area with light texture. As a result, the advantage of our proposed technique is in providing effects of emphasizing the beauty of blankness, as frequently used in Oriental painting techniques. For the information to create strokes, length of stork is determined by direction of principal coordination axis for each area and proportion between vertical vectors of principal coordination axis. Therefore, since strokes are created according to texture property, the most distinctive areas in traditional architectures such as roof tiles and edges of eaves can be depicted with fine and detailed pen touches, and even the areas with curvature property can be described with sharp and straight line pen touches in this method.

4 Conclusions

We propose an expression technique as one of the non-photo realistic rendering techniques to produce the sense of pen and ink sketches for the images of traditional architecture. The sketching method of pen and ink sketches is different from other drawing methods in that it does not consider color information and depicts light and shade, texture and morphology of an image with only thin and sharp lines using pen.

Therefore, it is difficult to maximize the feeling of pen and ink sketch for traditional architecture with the existing pen and ink sketch expression techniques. There are areas with repetitive and strong linear elements in traditional architecture images in comparison with general image, and it is difficult to effectively describe exquisite lines of roof and eaves by extracting only general outlines. Therefore, the proposed method generates the feeling of depth in a pen and ink sketch by emphasizing the characteristics of light and shade in the given image and defines stroke directions from texture property information for each area.

Acknowledgments This research is supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2010-0021154).

References

1. Gooch A, Gooch B, Shirley P, Cohen E (1998) A Non-photorealistic lighting model for automatic technical illustration. In: Proceedings SIGGRAPH'98, Computer graphics proceedings, Annual Conference Series, 447–452 1998
2. Gooch B, Gooch A (2001) Non-photorealistic rendering. AK Peters, Natick
3. Tao D, Li X, Wu X, Maybank S (2007) General tensor discriminant analysis and Gabor features for gait recognition. IEEE Trans Pattern Anal Mach Intell 29(10):1700–1715
4. Hotz I, Feng L, Hagen H, Hamann B, Jeremic B, Joy KI (2004) Physically based methods for tensor field visualization. Proceedings 15th IEEE conference, visualization (VIS'04), pp 123–130
5. Kindlmann G, Westin CF (2006) Diffusion tensor visualization with glyph packing. IEEE Tran. Vis Comput Graph 12(5):1329–1336

Context-Aware Statistical Inference System for Effective Object Recognition

Sung-Kwan Kang, Kyung-Yong Chung, Kee-Wook Rim
and Jung-Hyun Lee

Abstract This paper proposes a statistical ontology approach for adaptive object recognition in a situation-variant environment. In this paper, we introduce a new concept, statistical ontology, for context sensitivity, as we found that many developed systems work in a context-invariant environment. Due to the effects of illumination on a supreme obstinate designing context-sensitive recognition system, we have focused on designing such a context-variant system using statistical ontology. Ontology can be defined as an explicit specification of conceptualization of a domain typically captured in an abstract model of how people think about things in the domain. People produce ontologies to understand and explain underlying principles and environmental factors. In this research, we have proposed context ontology, context modeling, context adaptation, and context categorization to design ontology based on illumination criteria. After selecting the proper ontology domain, we benefit from selecting a set of actions that produces

S.-K. Kang

HCI Laboratory, Department of Computer Science and Engineering,
Inha University, Yonghyeon 1,4-dong, Nam-gu, Incheon, Korea
e-mail: kskk1111@empas.com

K.-Y. Chung (✉)

School of Computer Information Engineering, Sangji University,
83 Sangjidae-gil, Wonju-si, Gangwon-do, Korea
e-mail: dragonhci@hanmail.net

K.-W. Rim

Department of Computer Science and Engineering, Sunmoon University, Galsan-ri,
Tangeong-myeon, Asan-si, Chungcheongnam-do, Korea
e-mail: rim@sunmoon.ac.kr

J.-H. Lee

Department of Computer Science and Engineering, Inha University,
Yonghyeon 1,4-dong, Nam-gu, Incheon, Korea
e-mail: jhlee@inha.ac.kr

better performance on that domain. We have carried out extensive experiments on these concepts in the area of object recognition in a dynamic changing environment, and we have achieved enormous success, which will enable us to proceed on our basic concepts.

Keywords Object recognition · Context-awareness · Context modeling

1 Introduction

Rapid advancements in sensor information processing evince the promise of providing information reliably and inexpensively that can be employed for intensive research and investigation. After all, pervasive, home net telematics and ubiquitous computing systems all use visual sensors to read and input data. To make a robust system for the purposes mentioned above, those applications and services must be aware of and adapt to their changing contexts in highly dynamic environments. The aim of all research on this topic is to improve the accuracy of vision systems using any method. However, improvement of accuracy is not an easy task, especially when a small change in the environment creates many of changes in the environmental context. The aim of this paper is to build a statistical ontology on illumination [1, 2], to improve the accuracy of adaptive object recognition systems with the help of ontology for context-aware systems. Most recognition systems, ubiquitous systems, or pervasive computing systems are adversely affected by illumination as a small change in illumination in dynamic environments produces a varying image, which leads to a loss or degradation of a system's performance.

In this paper, we present a statistical ontology method to introduce the concept of environmental situation organization based on illumination in terms of ontology to achieve highly efficient object recognition. Here, ontology distinguishes the illumination variations of input images, repeatedly using an unsupervised learning method, and derives an illumination image category that is organized in ontology fashion and referred to as illumination ontology. The system also constructs a classifier set for each illumination category for effective exploration of the genetic algorithm [3, 4], search space of various classifier systems. The classifier set is encoded in terms of an artificial chromosome called action reconfiguration chromosome. The genetic algorithm is used to explore the most effective classifier system structure for each identified data context category. The knowledge of an individual context category and its associated chromosomes of effective classifiers are stored in the knowledge base. Once sufficient context knowledge is accumulated, the method can react to such variations in real time. This illumination ontology concept is applied for context-awareness in dynamic changing environments, the classifier set is selected according to that context, and the selected classifier set is applied for object recognition.

2 Context-Aware Inference System

A context-aware system requires context information to be exchanged and used between different entities such as users, devices, and services in the same semantic understanding. Ontology is an explicit specification of a conceptualization [5]. It has a long history in philosophy, in which it refers to the concept of existence. Ontology can be thought of as a description of the relational structure of concepts for the purpose of enabling knowledge sharing and reuse. Ontology is often equated with taxonomic hierarchies of classes regarding its usage.

Context-aware ontology should be able to capture all the characteristics of context information. First, it is responsible for capturing a great variety of contexts. Second, the separation of domains would be in a meaningful way. In this research, environmental context ontology—the taxonomic hierarchies in terms of some criteria—performs domain separation. It allows distinguishing between different types of environmental properties.

2.1 Environmental Context Ontology

Environmental context ontology can be employed making context awareness based on the possible context of environments, e.g., sunlight, rain, or snow. Environmental data are categorized as context data and action data. For example, all the relevant visual information in an environment is context data, while only limited data that produce an operation are action data. If once c_1, c_2, \dots, c_n are context data for an environment, then a_1, a_2, \dots, a_m would be action data if $a_i \notin c_i$. Input context data need to be identified (context identification) and used to validate the most effective classifier for given action data. Thus, context data should be modeled in association with input action data as much as possible. Environmental context ontology can be built using context modeling, context adaptation, and context identification.

2.2 Context Modeling

Context modeling is carried out with the combination of domain taxonomy acquisition, domain knowledge base, context data sampling, and manually image fiducial points [4]. In this research, the ontology is composed of FERET synthesized images (described in Sect. 5). Depending on the application domain, this ontology can be extendible. Numerical features represent the basic contents of an image. Twenty-eight fiducial points with 32×32 kernels [4] are associated with feature extraction. Table 1 shows example of high level description of domain class Right direction (Right directed illumination image).

Table 1 Example of high level description of domain class right direction

Domain class	Right direction
Super class	Image
Subparts	{right_top, right_bottom, right_middle, right_zero}
Attributes	{direction, moment, mean}
Direction	{negative, 0, positive}
Moment	{25, 50, 95, 120}
Mean	Average mean

2.3 Context Adaptation

There can be several situations in a dynamic changing environment; we call this context situation. Context adaptation assists in learning different context situations while the system performs the knowledge acquisition phase. The context adaptation task, for example, reveals how to detect a particular context at a certain moment when the system is in action.

Ontology construction can be carried out based on different context situations. If some situations are similar to each other, they will be associated with the same concept in ontology, which we refer to as a cluster. For this we would like to use a clustering algorithm to separate clusters in order to build a tree structure. Then we need to assign concepts for each node in the tree structure to produce ontology. Figure 1b shows the way of context categorization. Though there are many factors that affect environment to change the situation, we prefer sunlight, i.e., illumination. As the context situation varies with illumination, and as ontology is structured on context situation, we call the ontology illumination ontology. Figure 1a shows how to produce illumination ontology, and Algorithm 1 shows how to make contexts from environmental context data.

The system's performance and effectiveness depends on the success of clustering. No clustering algorithm is error free. This is why we have tested our clustering measurement using the popular cluster validity called Dunn's index [6, 7].

2.4 Context-Aware Classifier Selection for Adaptive Object Recognition

Context categorization is necessary to make the system aware of the environmental situation in execution time. Based on this awareness, the system can select a classifier set, which provides maximum accuracy in that situation. Cosine distance [8], K-NN [8], and Euclidean distance [8] are popular distance measurement techniques.

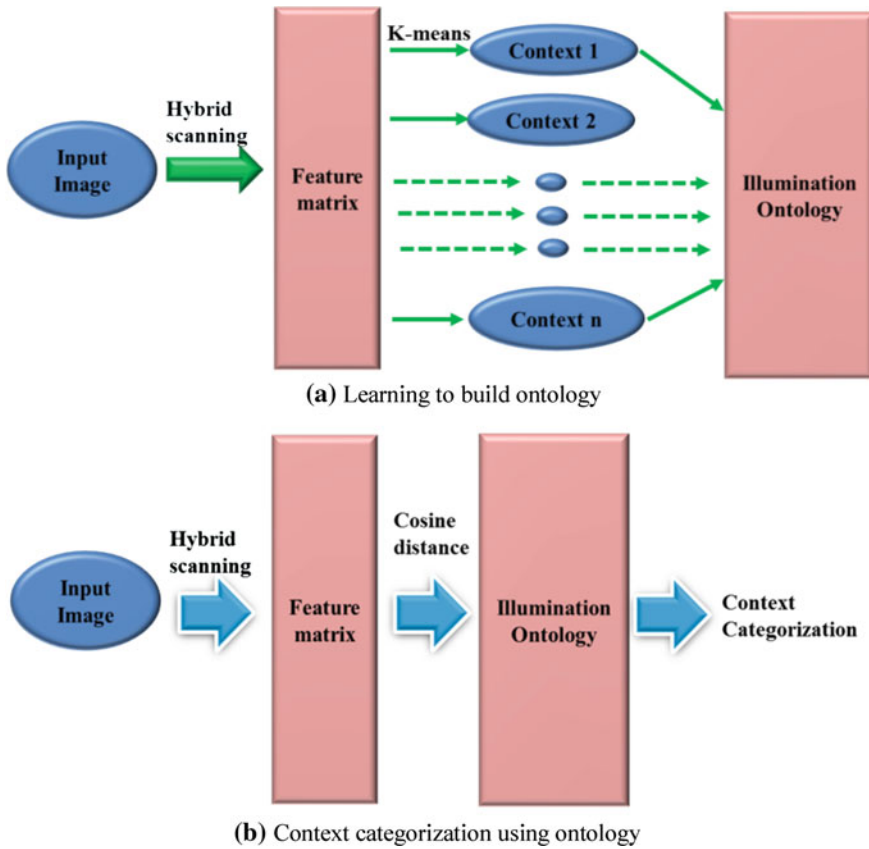


Fig. 1 a Learning to build ontology. b Context categorization using ontology

The system recognizes the current environmental contexts or situations while it performs any operation. It acts according to the minimum distance techniques [9] using Eq. (1).

$$Dis_{\min} = \sqrt{(p_x - q_x)^2 + (p_y - q_y)^2} \quad (1)$$

The classifier selection plays a vital role in the system's accuracy. It is found that some preprocessing tools work properly in different situations but, if a preprocessing tool is wrongly selected, it results in a poor success rate. For example, Retinex [4] and Histogram equalization work well on dark images and normal images, respectively. If this pre-processing is selected in reverse order, it produces a very bad success rate. As context recognition senses the system's experimental environment situation, it is easy to understand which classifier set is appropriate. The framework of classifier combination can be formalized according to Ref. [3].

3 Design of Context-Aware Statistical Inference System

In this research, ontology is built based on the illumination of the environmental context. As ontology representation methods are diverse and depend on the required level of detail and logic, we represent ontology as a hierarchy. For example, in our discussion, illumination ontology is a collection of context situations organized in such order where images are organized according to their illumination, i.e., normal illumination, dark illumination, left-directed illumination, right-directed illumination, and so on.

In this research, the input images are used as the action data as well as the context data. In the evolution mode, the context modeling unit clusters, or models, object data images into several data context categories according to the previous section. Each cluster denotes one data context category, and the cluster sets are organized in hierarchy fashion to produce ontology. The logical structure of illumination ontology is shown in Fig. 2.

The system makes context-awareness from illumination ontology and selects the best classifier combination set, which is applied in the object recognition process. We have organized the experimental environment into three parts: context modeling unit (CMU), classifier selection unit (CSU), and action unit (AU). CMU performs the production of illumination ontology, CSU selects the best classifier combination, and AU performs object recognition. CMU is associated with hybrid scanning [3] for feature extraction on manually sample images, and the K-means algorithm [8] is for learning, i.e., producing clustering and cosine distance measurements for identifying the clusters. CSU selects a classifier system accordingly using the genetic algorithm (GA). It stores its experiences in terms of the data context category and the artificial chromosome in a knowledge base (KB) so that the context knowledge can be accumulated and used later. Each chromosome represents the encoding of the structure of an optimal CSU for a corresponding data context category.

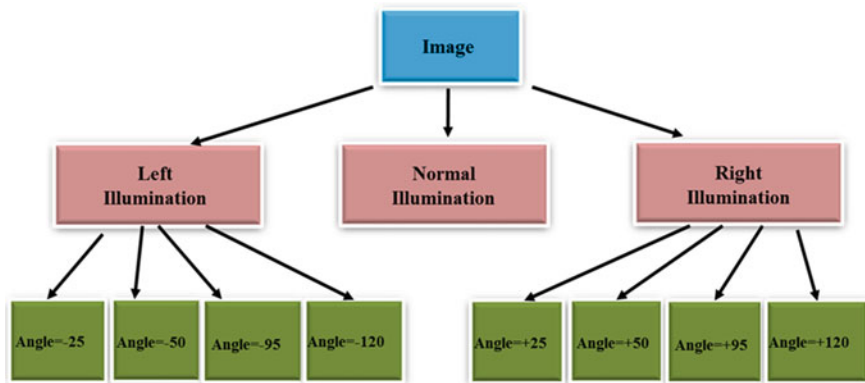


Fig. 2 Logical structure of illumination ontology

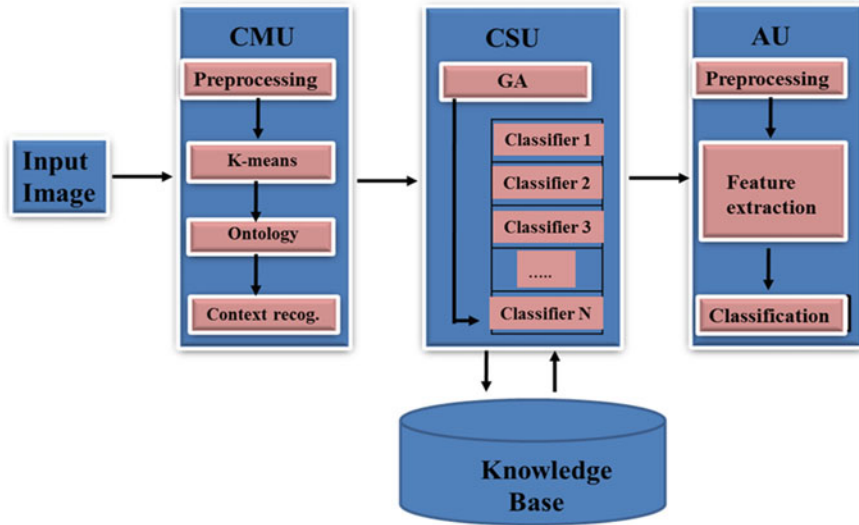


Fig. 3 Ontology construction and applied to object recognition

AU takes care of object recognition, which consists of three stages: pre-processing, feature representation, and class decision. Pre-processing is performed for providing stable quality images as much as possible for object recognition. The action primitives employed for the pre-processing stage here are the histogram equalization, the Retnix [4], and end-in contrast stretching. We adopt Gabor vectors [3, 10], with different weight values of individual fiducial points as the action primitives of feature representation. For simplicity, we adopt the non-parametric classification method K-NN with different threshold values as the action primitives of the class decision stage. The architecture of object recognition is based on ontology as shown in Fig. 3.

We have processed the images with hybrid scanning and fed them into a K-means algorithm for clustering. We have organized the clusters into hierarchy fashion in terms of illumination level resulting from illumination ontology. Then we evolved the genetic algorithm to evaluate the best classifier set for each context model using images in the corresponding cluster, which are then stored into a knowledge base by chromosome. At the time of evaluation, the system recognizes its environmental situation through illumination ontology and then tries to recognize it in association with the knowledge base.

4 Experimental Result

The international standard FERET [11] data set is used for object recognition. However, we have constructed the illumination ontology based on a synthesized FERET dataset. The synthesized FERET dataset is generated by distinguishing the

high, medium, and low brightness levels and left, front, and right coarse lighting direction for each FERET image.

This research shows one method of object recognition in a dynamic changing environment using statistical ontology. The main focus is to develop ontology on environmental situations. As there is no standard database on environmental situation, we have decided to implement on object recognition where environmental situation is artificially created using synthesized images. The system first produces illumination ontology on synthesized images. As we are supposed with 9 possible environmental situations, so our ontology produces 9 nodes. Different nodes work on different classifier sets. According to the chromosome of the knowledge base, we have selected a classifier set for each node and applied this classifier for recognition. For each different node, we have evaluated an enormous image set. Some nodes produce very good results; however, some are in still at the research level.

We have considered normal images as enrollment images. We have learned the system using independent situations and have evaluated it in every situation in random order. Figure 4 shows the recognition rate for each node according to node. The more illumination node on right sub tree produces less success rate than its own, thus disobeying expansion rule of ontology construction. As a result, this node can be further expanded. The average recognition rate is 91.42 %.

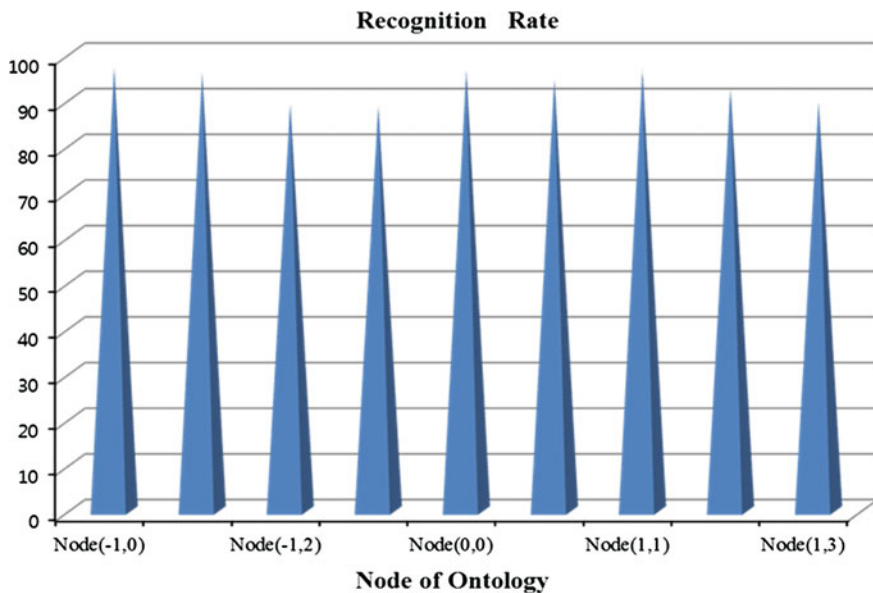


Fig. 4 Recognition rate for each node

5 Conclusions

This paper presents an approach to making statistical ontology based on illumination to assist adaptive object recognition systems in dynamic changing environments. Based on this ontology, an adaptive object recognition technique is carried out with the help of the best classifier selection using genetic algorithm to adapt to different environmental situations. The genetic algorithm-assisted learning builds a knowledge base in terms of artificial chromosomes that represent the classifier components. Consequently, we came up to conclude that the performance of context-aware classifier selection methods for adaptive object recognition is highly depending upon application environments. Although there is a role of context-awareness for robust object recognition, however, there is no way to deny the contribution of adaptability in the case of context-aware classifier selection. Most of the cases, we have achieved higher success by adaptable system with the concept of genetic algorithm. We have applied this illumination ontology approach to object recognition schemes on synthesized FERET data sets and we found that our system is very much competitive with other existing object recognition systems.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (No. 2012-0004478).

References

1. Liu DH, Lam KM, Shen LS (2005) Illumination invariant object recognition. *J Pattern Recognit* 38:1705–1716
2. Celentano A, Gaggi O (2006) Context-aware design of adaptable multimodal documents. *Multimed Tools Appl* 29:7–28
3. Ng CW, Ranganath S (2002) Real-time gesture recognition system and application. *Image Vis Comput* 20(13–14):993–1007
4. Schneiderman H, Kanade T (2004) Object detection using the statistics of parts. *Int J Comput Vis* 56(3):151–177
5. Gomez A, Fernandez M, Corch O (2004) *Ontological engineering*, 2nd edn. Berlin, New York
6. Bezdek JC, Li WQ, Attikiouzel Y, Windham M (1997) A geometric approach to cluster validity for normal mixtures. *Soft Comput* 1:166–179
7. Cootes TF, Taylor CJ, (2004) *Statistical models of appearance for computer vision*. University of Manchester, Manchester (M13 9PT)
8. Duda R, Hart P, Stork D (2001) *Pattern classification*, 2nd edn. Wiley, New York
9. Qing L, Shan S, Gao W, Du B (2005) Object recognition under generic illumination based on harmonic relighting. *Int J Pattern Recognit Artif Intell* 19(4):513–531
10. Wang X, Tang A (2004) Unified framework for subspace object recognition. *IEEE Trans PAMI* 26(9):1222–1228
11. Phillips P (1999) The FERET database and evolution procedure for object recognition algorithms. *Image Vis Comput* 16(5):295–306

Adaptive Skinny Smudge Tool

Noyoon Kwak and Eunyoung Ahn

Abstract This paper is related to an adaptive skinny smudge tool deciding a radius of the master adaptively as well as using an arbitrary master shape. The smudge tool should seem familiar to finger paintings in kindergarten. We can use the smudge tool by selecting its icon on the toolbox of Adobe Photoshop CS6 and dragging in the direction you want to smudge while holding the mouse button down on the image. As the smudge tool blends all the pixels within a fixed radius of the master to generate the result image, its disadvantages are to smudge even the pixels in the undesired region, and to often vary the radius of the master manually. In this paper to reduce the disadvantages, an adaptive skinny smudge tool is proposed. The proposed adaptive skinny smudge tool not only uses the arbitrary master shape adhered closely to the contour shape, but also is able to decide adaptively the radius of the master according to the characteristics of the pixel distribution. Thus, the proposed skinny smudge tool has the advantage of automatically varying the radius of the master and applying the smudge effect to the desired region regardless of the background.

Keywords Smudge tool · Skinny smudge blending · Arbitrary-shaped master · Virtual plastic surgery · Adobe photoshop CS6

N. Kwak (✉)

Division of Information and Communication Engineering, Baekseok University, 115, Anseo-dong, Dongnam-gu, Cheonan, Chungcheongnam-do, South Korea
e-mail: nykwak@bu.ac.kr

E. Ahn

Department of Multimedia Engineering, Hanbat National University, San 16-1, Deokmyeong-dong Yuseong-gu, Daejeon, South Korea
e-mail: aey@hanbat.ac.kr

1 Introduction

A smudge tool is a popular graphic tool included in Adobe Photoshop CS6 [1] and is used to change the contour shape of an object by using smudge effect that makes paints as if the paints are smeared with a finger.

The smudge tool of Adobe Photoshop CS6 is used to smear paints on a picture and makes an effect very similar to finger painting in kindergarten. The smudge tool can be used in such a manner that a user selects the 'smudge tool' in the toolbar of Adobe Photoshop CS6 and clicks and holds down the mouse in the area that you want to have smudged, and then makes strokes in the direction you would like the smudge to be going. The smudge tool is helpful when you want to remove dust and scratches on old photos. Also this tool is widely used for to change the contour shape of an object by using the smudge effect. Owing to this characteristics, even Reallusion iClone5 [2], that is a real-time 3D animation software, employs the smudge tool.

Figure 1 illustrates an exemplary process of smudge-blending a segment of a line by using a conventional smudge tool. Figure 1 shows a result obtained by upwardly smudging a line segment having a thickness of 5 pixels while increasing a smudged distance step by step with a master diameter of 20 pixels and hardness of 100 % in Adobe Photoshop CS6.

In Fig. 2b these are the result of upward, downward, left, and right strokes. The arrows indicate what direct the strokes were made. Figure 3 illustrates a process of smudge-blending a concentric circle with the smudge tool. Figure 3a shows that a master center is designated and then a fixed circular master is formed. Figure 3b shows the result of smudge blending with the fixed circular master. Through Figs. 1 and 3, it can be known that the smudge tool performs smudge blending along the trajectory of moving circular master.

Smudge blending first copies all the pixel values included within the radius of the master located to the start point designated by a user into the master buffer, and pixel by pixel blends the pixel values of the master buffer with the pixel values under the master at the next position of the master, and then replaces the pixel values under the master and the pixel values of the master buffer with the blended pixel values. And the smudge blending as mentioned above along a trajectory of moving master is repeatedly carried out whenever a center of the master shifts from a pixel to another one. Normally the smudge tool is so useful in changing the shape of the visual objects as well as drawing and image touch-up.

However, as the smudge tool blends all the pixels within a fixed radius of the master to generate the result image, one of its disadvantages is to smudge even the pixels in the undesired region in changing the shape of the visual objects. For

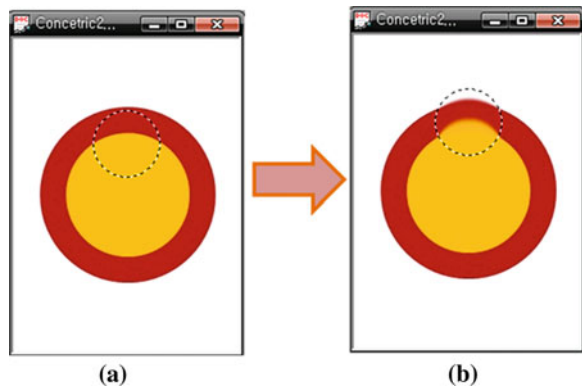
Fig. 1 An exemplary process of smudge blending a line segment with the smudge tool





Fig. 2 An example of smudge blending the water lilies image with the smudge tool. **a** Original water lilies. **b** The water lilies smudged

Fig. 3 An exemplary process of smudge blending a concentric circle with the smudge tool. **a** Master center designated. **b** Smudge blending

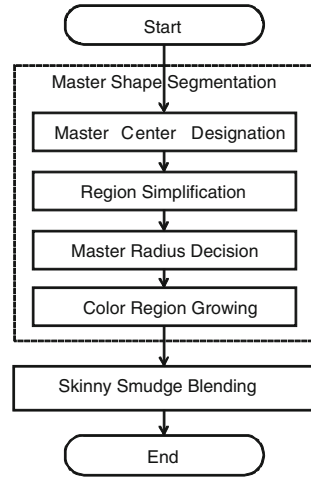


example, when a user does not want to change the shape of the outer circle shown in Fig. 3, this user's intention cannot be easily accepted [3]. And another disadvantage is to often vary the radius of the master manually. In this paper to reduce the disadvantages, an adaptive skinny smudge tool deciding a radius of the master adaptively as well as using an arbitrary master shape is proposed. The proposed skinny smudge tool has the advantage of automatically varying the radius of the master and applying the smudge effect to the desired region regardless of the background.

2 Proposed Adaptive Skinny Smudge Tool

The proposed adaptive skinny smudge tool consists of master shape segmentation step and skinny smudge blending step as follows. Figure 4 is the flowchart showing the proposed adaptive skinny smudge blending.

Fig. 4 Flowchart of the proposed adaptive skinny smudge blending



2.1 Master Shape Segmentation

The master shape segmentation step consists of four steps as follows: master center designation step, region simplification step, master radius decision step, and color region growing step.

- **Master center designation:** In the master center designation step, if the master center is designated at a desired point by using the user input, a circular master candidate region is designated using the predetermined radius of the master.
- **Region simplification:** In the region simplification step, the open-close by reconstruction [4] that is a kind of mathematical morphological operation is performed on R, G and B components of the master candidate region in order to simplify the pixel distribution while maintaining strong edges in the master candidate region.
- **Master radius decision:** In the master radius decision step, the proposed method calculates the standard deviation of each color component of the sample region that includes the master center and ± 4 blocks around the master center, and then decides a radius of the master according to the standard deviation of each color component using the Eq. (1).

$$R_n = R_0 \left(1 + \frac{1}{2} f \left(\sqrt{\sigma_r^2 + \sigma_g^2 + \sigma_b^2} \right) \right), \quad f(t) = \frac{1 - e^{-a(t-b)}}{1 + e^{-a(t-b)}} \quad (1)$$

In the Eq. (1), R_0 and R_n are the newly decided radius and the predetermined radius of the master, respectively. And σ_r , σ_g , and σ_b represent the standard deviations of each R, G, and B pixel value in 9×9 sample region, respectively and $f(t)$ denotes a sigmoid function.

- **Color region growing:** As shown by the Eq. (2), the color bounding box determined in a range in proportion to the standard deviation of each color component is obtained. For all the pixels included within the newly decided radius of the master M_i that is the master candidate region, Color image segmentation is performed using the region growing process that merges neighboring pixels having similar pixel values based on the color bounding box. Segmented regions extracted in this manner are used as the arbitrary-shaped master region m_i [5].

$$m_i = \begin{cases} 0 & \text{if } [|I_r(x,y) - u_r| > \frac{W_r}{2}] \vee [|I_g(x,y) - u_g| > \frac{W_g}{2}] \vee [|I_b(x,y) - u_b| > \frac{W_b}{2}] \\ 1 & \text{otherwise} \end{cases} \quad \text{for } I(x,y) \in M_i,$$

where $W_r = s\sigma_r + o$, $W_g = s\sigma_g + o$, and $W_b = s\sigma_b + o$

(2)

In the Eq. (2), $I_r(x,y)$, $I_g(x,y)$, and $I_b(x,y)$ represent each R, G and B pixel value in the master candidate region M_i , respectively. u_r , u_g , and u_b denote mean values of each R, G and B pixel value, respectively. W_r , W_g , and W_b denote widths of each R, G and B axis of the color bounding box, respectively. In addition, s and o represent a scale factor and an offset factor used to calculate widths of the color bounding box, respectively and ' \vee ' denotes a logical OR operator of C language.

2.2 Skinny Smudge Blending

In the skinny smudge blending step, the proposed method repeatedly performs the skinny smudge blending [3] for all the pixels included within the arbitrary-shaped master along a trajectory of moving arbitrary-shaped master whenever a center of the arbitrary-shaped master shifts from a pixel to another one. The skinny smudge blending will now be explained in more detail with reference to the Eq. (3). In the Eq. (3), $\alpha(x_i, y_i)$ represents opacities of $I(x_i + \Delta x, y_i + \Delta y)$, and (x_c, y_c) denotes the coordinate of the master center, and h represents hardness of the master.

$$ssb_i(x,y) = \alpha(x,y) \times I(x + \Delta x, y + \Delta y) + (1 - \alpha(x,y)) \times m_i(x,y)$$

$$\text{where } \alpha(x,y) = 0.5 \times \left(\frac{\sqrt{(x - x_c)^2 + (y - y_c)^2}}{R_n} \right)^h \quad (3)$$

In the smudge blending step, the proposed method first copies all the pixel values $m_i(x,y)$ included within the radius of the arbitrary-shaped master located to the start point designated by a user into the master buffer, and pixel by pixel blends the pixel values $m_i(x,y)$ of the master buffer with the pixel values $I(x_i + \Delta x, y_i + \Delta y)$ under the master at the next position of the master, and then replaces the pixel values

under the master and the pixel values of the master buffer with the blended pixel values $ssb_i(x, y)$. Subsequently, the pixel values of the master buffer would have been updated into the skinny smudge blended pixel values, and then the skinny smudge blending as mentioned above along a trajectory of moving arbitrary-shaped master is repeatedly carried out whenever a center of the master shifts from a pixel to another one.

3 Simulation Results

In order to evaluate the performance of the adaptive skinny smudge tool, a computer simulation was performed using MS Visual C ++.NET 2008 on the computer of Intel Core i7 CPU Q740 1.73 GHz and 4 GB RAM with Windows 7 operating system.

Figure 5 is the results changed the contour shape of a concentric circle with the proposed skinny smudge tool. As shown Fig. 5, the proposed tool changes the contour shape of the inner circle without changing the contour shape of the outer circle.

Figure 6 illustrates the arbitrary-shaped master obtained using the proposed skinny smudge tool near a chin. The proposed skinny smudge tool is so useful in



Fig. 5 A smudge blending process concentric circles with the proposed skinny smudge tool

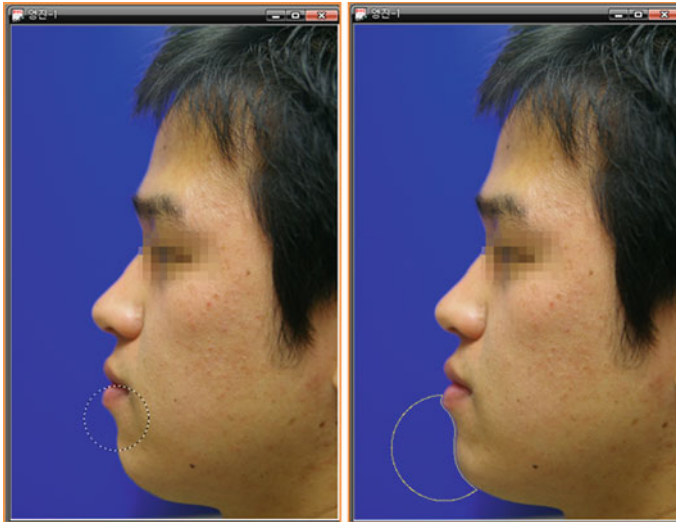


Fig. 6 The arbitrary-shaped master obtained using the proposed skinny smudge tool

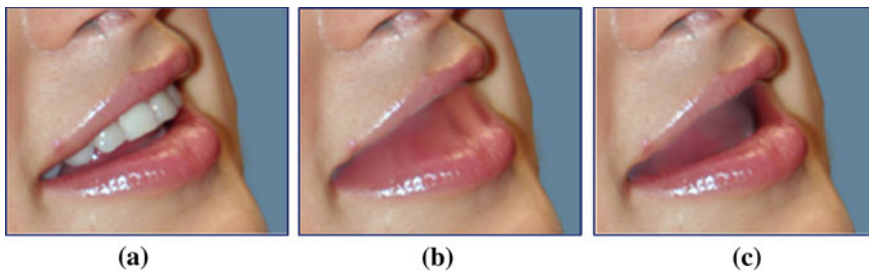


Fig. 7 Tooth removal results obtained using the smudge tool and proposed smudge tool. a Original image. b The smudge tool. c Proposed smudge tool

changing the shape of the visual objects in 2D virtual plastic surgery system. It can be confirmed through Fig. 6 that the arbitrary-shaped master are appropriately extracted such that the master shape regions are in contact with a contour shape formed by pixels existing in the newly decided radius from a master center. The propose method adaptively decides a radius of the master according to the characteristics of the pixel distribution, and then extracts the arbitrary master shape adhered closely to the contour shape within the decided radius of the master using the technique of color image segmentation. And the proposed method repeatedly performs smudge blending all the pixels included within the arbitrary-shaped master along a trajectory of moving arbitrary-shaped master whenever a center of the arbitrary-shaped master shifts from a pixel to another one. Figure 7 illustrates tooth removal results obtained using a previous smudge tool and the skinny

smudge. Referring to Figs. 5, 6, and 7, it can be confirmed that the proposed skinny smudge tool can automatically vary the radius of the master and apply the smudge effect to the desired region regardless of the background.

4 Conclusions

An adaptive skinny smudge tool deciding a radius of the master adaptively as well as using an arbitrary master shape has been proposed in this paper to reduce the disadvantages of the conventional smudge tool. The proposed adaptive skinny smudge tool has the advantage of automatically varying the radius of the master and applying the smudge effect to the desired region regardless of the background. The proposed smudge tool can extract the arbitrary-shaped master in contact with a contour shape to apply the smudge effect only to a desired part irrespective of a background. Therefore, the desired result can be acquired while minimizing a shape change of an undesired part. The proposed smudge tool can be performed in real time, and provide natural and realistic skinny smudge blending results, and also minimize manual operations of a user. The proposed smudge tool is very useful in changing the shape of the visual objects, especially in 2D virtual plastic surgery system.

The technique of color image segmentation not only is a base of the proposed smudge tool, but also decisively affects a realistic smudge blending result. Although a number of image segmentation algorithms have been proposed, the image segmentation technique capable of automatically extracting a desired master shape irrespective of complexity or contrast of a background does not exist due to the characteristics of a 2D image. However, the proposed smudge tool could provide relatively practical and satisfied results by using the color region growing with the standard deviation of each color component.

References

1. Adobe Photoshop CS6 (2012) <https://www.photoshop.com/>. Adobe Systems Inc
2. Reallusion iClone 5 (2012) <http://www.reallusion.com/iclone/>. Reallusion Inc
3. Kwak N-Y, Ahn E-Y (2009) Skinny smudge blending based on master shape segmentation. In: Proceedings of international conference on convergence content 2009, vol 7, no 2, 2009
4. Salemier P, Pardas M (1994) Hierarchical morphological segmentation for image sequence coding. *IEEE Trans Imag Process* 3(5):629–651
5. Gonzalez RC, Wood RE (2002) *Digital image processing* 2nd edn. Prentice Hall, Englewood Cliffs, pp 320–335

Modeling Student's Handwritten Examination Data and Its Application Using a Tablet Computer

Youngjae Kim, Cheolil Lim, Haewook Choi and Minsoo Hahn

Abstract In this work, we propose a model of students' examination data which can contain timestamp and handwriting information using a tablet computer. Stored data is consumed by students' achievement recording software so that both teachers and students can trace their records at any time. The handwriting data in this work can be expressed as XML description so that other applications can use the data easily. The scope of this work is limited to store students' examination information so that it is recommended to be combined with students' profile or other valuable data. Inside the data, this work treats the examination information, question information, students' response time of each question, and moreover, students' behavioral factors including page navigation, removal and writing activities. With the students' behavioral data, a visualization software can give user to suppose the student's activity easily and vividly. In this work, we describes an example of visualization software which allows user to review recorded

Y. Kim

F309, Digital Media Lab, Korea Advanced Institute of Science and Technology,
Moonji-dong, Daejeon 305-732, Republic of Korea
e-mail: yj_kim@kaist.ac.kr

C. Lim

Department of Education, Seoul National University, Seoul 151-742, Republic of Korea
e-mail: chlim@snu.ac.kr

H. Choi · M. Hahn (✉)

Department of Electrical Engineering, Korea Advanced Institute of Science
and Technology, Gusung-dong, Daejeon 305-701, Republic of Korea
e-mail: mshahn@ee.kaist.ac.kr

H. Choi

e-mail: hwchoi2@ee.kaist.ac.kr

secondary students' mathematics examination. Not only score-based assessment, our work may contribute to improve student's assessment and feedback with various educational insights.

Keywords User modeling · Assessment · XML · Education

1 Introduction

The tablet computer is commercialized and widely used in these days, education field also takes advantage to use the tablet. In Republic of Korea, the government announced plans to switch paper-based textbooks from elementary to high school to digitalized textbooks [1]. The advantages of digital texts is continuous updates of contents, ease of distribution, non-linear, resource-rich learning tools, and the inability to lose or damage texts [4]. In detail, it gives exploratory learning and collaborative learning than previous classroom-based study and gives vivid multimedia contents. Common approach to implement a digital textbook is to develop contents based on the tablet computer.

However, many studies have been done with reading activity with a tablet computer, but little have been researched about writing and its assessment. This is because, tablet computer which equipped with digital pen (a.k.a. stylus) is relatively expensive and rare than touch-only tablet computer, and educations in school are mainly performed by reading activity. The writing activity is important because it reflects students' knowledge and, moreover, it reflects more than knowledge itself, for example, the ability of constructive or logical thinking. Previous educational feedback had been performed score-based only, a student hardly trace their past achievements. If we provide a system which stores and displays assessment vividly, the quality of educational feedback may be improved so that the value of education may be enhanced. In this reason, it is important to record and to review students' activity during learning and assessment.

The proposed system provides students to carry out problem solving activities with a digital pen on a paper-like screen and the system records their activities with millisecond precision. The system gathers data on four factors: (1) basic examination information, (2) writing and erasing strokes with the timestamp, (3) browsed question number and timestamp, and (4) subjective measurement of difficulties for each problem. The gathered information can be exported so that it can be consumed by other learning management system. With the proposed data scheme, both students and teachers easy to review and to assess students.

2 Design

2.1 Requirements

In this chapter, we describe our approach to model a student's examination data. There are many types of examination like pencil-and-paper exam, or 1:1 dialogue-based exam, or physical exam. Based on our survey, we organize commonalities of examination to be performed. The common examination data can be summarized as follows.

- Timespan and timestamp during the exam: start time and end time of the data. It is recommend to be recorded in UTC time not a local time. Because we found that local time-based timestamp is required to process additional condition branch with regional information.
- Target students: generally, single student performs the test. But it is required to consider a group of students to be assessed.
- Basic exam metadata. For example, exam code, ID, subjects, etc.
- Recorded data: an activity data during the examination such as handwriting information.
- Score: A numerically expressed the assessed result of the question.
- Comment: an opinion from teachers or colleagues.

Based on the above requirements, we design the flexible and extensible data structure.

2.2 XML Description

The XML specification is important in modern computer data. In modern computer industry, it is strongly required that the data can be easily exported and be consumed with other learning management system [2, 3]. In this reason, XML specification is widely used in data expression. As XML format has drawbacks, we adopt and define XML specification [5]. Figure 1 below describes an example of XML data of recorded examination.

One of the drawbacks in XML format is binary data expression. The root *LoExam* contains information of specification version, student ID, supervisor ID, exam-related data (*ExamCode*, *ExamDateTime*, *ExamTimeSpan*, and *Description*), badge which stands for a tag and comment from teachers or colleagues. The *Q* stands for *Question* which branch contains information about a question. It has question ID, score, rating of the subjective difficulty survey, and likewise, badge and comment. The *Session* stands for a time-based activity which reflects students' behaviors such as page flipping during the examination. The timestamp in this example is expressed with binary time which is based on the UTC time. As we can see, our work also contains Base-64 encoded binary information in *Strokes*

```

<?xml version="1.0" encoding="utf-8" ?>
<LoExam V="20110314" UserID="H8474" Supervisor="kaistudy" ExamCode="2012_HMI_GENERAL" ExamDateTime="
5246529199391345835" ExamTimeSpan="100963.79980000001" Badge="" Comment="" Description="2011 Final Exam">
  <Q ID="1634" Score="0" IsCorrect="" Rating="0" RatingScale="5" Comment="" Badge="" QuestionNumber="1"
  Strokes="AJIbATCIiu2h2uzXR5WZEdI4CovyzsH" CanvasWidth="800" CanvasHeight="1135">
    <Session>
      <Activity Time="5246529198382020796" Category="PageOn" />
      <Activity Time="5246529198429522007" Category="PageOff" />
      <Activity Time="5246529198455616674" Category="PageOn" />
      <Activity Time="5246529199163492572" Category="PageOff" />
      <Activity Time="5246529199179430671" Category="PageOn" />
      <Activity Time="5246529199391345835" Category="PageOff" />
    </Session>
  </Q>
  <Q ID="1635" Score="0" IsCorrect="" Rating="0" RatingScale="5" Comment="" Badge="" QuestionNumber="2"
  Strokes="AAYCA8AHwA" CanvasWidth="800" CanvasHeight="1135">
    <Session>
      <Activity Time="5246529198429678268" Category="PageOn" />
      <Activity Time="5246529198434522707" Category="PageOff" />
      <Activity Time="5246529198452804080" Category="PageOn" />
      <Activity Time="5246529198455616674" Category="PageOff" />
      <Activity Time="5246529199163492572" Category="PageOn" />
      <Activity Time="5246529199165368061" Category="PageOff" />
      <Activity Time="5246529199178337503" Category="PageOn" />
      <Activity Time="5246529199179430671" Category="PageOff" />
    </Session>
  </Q>
</LoExam>

```

Fig. 1 Block diagram of the proposed system

attribute at Q branches. In above example, the *Strokes* attribute contains student’s handwriting data. The *Strokes* attribute is an optional value and any other binary data can be stored with its appropriate attribute. In addition, we designed that many values are placed as attributes (not to add branches of deeper levels) in each branch. This is because, we maintain to preserve simplicity of XML description and to store customizable data. In next chapter, we will describe how the data can be acquired and to be visualized with given data.

3 Recording and Visualization of Examination

We implement a software which utilize the proposed XML description. Figure 2 is screenshot of the implemented software. The software is implemented by C# and runs based on the Windows 7 operating system.

As we can see, the left side mimics traditional pencil-and-paper examination. While students solve problems, the system gathers data on three factors: (1) writing and erasing strokes (a group of dots) with a timestamp, (2) the question number browsed with a timestamp, and (3) a five-point subjective difficulty measurement survey for each problem. Except for the top and bottom of the screen, almost all areas accept writing input. Visualizer has a list of given problems on the left panel, the selected problem with a review button in the center panel, and multi-functional panels on the right. With the two software, user can perform test and review student’ activity.

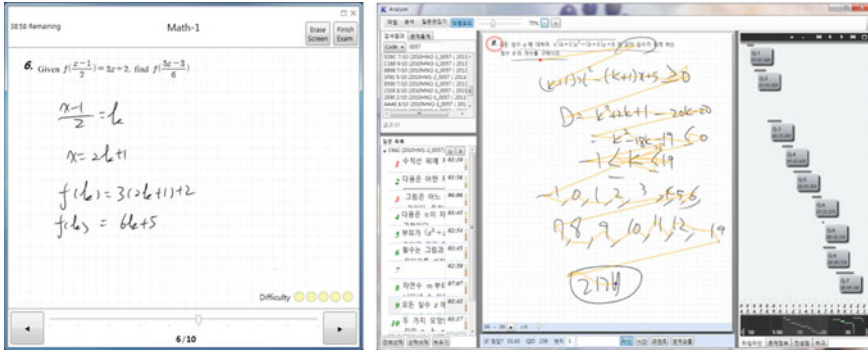


Fig. 2 Screenshot of the implemented system (left digital exam sheet, right visualizer)

4 Conclusion

The proposed system aims to describe students' examination data in order to review and visualize students' examination activity. The proposed scheme provides extensibility and flexibility in terms of data expression so that it can be consumed by other educational data. The digital exam sheet software is implemented that can record students' handwriting information with the proposed scheme. And the visualizer software is for displaying students' activity. The implemented system can be used with the tablet computer so that handwriting information is stored during the examination. We hope our system helps to improve students' assessment and feedback.

References

1. Kim JHY, Jung HY (2010) South Korean digital textbook project. *Comput Sch* 27(3-4):247-265
2. Kim M, Yoo KH, Park C, Yoo JS (2010) Development of a digital textbook standard format based on XML. *Adv Comput Sci Inf Technol* 363-377
3. Kim P, Podlaseck M, Pingali G (2004) Personal chronicling tools for enhancing information archival and collaboration in enterprises. In: *Proceedings of the 1st ACM workshop on continuous archival and retrieval of personal experiences*, ACM, New York pp 56-65
4. Lamb B, Bleecker J (2007) *Textbook Reloaded: rise of the digital text*
5. Roy J, Ramanujan A (2001) XML schema language: taking XML to the next level. *IT Prof* 3(2) 37-40

Advanced Media Measuring Method Using MPEG-2 Transport Stream for High Quality Broadcasting Management System

Sangkeun Kim

Abstract IPTV is a system where a digital television service is delivered by using Internet protocol over a network infrastructure, which may include delivery by a broadband connection. This paper proposes advanced FR-based measuring methods between original and processed media while transport streams are delivery from headend system. The proposed measuring scheme uses the brightness and edge of digitized contents blocks per each frame of MPEG-2 transport stream to evaluate contents in realtime. The proposed algorithm is effective video measuring as evidenced in the experimental matching results. The method in this paper performed in a high precision degree.

Keywords Broadcasting management system · Measuring · MPEG-2 transport stream

1 Introduction

QoS performance monitoring information collected from the network often does not accurately correlate with the perceived QoE by users owing to error correction and recovery mechanisms at the application layer. Also, aspects such as application control plane latency (e.g., channel change delay) and actual picture distortion may go unnoticed in the network. Therefore, to measure QoE requires the collection of performance monitoring data from various sources to determine actual

S. Kim (✉)
Department of Broadcasting Media, Korea Nazarene University,
Chungnam 331-718,
South Korea
e-mail: kimsk@kornu.ac.kr

service state and quality, as well as root causes of quality degradation. IP-media system can be classified by three main factors: headend system, transport network, customer network. We can define headend system which is including until process after muxing of TS source. IP-core network and access network can be defined as transport networks. Finally, customer network include full processing after access network which is included set-top box, HDTV, etc. In order to evaluate QoS/QoE over IP network environment, various measurements solutions is deployed. However, early deployments sometimes use expensive dedicated test probes. And then, these methods are unsuitable for the real-time measurement or monitoring in order to guarantee high quality of IPTV service. FR [2] methods use the entire set of pixels while RR methods extract key information from the encoder and decoder pixels. Both FR and RR require reliable transport of information to the video quality monitor from the transmitting and receiving ends. The rate of this transfer is very high for the FR methods, although they may have the greatest accuracy because they use the most information. No-Reference (NR) methods do not require the original video, because they predict video quality using only measurements at either before decoder or after decoder processing [6]. Using the picture activity measure, the slope relating the PSNR and the logarithm of the quantizer-scale is obtained. Then, using the logarithm of the quantizer-scale and the obtained slope, the PSNR incurred by compression only is estimated. They extend this work by estimating a separate statistical distribution for each DCT coefficient, using quantized data and assuming the distribution is constant across the frame [8]. However, those are often not scalable and may ignore aspects such as user perceived quality of experience based on human vision model although managing Quality of Experience (QoE) is one of the most challenging aspects for IPTV service requirements [9]. In this paper, to measure the quality of video provided by IP-media service, we propose realtime measurement scheme using original video source and determine QoE indicators over IP networks, which are robust designed for distortion problem especially comparing reference with processed video sequences base on FR method. Hence, the proposed techniques make that both methods can apply to measure in real-time between TS and original video, possibly. To measure the degradation of video quality, we set the three monitoring points in the IPTV service structure, which are headend, transport, and customer networks. Developed methods in this work are applied for measuring the quality of experience on video source. From the experimental results, with our proposed method for measuring QoE indicator, the degradation of video quality can be measured efficiently and robustly. From the experimental results, with our proposed method for matching of distortion TS frames, the degradation of video quality can be measured efficiently and robustly in real-time. The rest of this paper is organized as follows. In Sect. 2, we describe real-time distortion frame realtime measurement scheme. In Sect. 3, the experiments for performance evaluation of the proposed model are shown the proposed IP-media quality measurement system. Finally, in Sect. 4, we draw the conclusions including further studies (Fig. 1).

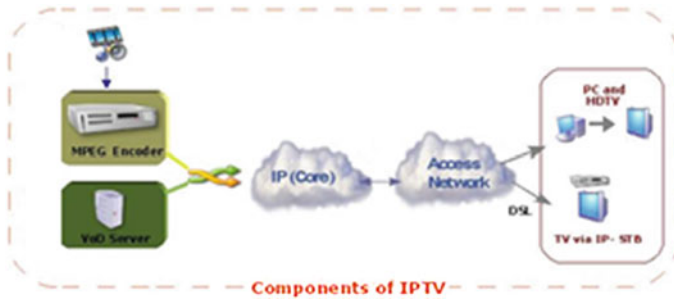


Fig. 1 Components of broadcasting media transport networks

2 Headend Broadcasting System and Measuring Scheme

For the measurement for broadcasting media service with considering both network and video levels, we should consider with both QoS factors and QoE factors together [3, 10]. Network performance can be monitored directly to ensure video service delivery. Some necessary performance metrics for IPTV surveillance can be collected from the IP/MPLS network elements and element management systems, including network traffic loss/delay/jitter statistics. Furthermore, we should measurement not just end-to-end, but at any possible points on networks such as headend, transport network, customer networks to provide QoS/QoE guaranteed IP media service and to find problems as fast as possible.

The Fig. 2 shows broadcasting media system with the possible checking points from three different areas for measuring. On the headend area, we can assess at the points which are before/after encoder processing and before muxing. At the measuring points, block distortion and blurring effects mainly happen on the TS source. From the transport area, we can assess at the points which are before/after IP network and before/after access network. TS source is affected by packet loss, delay, jitter, etc. and then color error, jerkiness, edge busyness, etc. Set-top box in customer network area which is after final access network measures all of QoE indicators which can be mainly affected by channel zapping time. Basically, we use network-QoS aware video-QoE indicators estimation based FR video measurement and monitoring method in real-time. The current issue in the area is to measure in real-time with face value which service providers really want the greatest accuracy. The content-based video retrieval systems normally use color histogram-based matching methods [1]. However, they are not suitable for distorted TS of IPTV service and have problem with color distribution between original and processed TS especially since the color histogram does not preserve information about the spatial distribution of colors. Also, in the cases of dynamic motion and various shot changing for short time, there are usually occurred serious distortion problem on TS.

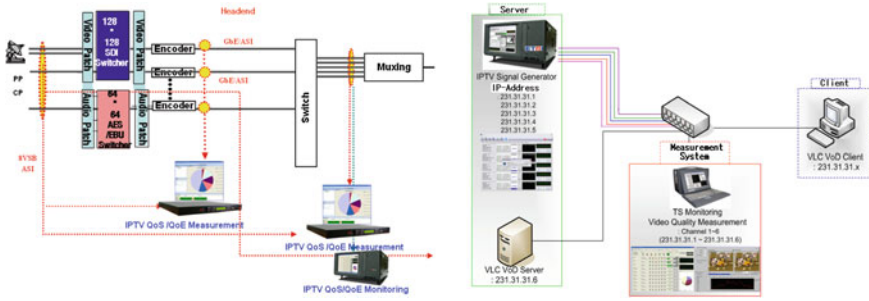


Fig. 2 Components of broadcasting media System

For applying to various distortion frames of TS without above problem, we use partition based [7] ordinary measuring approach method. The proposed scheme uses color classification in necessary for training of real-time matching using digitized TS source, first of all. Second, the proposed scheme is also robustness to matching and measuring in accuracy with even serious distortion frames on TS source by several reasons such as delay, jitter, packet loss, encoding/decoding processes, etc. The color classification based training of current TS is investigated in RGB/HSV [4, 5]. We analyze colors which are specified in terms of hue (H), saturation (S), and intensity value (IV) and show the relationship with RGB in order to match and measure in real-time considering various distortion errors. We assign H bins for hue channel, S for saturation channel, and IV for intensity value channel. We set 164, 164, and 256 to H, S, and IV, respectively. We define accumulated histograms HUE, SATURATION, INTENSITY_VALUE, in which the values in each bin is accumulated for the first 35 TS frames per second. For the computation of partition based ordinary measuring, since the considering IPTV resolution [3], each TS video frame is partitioned into 16*16 images. That is, each frame is partitioned as the same 164 blocks. After partition the TS source, the mean brightness of color is expressed as

$$M_i = \frac{1}{HW} \sum_{x,y \in B_i} I(x,y), \quad i = 1, 2, \dots, 164 \quad (1)$$

where B is block, H*W are height * width of blocks from the partitioned TS sources. I(x, y) is brightness degree values, respectively. From Eq. (1), M which is mean brightness value, is saved as 16 * 16 rank maps to measure distance between original source frame and TS source frame.

$$d_{n,m} = \frac{1}{164} |TM_i^n - OM_i^{n+m}|, \quad i = 1, \dots, 164 \quad m = 0, 1, \dots, L - P \quad (2)$$

where TM is processed TS source, OM is original source. L is length of TS source, and P is length of original source. After the processing of each frame, distance value D is measured from Eq. (3), respectively.

$$D = \frac{1}{P} \sum_{n=1}^P d_{n,m}, \quad m = 0, 1, \dots, L - P \tag{3}$$

With dynamic motion shots such as dancing, sports, etc., there are usually occurred serious distortion problem on current TS. For the case of serious distortion areas happen, to detect more closely, the video sources are partitioned again into $(16 * 16) * n$ times to matching in accuracy and check distance until the value is less than threshold values which is less than threshold 5 in this case. Finally, we match the distortion frames of current TS in real-time to support QoE indicators errors. After the processing procedures, all of the QoE indicators are measured by the proposed FR-based measurement methods.

3 FR Model Based Quality Measurement Algorithm

The framework for evaluating performance has been implemented by using Visual Studio 2010 under FFmpeg library, which has been utilized for MPEG decoding. Various videos are used for our experiments. To analyze the encoding effect, original video is encoded with different QP value by using MPEG-2.TS. For real IPTV service environment, we also use the sample video, which is degraded by noise and packet loss artificially on the headend system. We use two videos, which are shown in Table 1, to evaluate the performance.

Our advanced VQM is designed by QoE indicators such as edge and block. Finally, we get MOS result which is graded from 1 to 5 levels. It is also correlated between VQM and subjective quality measurement. The Fig. 3 shows procedures of media quality assessment.

Table 1 Evaluating performance

Test videos	Characteristic	Total frames
TS 1(Music video)	Network effect	50 frames
TS 2 (CF)	Encoding effect	100 frames

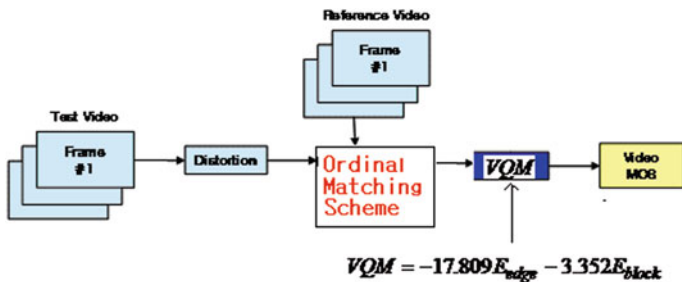


Fig. 3 FR model based quality measurement scheme

4 Conclusions

In this paper, we have proposed advanced media measuring method of MPEG-2 Transport Stream over headend broadcasting system. Also, the FR-based video measurement method has developed in order to guarantee accuracy and reliability with realtime measurement. The proposed algorithm is effective video measuring as evidenced in the experimental matching results. The method in this paper performed in a high precision degree.

Acknowledgments This work has been supported by the research program of Korea Nazarene University, South Korea, 2012. Also, I am very thanks to Prof. Jinsul Kim for his great technical advice.

References

1. ATIS (2007) A framework for QoS metrics and measurements supporting IPTV services. ATIS-0800004
2. Hemami S, Masry M (2002) Perceived quality metrics for low bit rate compressed video. In: Proceeding of ICIP, pp 721–724
3. Jinsul K, Tae-Won U, Won R, Byung Sun L, Minsoo H (2008) Heterogeneous networks & terminals-aware QoS/QoE-guaranteed mobile IPTV service. *IEEE Commun Mag* 46(5):740
4. Kim Byung-Gyu et al (2006) A fast intra skip detection algorithm for H.264/AVC video encoding. *ETRI J* 28(6):721–731
5. Lee MS et al (2007) Techniques for flexible image/video resolution conversion with heterogeneous terminals. *IEEE Commun Mag* 45(1):61–67
6. Reibman AR, Vaishampayan VA, Sermadevi Y (2009) Quality monitoring of video over a packet network. *IEEE Trans Multimed* 6(2):327–334
7. Swain M, Ballard D (1991) Color indexing. *Int J Comput Vis* 7(1):11–32
8. Turaga et al D (2002) No reference PSNR estimation for compressed pictures. In: Proceeding of ICIP, pp III.61–III.64
9. Girod B () What's wrong with mean-squared error. In *Digital Images and Human Vision*, A. B. Watson ed, MIT Press pp 207–220
10. Yuchul J, Yoo-mi P, Hyun JB, Byung SL, Jinsul K (2011) Employing collective intelligence for user driven service creation. *IEEE Commun Mag* 49(1):76–83

Proposed Media Signal Sharing Scheme Through NGN for Service Overlay Multimedia Framework

Jungdae Kim

Abstract The media sharing technique in the signal for multimedia communication is one of the necessary research issues to provide the quality-guaranteed services over IP-based NGN (Next Generation Network). This paper proposes noisy signal reduction method with considering packet loss for multi-user connected multimedia communication. In order to reduce noisy signal without packet distortion efficiently for applying to All-IP based convergence environment, we provide a shared signal reduction and recovery scheme which use an optimized Wiener filtering technique based on input-SNR estimation with adaptive both-side information utilization according to the average magnitude variation across the frames. Also, this paper presents the policy-based service overlay IPTV framework for user experiences in the Open IPTV. We suggest the EPG ought to be selectable in the Open IPTV framework for requirements. The framework is proposed with the basis of the ITU-T IPTV NGN framework with IMS (IP Multimedia Subsystem).

Keywords Media signal sharing · Packet-based IP networks · Framework · Overlay

1 Introduction

To ensure further growth of the IPTV (Internet Protocol TeleVision) market, it is one of the big problems to solve that the IPTV service heavily depends on the service providers. They gather subscribers, aggregate contents, provide the IPTV

J. Kim (✉)

Korea Nazarene University, Chungnam 331-718, South Korea
e-mail: jdkim@kornu.ac.kr

platform, and make the charge for the services. As the interests in supporting IP communication tremendously increase, a variety of IP-based real time multimedia application services have been developed rapidly in residential and enterprise communication markets because of their attractive service enhancements and cost savings. For the future All-IP based convergence networks and systems in providing high quality, useful quality-guaranteed techniques with considering multimedia and multi-user are researched to support pervasive communication services. As the general quality factor in ubiquitous environment, environmental noise impediment is inevitable in case of that hardware- or software-based communication system located in anywhere and we make phone-call in anytime. Research on noise reduction/speech enhancement can be traced back to about 40 years ago [1–4]. The noise reduction techniques are used to a wide range of applications such as communication, automatic speech recognition, and sound source localization systems today [5–8].

We think that the Open IPTV ought to be beneficial to service, contents, and terminal providers, and subscribers most of all. The important requirements of our Open IPTV design are as follows. Subscribers are able to enjoy every service providers' contents freely and cheaply with competitions. Service providers maintain their priority to profit from their customers. Thus, we make the specific service overlay environment for the Open IPTV called the Neighbored Garden.

And then, we are able to perceive that in the Open IPTV environment the providers' main interest is the management of EPG (Electronic Program Guide). Because EPG influences the subscribers to watch the intended VODs, and the actions is directly connected with the profit of the firm.

In the Open IPTV, the service selection problem is very important, because there are numerous overlapped contents and providers. In the service selection problem, the policy [9] is very useful to make the priority of contents [10]. The existing paper about the policy in the walled garden IPTV mainly proposed QoS (Quality of Service) and charge management [10]. And the service provider needs to record the location of subscriber in the service space to manage the authentication, and the billing with the service profile [11] in the Open IPTV.

The monitoring component is essential to our framework to search for the list of the same services and the server condition variables. And the monitoring component can detect contexts [12, 13], so this component is also able to the context awareness and the service managements services for adaptation [14, 15].

The proposed structure is based on the ITU-T IPTV NGN Framework with non-IMS, and similar with the service overlay network [16, 17]. And the NGSON (Next Generation Service Overlay Network) shows the overlay structure.

The EPGs (Electronic Program Guide) must be selectable by subscribers for the IPTV openness [18]. The various and specialized EPG services of several providers such as the EPG for children, adult, etc. will promote user experiences. It's similar with the internet portal service or the IPTV personalized EPG [19, 20].

2 Media Signal Sharing Scheme with Noise Reduction

The Noise reduction is one of the significant research factors because of degrading the speech through IP network. A clean speech signal $d(n)$ which is a zero-mean clean speech signal, an additive noise $v(n)$ which is a zero-mean noise process, and recent output of an observation signal as the noisy speech signal at the discrete time sample $y(n)$ is given by

$$y(n) = d(n) + v(n) \quad (1)$$

We propose an optimized Wiener filter method using estimated SNR (Signal-to-Noise Rate) ratio for speech enhancement. The signal power spectrum is computed for the windowed signal using the 256-FFT. Based on the VAD decision, the noise power spectrum is updated only for non-speech intervals in the Wiener filter design. For speech intervals, the last noise power spectrum is reused. And the speech power spectrum is estimated by the difference between the noise power spectrum and speech power spectrum. The designed Wiener filter coefficients in the frequency domain are transformed into the time-domain ones by the IDCT (Inverse Discrete Cosine Transform). The noise is suppressed by the convolution sum between the impulse response of the proposed an optimized Wiener filter which is estimated by SNR measuring and the noisy speech. In the proposed optimized Wiener filter, the frequency response is given by

$$W(k) = \frac{\zeta^\alpha(k)}{1 + \zeta^\alpha(k)}, 0 < \alpha \leq 1 \quad (2)$$

and $\zeta(k)$ is defined as

$$\zeta(k) = \frac{P_d(k)}{P_v(k)} \quad (3)$$

where k is the frequency bin, $\zeta(k)$, $P_d(k)$, and $P_v(k)$ are the SNR, the speech power spectrum, and the noise power spectrum, respectively.

The filtering can be controlled by the parameter α . As α increases, $\zeta^\alpha(k)$ also increases for $\zeta(k)$ greater than one, while $\zeta^\alpha(k)$ decreases for $\zeta(k)$ less than one. Therefore, the signal is more strongly filtered out to reduce the noise for small $\zeta^\alpha(k)$. On the other hand, the signal is more weakly filtered with little attenuation for large $\zeta^\alpha(k)$. To analyze the effect of α , we evaluate the performances for α values from 0.1 to 1.0. We can adaptively select the optimal α according to the estimated SNR by a logistic function. The logistic function is trained to decide the optimal α for the estimated SNR. The logistic function used in this paper can be expressed as

$$p(SNR) = Min + \frac{2(Max - Min)}{1 + e^{(|n-1|/\beta)}} \quad (4)$$

Because the shape of the logistic function changes with the variation of β , if the maximum and the minimum values of the logistic function are fixed, we should find the appropriate β .

The appropriate β value is decided by the simple gradient search algorithm. At the first iteration, for the initial β value, the corresponding α as the output of the logistic function is calculated with the estimated SNR as the input of logistic function at each frequency bin for each frame. The average spectral distortion J for all frames is measured with the log spectral Euclidean distance defined as

$$J = \frac{1}{N} \sum_{i=0}^{N-1} \frac{1}{L} \sqrt{\sum_{k=0}^{L-1} [\log|X_{ref_i}(k)| - \log\{|X_{in_i}(k)|W_i(k)\}]^2} \quad (5)$$

where i is the frame index, N is the total number of the frames, k is the frequency bin index, L is the total number of the frequency bins, $|X_{ref_i}(k)|$ is the spectrum of the clean reference signal, and $|X_{in_i}(k)|W_i(k)$ is the noise-reduced signal spectrum after filtering with the designed Wiener filter. At the second iteration, β is updated by the simple gradient search procedure. The average spectral distortion is measured with the new logistic function defined by the updated β . Until the termination condition is satisfied, the iteration is repeated. Finally, β is decided after the final iteration.

3 Scenario for Service Overlay Multimedia Framework

In the design of the Service Overlay Multimedia framework in the service overlay control we mainly point two things. One is that subscribers can select any EPG in the network, and the other is that the final decision of service selection depends on the subscriber's policy.

The framework design with the ITU-T IPTV NGN framework with non-IMS, and it is composed of the Application Function, the Service Control Function, the End-User Function, the Content Delivery Function, and the Data Transport Function.

We mainly re-designed the Service Control Function with 5 functional blocks: the Service Overlay Control Block (SOC) mainly connects other service providers to provide the Open IPTV services, and the User Profile Management Block (UPM) manages the user profiles, they are not shared with other service providers.

The Service Monitoring Block (SM) works for the gathering information in the service overlay network, the Service Policy Decision Block (SPD) is the brain to make decision with personal policy, and the Service Routing/Negotiation Block (SRN) control the transmission on the overlay network. Figure 1 and Table 1 shows the configuration and the capability of the framework blocks.

The SOC, the UPM, and the SRN work on the service overlay network and the SOC is the Gateway to connect with the other service providers so the UPM and the SRN work through the SOC but we omit the SOC in the sequence diagram of the SM, or the SRN. Figure 2 shows the overlay network for the Open IPTV network.

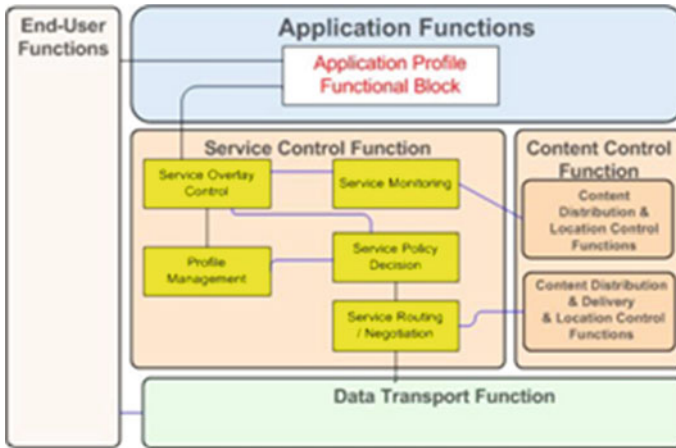
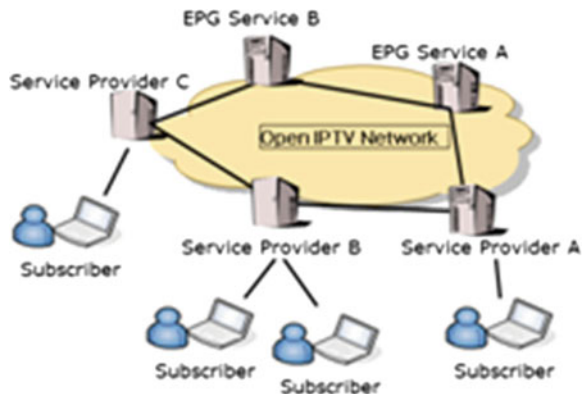


Fig. 1 Framework for open IPTV

Table 1 Capability of the Blocks

Block name	Functions
Service overlay Control (SOC)	Gateway of message and control the process on the service overlay network
User Profile Management (UPM)	Keeping user profiles and sending user authentication information, etc.
Service Monitoring (SM)	Searching service list and monitoring network, context, and user location on the service overlay network
Service Policy Decision (SPD)	Deciding priority with personal policy
Service Routing/Negotiation (SRN)	Requesting NGN network resource and service transmission on the service overlay network

Fig. 2 Service overlay multimedia network



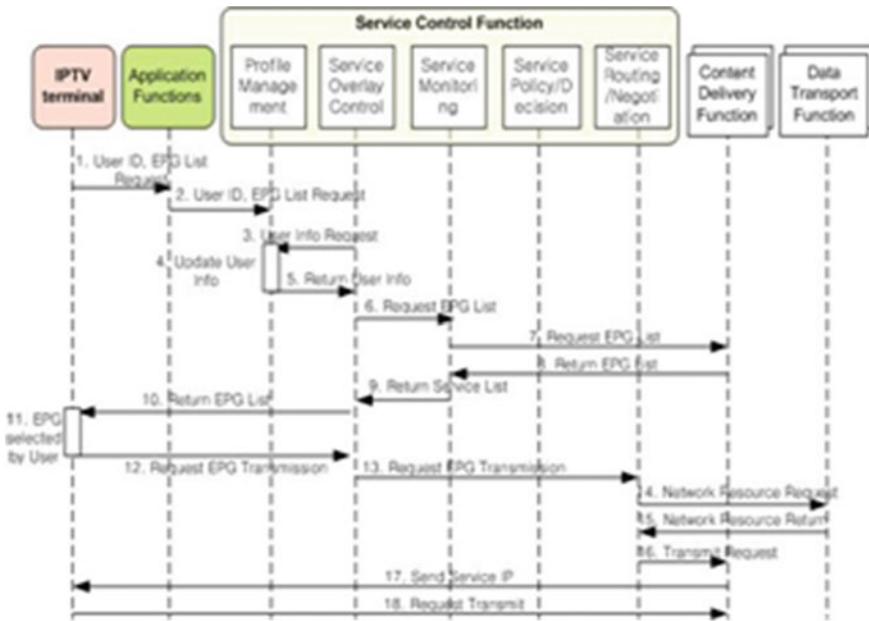


Fig. 3 Sequence of EPG selection

The framework has advantages to works on the service overlay network and be adaptive to the personal policy compared with ITU-U NGN framework with non-IMS.

4 Search EPGs and Transmit EPGs

The Open IPTV subscriber can select the EPG which is provided in the network, and receive the EPGs list. When the subscriber request EPG list, the SM collect the EPG list on the service overlay network. And after a EPG is selected by subscriber, the SRN request the transmission of EPG. Figure 3 shows the sequence diagram of the EPG selection scenario.

5 Evaluation

We implemented our test system of the Open IPTV based on the service scenarios. Our test system used the ASP.NET framework, and Java for the XML (eXtensible Mark-up Language) Web Service (Tables 2, 3).

Table 2 Main functions for EPG selection

Seq. Num.	Function name	Input	Output
1	<i>ServiceRequest</i>	UserID, ServiceInfo, Policy	None
6	<i>ServerListRequest</i>	ServiceInfo	Server List, Variables
10	<i>DecisionRequest</i>	Server IP List, Server state, Policy	None
12	<i>TransmitApproval Request</i>	Provider ID	None

Table 3 Main functions for VoD selection

Seq. Num.	Function name	Input	Output
1	<i>EpgListRequest</i>	UserID, IP	None
3	<i>UserDataRequest</i>	UserID, IP	User Info, SLA,
6	<i>Request EpgList</i>	UserID, IP	None
12	<i>RequestEpgTransmit</i>	EPG Server IP	None
14	<i>ResourceRequest</i>	Server IP, User IP, Bandwidth	None
16	<i>TransmitRequest</i>	Server IP, User IP	None

6 Conclusion

In this paper, the performance evaluation of speech quality confirms that our proposed shared scheme outperforms more efficiently than the original algorithm in the speech codecs. The performance results in this paper have established and confirmed that our proposed shared packet loss-aware robust noise reduction scheme using estimated input-SNR outperforms more efficiently than the existing method, which control noisy signal considering speech distortion problem with packet loss. To sum up, our proposed method allows and shows the strength with the improved performance results for the noisy signal reduction at the same post-processing time while the noisy signal is removed over packet-based IP networks. Also, we proposed the service overlay media framework. And we show that subscribers are able to select the provider freely on the service overlay network. So we believe that the neighbored garden Open IPTV would be a new chance to not only the customers but also the providers.

Acknowledgments This work has been supported by the research program of Korea Nazarene University, South Korea, 2012. Also, I am very thanks to Prof. Jinsul Kim for his great technical advice.

References

1. McAulay RJ, Malpass ML (1980) Speech enhancement using a soft decision noise suppression filter. *IEEE Trans Acoust Speech Signal Process* 28(2):137–145
2. Boll SF (1979) Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans Acoust, Speech Signal Process* 27(2):113–120

3. Ephraim Y, Malah D (1984) Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. *IEEE trans Acoust, Speech Signal Process* 32(6):1109–1121
4. Ephraim Y, Malah D (1985) Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE trans Acoustics Speech Signal Process* 33(2):443–445
5. Park M, Kim HR, Yang SH (2006) Frequency-temporal filtering for a robust audio fingerprinting scheme in real-noise environments. *ETRI Journal* 28(4):509–551
6. Lee HC, Halverson DR (2007) Design of robust detector with noise variance estimation censoring input signals over AWGN. *ETRI J* 29(1):110–112
7. Choi BH et al (2006) An all-optical gain-controlled amplifier for bidirectional transmission. *ETRI J* 28(1):1–8
8. Bang KH et al (2006) Audio transcoding for audio streams from a T-DTV broadcasting station to a T-DMB receiver. *ETRI J* 28(5):664–666
9. Sreenivas D (2007) Policy control for IPTV domain, IP multimedia subsystem architecture and applications. In: 2007 international conference on, pp 1–5
10. Lymberopoulos L, Lupu E, Sloman M (2003) An adaptive policy-based framework for network services management. *J Netw Syst Manag* 11:277–303
11. Sheridan-Smith NB, Soliman J, Colquitt D, Leaney JR, O’Neill T, Hunter M (2004) Improving the user experience through adaptive and dynamic service management. In: Australian telecommunications networks and applications conference, Citeseer
12. Dey AK, Abowd GD (2000) Towards a better understanding of context and context-awareness. In: CHI 2000 workshop on the what, who, where, when, and how of context-awareness, pp 304–307
13. Schilit B, Adams N, Want R (1994) Context-aware computing applications. In: Proceedings of the workshop on mobile computing systems and applications, Citeseer, p 8590
14. Muntean CH, Muntean G (2006) Framework for interactive personalised IPTV for entertainment. W3C WWW conference, Edinburgh, Scotland
15. Keeney J, Cahill V (2003) Chisel: a policy-driven, context-aware, dynamic adaptation framework. *POLICY* pp 4–6
16. Hwang J, Park MS (2003) Adaptive service profile replication scheme for next generation personal communication networks. *IEICE transactions on communications e series B*, vol 86, pp 3344–3351
17. Duan Z, Zhang ZL, Hou YT (2003) Service overlay networks: SLAs, QoS, and bandwidth provisioning. *IEEE/ACM Trans Netw (TON)* 11:883
18. Obrist M, Kepplinger S, Beck E, Tscheligi M, Muzak P (2008) Usability and user experience: preliminary results from evaluating an IPTV community platform. *Lect Notes Comput Sci* 5066:303–307
19. Lyu J, Pyo S, Lim J, Kim M, Lim S, Kim S (2007) Design of open APIs for personalized IPTV service. In: Advanced Communication Technology, the 9th International Conference on, 2007
20. Mas I, Berggren V, Jana R, Murray J, Rice C (2008) IMS-TV: an IMS-based architecture for interactive, personalized IPTV. *IEEE Commun Mag* 46:156–163

Realtime Sport Analysis Methodology for Extracting Target Scenes on Mobile Environment

Chung Young Lee and Jung Mo Kim

Abstract In this paper, we propose realtime sport analysis method to extract target scenes in mobile environment considering heterogeneous devices. The proposed method extracts a specific context and displays the context of the target scene from the whole scenes with considering optimal viewer visual sight on devices. We use color-based classification with satisfying viewer visual perception in terms of the number of fast frames within the sport video streaming.

Keywords Sport analysis · Mobile devices · Video streaming · Scene extracting

1 Introduction

In this paper, we propose realtime extract target scenes in mobile environment which is absolutely useful technique for sport video analysis as today's rapidly increasing of sport market. Technically, mobile devices use all of area in the world today. That is, the people would like to broadcast and display multimedia streaming content [1] and transmit the contents over IP networks in realtime to share with others [2].

Currently, there are several problems without the proposed scheme in necessary techniques. As providing sport video contents on many different devices such as cellular phone, PDA, computer, etc., which has various screen sizes from small to

C. Y. Lee · J. M. Kim (✉)
Department of Takwondo, Korea Nazarene University,
Cheonan, ChungNam 331-718, South Korea
e-mail: kjmtkd@kornu.ac.kr

C. Y. Lee
e-mail: cylee@kornu.ac.kr

big, the contents just broadcasts by first captured scene without considering device's capability with display size (Fig. 1). Viewer feels that it is very uncomfortable and cannot recognize scene [5]. Current sports broadcasting system like soccer, football, etc., does not consider the function which extracts viewer's targeted context with visual sight none the less. Color classification based processing is necessary for training of realtime sport video stream. To solve those problems, we present useful methodology which is color and contexts classification-based advanced scenes extracting scheme for sport analysis system with considering terminal capability.

2 Target Scenes Extracting Method

The proposed scheme extract target scene for sport analysis system. We believe that viewers have more interesting area in each scene. To provide viewer centric contexts on a scene, we extract viewer targeted scene using dynamic motion vector and static image information of each object on source level in the proposed method. In order to reflect and consider viewer, the proposed scheme provides with four different blocks: video source size controls with considering viewer's visual sight and terminal capability, color classification based training of current video stream which is investigated in RGB/HSV, target context analysis and display, and specific scenes-based background processing from camera's original shot (Fig. 1).

First of all, we consider that a sport video streaming is defined as a sequence of shots, in which a shot is an image sequence captured by a single and by several cameras.



Fig. 1 Display sport video scene (e.g., football game): compare same scene with heterogeneous devices having different LCD screen sizes

The *SBS1* (*Sport Broadcasting Shot by a Single*) $S1_n$ is number of frames in a video which is given as

$$SBS1 = \{S1_0, S1_1, \dots, S1_{n-1}\} \tag{1}$$

Also, the n th shot can be denoted by a sequence of frames which can be written as

$$S1_n = \{f1_n^0, f1_n^1, \dots, f1_n^{I-1}\} \tag{2}$$

where $f1_n^i$ is the i th frame of $S1_n$ and I is the number of frames in the shot by a single.

Also, *SBS2* is sport broadcasting shot by double camera and $S2_n$, the n th shot can be denoted by a sequence of frames which can be written as

$$S2_n = \{f2_n^0, f2_n^1, \dots, f2_n^{I-1}\} \tag{3}$$

where $f2_n^i$ is the i th frame of $S2_n$ and I is the number of frames in the shot by a second camera. We can define several cameras as *PBSN* and SN_n .

Several cameras shot same context at the same time in various angles. For instance, lots of spectators shot the singer at their location in concert hall or the player at their location in a stadium. In this case, in order that user wants to gathering the target item with removing background from several shots by several cameras based on several angles, we provide the following steps: stitching images, drawing boundary lines, removing the background, feature extraction and classifying objects.

For stitching shots, we used Lukas–Kanade (LK) tracker to compute displacements between successive shots. LK tracker first selects feature points using a corner detector, and, when a new frame arrives, LK tracker computes the new positions of the feature points. The displacement of a new frame with respect to an old frame is determined by the average of the displacement vectors of the feature points. A new frame is combined with the previous shot image simply by translating the other shot image by the calculated displacement and “overwriting” it on the previous shot. We could use a better registration technique to combine shots, but we decided to use the simplest one because classification algorithm in the latter step utilizes only the color distribution of a target context. For drawing boundary lines, the outline of a target object is indicated by a circling gesture of mouse and then can be constructed by connecting the centers of shot image frames. After the processing, for removing background a boundary lines drawn in the previous step forms a closed curve. Given a closed curve surrounding a target, it is straightforward to eliminate the background. The resulting image is cropped and resized to a 640×480 image in our experiment. For extraction and classification, as the image features of a target context, we decided to use the number of gray pixels (pixels whose saturation value is smaller than 0.1) and the 32-bin hue histogram of the remaining pixels. These features are invariant under translation and rotation, not too sensitive to the change of an ambient illumination, and, above all, fast to

compute. The target image context was obtained by substituting black pixels for gray pixels in the original image and retaining only the hue component of other pixels in the original image. Counting the number of gray pixels followed by a 32-bin hue histogram computation gives 33 numbers, and these numbers are then normalized (divided by their sum) in order to ensure scale invariance. The first 32 columns in each plot correspond to a hue histogram and the last column represents the number of gray pixels.

We present a comprehensive study of important issues of the color pixel classification approach to possible long-shot segmentation in football stadium or in concert hall, etc. for sport analysis broadcasting service in this paper. We investigate how the choice of color space and the use of chrominance channels affect football ground long-shot segmentation while personal service provider broadcasts the football game in football stadium. We should consider that there exist numerous color spaces however many of them share similar characteristics. Also, since specific color of an object share with other objects in many cases, it is not easy to analyze and need learning the specific color. Colors are specified in terms of the three primary colors: red (R), green (G), blue (B), basically. In this paper, we analyze colors which are specified in terms of hue (H), saturation (S), and intensity value (V) and show the relationship with RGB in our target shot in order to reflect the degree of human perception. We consider distortion of all or a portion of the final characterized by the appearance of unnatural or unexpected hues or saturation levels which were not present in the original image on realtime sport broadcasting service. The HSV histogram obtained from a long-shot frame in a football video. We see hue histogram is dominated by yellow-green bins, while histograms for saturation and intensity value are pretty spread out. We assign H bins for hue channel, S for saturation channel, and IV for intensity value channel. We set 64, 64, and 256 to H , S , and V , respectively. Therefore, each histogram for the i th frame is defined as in (4).

$$\begin{aligned} 0 \leq k \leq H \text{ for } Hue_i[k] \\ 0 \leq k \leq S \text{ for } Saturation_i[k] \\ 0 \leq k \leq IV \text{ for } IntensityValue_i[k] \end{aligned} \quad (4)$$

In addition, we define accumulated histograms HUE , $SATURATION$, $INTENSITY_VALUE$, in which the values in each bin is accumulated for the first 30 frames of the video.

$$\begin{aligned} HUE[k] &= \sum_{i=0}^{29} Hue_i[k] \\ SATURATION[k] &= \sum_{i=0}^{29} Saturation_i[k] \\ INTENSITY_VALUE[k] &= \sum_{i=0}^{29} IntensityValue_i[k] \end{aligned} \quad (5)$$

by utilizing the definitions in (5), two variables are defined as follows. *ValuePeakIndex* = i , where

$$INTENSITY_VALUE[i] = INTENSITY_VALUE[p] \text{ for all } 0 \leq p \leq IV, \text{ and}$$

$$SaturationMean = \frac{\sum_{i=0}^{NS-1} i.SATURATION[i]}{\sum_{i=0}^{NS-1} SATURATION[i]} \tag{6}$$

ValuePeakIndex denotes the index of the bin, which is the peak point of the *IntensityValue* histogram. We regard that there exist enough information for ground learning if the summation of $HUE[i]$, $9 \leq i \leq 21$, takes more than 70 % of whole pixels, where the 9th to 21st bins correspond to yellow-green area, determined through enough observations. If the condition is not satisfied, the same process is performed for the next 30 frames are satisfied. Basically, it is noted that there is a relationship of $g > r > b$ on the ground color in sport stadium, which is obtained by observing many football or soccer videos of sport analysis broadcasting system. We want to set more refined conditions to minimize the false including of a ball or gray tone pixels into the ground detection results. The equations to compute saturation S and intensity value IV in HSV color space from R, G, B values in RGB space.

$$S = \frac{Max(R, G, B) - Min(R, G, B)}{Max(R, G, B)} \text{ for } (0.0 \leq S \leq 1.0) \tag{7}$$

$$IV = Max(R, G, B) \text{ for } (0.0 \leq IV \leq 1.0)$$

where R, G, B are all normalized values from 0 to 1. By using (7) and the observed relationship $g > r > b$ for the ground area, we can denote IV as $g \cdot IV$ since $G \approx Max(R, G, B)$ and $V = 256$, whereas $S = \frac{g-b}{g} \cdot S$ since $G \approx Max(R, G, B)$ and $B \approx Min(R, G, B)$. Finally, the refined condition, described in rgb space, to distinguish each pixel whether or not it belong to the ground $G(x,y)$ are expressed in (8).

$$G(x, y) = \left\{ \begin{array}{l} 1, \text{ if } \left\{ \begin{array}{l} g > 0.95.r, \quad \text{and} \\ r > 0.95.b, \quad \text{and} \\ g < ValuePeakIndex + \theta_1, \quad \text{and} \\ \frac{g-b}{g} \cdot S > SaturationMean - \theta_2 \end{array} \right\} \\ 0, \text{ otherwise} \end{array} \right\} \tag{8}$$

where r, g, b denotes the RGB values at $G(x,y)$ respectively, ranging from 0 to 255. We set the value θ_1 and θ_2 to be $V/6$ and $S/4$, respectively. If a pixel's *IntensityValue* is too higher than *ValuePeakIndex*, the pixel is not assumed to be in the ground. If a pixel's Saturation is too lower than *SaturationMean*, the pixel is not

assumed to be in the ground, either. Using this method is faster than those proposed from [3], and robust to the case where a part of the football ground is shadowed [4].

Although we have a list of objects and discrimination rules, it is not easy to determine which object is the target context [5]. For instance, if there are existed many similar objects in a scene, it is really hard to find the target object. Furthermore, usually objects from video stream move unsystematically and locate at different position. Initial object position and their trajectory information have been used to obtain for future frames by some schemes. However, it is not quietly suitable for realtime application. For extracting of the target context, we suggest a simple and causal method which can be used for realtime processing. First, we assume that the longest tracked remarkable object-candidate has the highest probability to be a target context. This strategy can minimize the influence of sudden noise. We maintain candidates list and keep adding the newly found candidates to our decision tree. Each candidate has its own age. In the next frame, candidates are succeeded by the closest object in terms of both spatial distance and attributes. If a successor has an attribute of an object, it is kept in the list and its age increases. On the contrary, if a successor does not have a similar attribute of a target context, its age in the list is decreased. We choose the oldest candidate as the most probable candidate. An object whose age is less than zero is removed from the list. With this scheme, we can detect the specific target context with high accuracy, finally. Figure 2 shows target scenes for sport analysis broadcasting service after applying the proposed method.

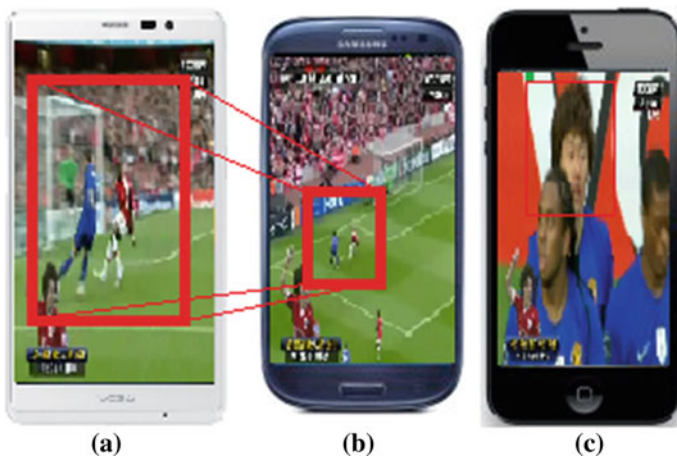


Fig. 2 Display target scenes for sport analysis broadcasting service (e.g., Soccer): **a** display by considering terminal capability and viewer visual sight, **b** display without considering terminal capability and viewer visual sight, **c** display by extracting target contexts

3 Conclusion

In this paper, we propose realtime sport analysis method to extract target scenes in mobile environment considering heterogeneous devices. The proposed method extracts a specific context and displays the context of the target scene from the whole scenes with considering optimal viewer visual sight on devices. We use color-based classification with satisfying viewer visual perception in terms of the number of fast frames within the sport video streaming. The color-based classification to provide realtime analysis and extraction targeted scenes with considering end user satisfied service. Also, we expected that the method is very useful for service provider broadcasts system.

Acknowledgments This work has been supported by the research program of Korea Nazarene University, South Korea, 2012. Also, I am grateful to Ph.D Jins. Kim, for his professional supports.

References

1. Knoche H, McCarthy JD, Sasse MA (2005) Can small be beautiful: assessing image resolution requirements for mobile tv, In: Proceedings of the 13th annual ACM international conference on multimedia (MULTIMEDIA'05). ACM Press, New York, pp 829–838
2. Cheng W.-H, Chu W.-T, Wu J.-L (2010) A visual attention based region-of-interest determination framework for video sequences. IEICE Trans Inform Syst E-88D:1578–1586
3. Wan K, Yan X, Yu X, Xu C (2008) Real-time goal-mouth detection in mpeg soccer video. In: Proceedings of the eleventh ACM international conference on multimedia (MULTIMEDIA'03). ACM Press, New York, pp 311–314
4. Yu X, Xu C, Leong HW, Tian Q, Tang Q, Wan K (2011) Trajectory-based ball detection and tracking with applications to semantic analysis of broadcast soccer video. In: Proceedings of the eleventh ACM international conference on multimedia (MULTIMEDIA '03). ACM Press, New York, pp 11–20
5. Kim J, Um T-W, Ryu W, Lee BS, Hahn M (2008) Heterogeneous networks and terminals-aware QoS/QoE-guaranteed mobile IPTV service. IEEE Commun Mag 46(5):110–117

A Cost Effective Method for Matching the 3D Motion Trajectories

Hai-Trieu Pham, Jung-ja Kim and Yonggwon Won

Abstract 3D trajectory data have progressively become common since more devices which are possible to acquire motion data were produced. These technology advancements promote studies of motion analysis based on the 3D trajectory data. Even though similarity measurement of trajectories is one of the most important tasks in 3D motion analysis, existing methods are still limited. Recent researches focus on the full length 3D trajectory data set. However, it is not true that every point on the trajectory plays the same role and has the same meaning. In this situation, we developed a new cost effective method that uses the feature ‘signature’ which is a flexible descriptor computed only from the region of ‘elbow points’. Therefore, our proposed method runs faster than other methods which use the full length trajectory information. The similarity of trajectories is measured based on the signature using an alignment method such as dynamic time warping (DTW), continuous dynamic time warping (CDTW) or longest common subsequence (LCSS) method. In the experimental studies, we compared our method with two other methods using Australian sign word dataset to demonstrate the effectiveness of our algorithm.

Keywords Motion analysis · 3D trajectory · Similarity of trajectory · Sign words

H.-T. Pham

Chonnam National University, 77 Yongbong-ro, Buk-Gu, Gwangju 500-757, Korea
e-mail: haitrieupham@gmail.com

J. Kim

Chonbuk National University, 567 Baekje-daero, Deokjin-gu, Jeonju-si 561-756, Korea
e-mail: jungjakim@jbnu.ac.kr

Y. Won (✉)

Chonnam National University, 77 Yongbong-ro, Buk-Gu, Gwangju 500-757, Korea
e-mail: ykwon@chonnam.ac.kr

1 Introduction

Over recent years, thanks to the development of sensor technology and mobile computing, trajectory-based object motion analysis has gained significant interest from researchers. It is now possible to accurately collect location data of moving objects with less expensive devices. Thus, applications for sign language and gesture recognition, global position system (GPS), car navigation system (CNS), animal mobility experiments, sports video trajectory analysis and automatic video surveillance have been implemented with new devices and algorithms. The major interest of trajectory-based object motion analysis is the motion trajectory recognition. The motion trajectory recognition is generally achieved by a matching algorithm that compares new input trajectory with pre-determined motion trajectories in a database.

The first generation of matching algorithms only used raw data to calculate the distance between two trajectories, which is ineffective. Raw data of similar motions will appear differently because of various varying factors such as scale and rotation. To overcome this problem, local features of trajectory, called *signature*, were defined for motion recognition [1–3]. This signature performs better in flexibility than other shape descriptors, such as B-spline, NURBS, wavelet transformation, and Fourier descriptor. Trajectories represented by the signature and the descriptors are invariant in spatial transformation. However, computing the distances between trajectories using this signature is not enough for accurate recognition of 3D motion. To improve the performance, some matching approaches were used to ignore similar local shapes of different motion trajectories or to ignore outliers and noise.

‘*Matching*’ is an important process in motion recognition and classification, which have been studied for years and widely used in many fields. It is achieved by alignment algorithm, and the famous and efficient ones in motion recognition are dynamic time warping (DTW), continuous dynamic time warping (CDTW), and longest common sub-sequence (LCSS) [4–6].

Recent researches use the full length of trajectory data for motion recognition [1–3]. However, many points of the trajectory have similar signatures because they lie on a straight line, thus computing task for signatures can be useless. To eliminate this drawback, we developed a new method that computes the signatures only from the region of ‘*elbow points*’ to gain advantage of computing speed. Besides, we also present a set of descriptors and normalization process for invariant motion recognition.

2 Preprocessing Method

Due to the system noise, measurement noise or both, trajectory data may not be accurate. ‘*Smoothing*’ process is an important task because it enhances the signature’s computational stability by reducing the noise and vibration of motion.

However, trajectory shape may be affected by the smoothing process. To cope with the effect of noise, the derivatives of a smooth version of data using a smoothing kernel ϕ are considered, i.e. $x^{(j)}(t) = (x(t) * \phi(t))^{(j)}$. By the derivative theorem of convolution, we can have $x^{(j)}(t) = x(t) * \phi^{(j)}(t)$. For this paper, a B-spline $B(t)$ is taken to be the smoothing kernel $\phi(t)$. An odd degree central B-spline of degree $2h - 1$ with the integer knots $-h, -h + 1, \dots, 0, h - 1, h$ is given by

$$B(s) = \frac{1}{(2h - 1)!} \sum_{l=-h}^{h-1} (-1)^{l+h} \binom{2h}{l+h} (s - l)_+^{2h-1}$$

where the notation $f_+(s)$ mean $f(s)$ if $f(s) \geq 0$ and 0 otherwise. For a quantic B-spline, $h = 3$ [7].

Next, we normalize the location and the scale of a 3D trajectory so that all trajectories are transformed to a common domain. Trajectory normalization makes scale, rotation and translation invariant, which can produce better performance for the following processes. We applied the continuous principal component analysis (PCA) [8] to the trajectory points, where we assume that three distinct nonzero eigenvectors can be computed from the 3D trajectory coordinates. The continuous PCA ensures the invariance of the translation, the rotation, the reflection, and the scale.

3 Signature as a Trajectory Descriptor

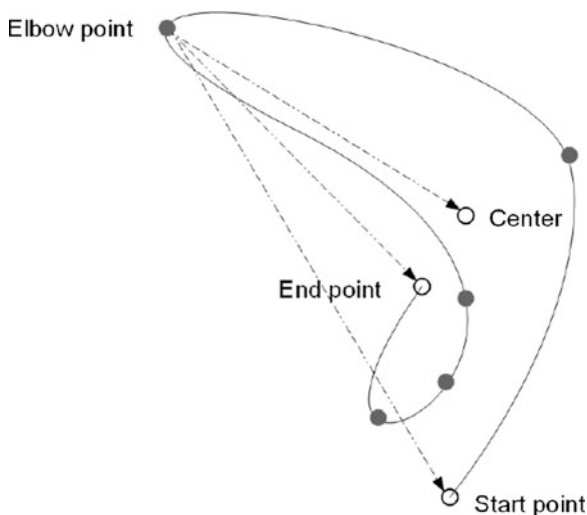
For trajectory matching, we need a descriptor that can well describe the shape of the trajectory. In our study, we use the *signature* for the descriptor. For a point t , the signature $S(t)$ is defined by five values: $\kappa(t)$, $\tau(t)$, $h(t)$, $e(t)$ and $c(t)$. $\kappa(t)$ is the ‘curvature’ that is a measurement for the turning amount of the contour, and $\tau(t)$ is the ‘torsion’ that presents its twist amount out of the tangent-normal plane. Other three values $h(t)$, $e(t)$ and $c(t)$ are the Euclidian distances from the point t to the start-point, the end-point, and the center-point of the trajectory, respectively. Note that the center-point is computed by the continuous PCA which is performed at the normalization process. Thus, for a motion trajectory in 3D space with N points $\Gamma = \{x(t), y(t), z(t) | t \in [1, N]\}$, the signature set D^* for the entire trajectory is defined in the following form

$$D^* = \{\kappa(t), \tau(t), h(t), e(t), c(t) | t \in [1, N]\}$$

where

$$\begin{aligned} \kappa(t) &= \|\dot{\Gamma}(t) \times \ddot{\Gamma}(t)\| / \|\dot{\Gamma}(t)\|^3 \\ \tau(t) &= (\dot{\Gamma}(t) \times \ddot{\Gamma}(t)) \cdot \ddot{\ddot{\Gamma}}(t) / \|\dot{\Gamma}(t) \times \ddot{\Gamma}(t)\|^2 \end{aligned}$$

Fig. 1 Illustration of elbow points (block dots) and 3 Euclidian distances



$$h(t) = \|\Gamma(t) - \Gamma(1)\|$$

$$e(t) = \|\Gamma(t) - \Gamma(N)\|$$

$$c(t) = \|\Gamma(t)\|$$

An *elbow point* is a point on the trajectory which have the curvature value $\kappa(t)$ larger than a threshold ϕ . If we know the coordinates of the elbow points and their sequential order, we can rebuild an approximated trajectory by connecting the elbow points with the straight lines of points. Consequently, information about the elbow points is good enough to align two trajectories for matching task. Therefore, a new literal set of signature only with the elbow points can be described as

$$D' = \{\kappa(t), \tau(t), h(t), e(t), c(t) | t \in [1, N], \tau(t) > \phi\}$$

This new set of the signature only with the elbow points has four or five times less number of elements (signatures, points) than D^* . As a result, computational burden for matching two trajectories can be dramatically reduced by using D' rather than D^* . An illustration for the elbow points (block dots) and the three distances are shown in the Fig. 1.

4 Signature Alignment

As mentioned in previous section, each trajectory is represented by a set of signature. Note that in our proposed method, we only compute the signatures at the elbow points. For each elbow point, as mentioned above, five signature elements are obtained: $\kappa(t)$, $\tau(t)$, $h(t)$, $e(t)$ and $c(t)$. In order to match two trajectories, two

corresponding signatures should be correctly paired. Since there are many noisy factors such as different number of signatures in two trajectories, a matching approach should consider methods to handle the noisy factors. There are many approaches to match two set of sequence data such as LCSS and DTW [4]. The LCSS is more adaptive and appropriate distance measurement for trajectory data than DTW [1]. We therefore choose LCSS for matching process in our study.

Given an integer δ and a real number $0 < \varepsilon < 1$, we define the $LCSS_{\delta,\varepsilon}(A, B)$ as follows:

$$\begin{cases} 0 \text{ if } A \text{ or } B \text{ is empty} \\ 1 + LCSS_{\delta,\varepsilon}(Head(A), Head(B)), \\ \quad \text{if } |a_{x,n} - b_{x,m}| < \varepsilon \\ \quad \text{and } |a_{y,n} - b_{y,m}| < \varepsilon \\ \quad \text{and } |a_{z,n} - b_{y,m}| < \varepsilon \\ \quad \text{and } |n - m| \leq \delta \\ \max(LCSS_{\delta,\varepsilon}(Head(A), B), LCSS_{\delta,\varepsilon}(A, Head(B))), \text{ otherwise} \end{cases}$$

The constant δ controls how far in time we can go in order to match a given point from one trajectory to a point in the other trajectory. The constant ε is the matching threshold. The similarity function S between two trajectories A and B , given δ and ε , is defined as follows:

$$S(\delta, \varepsilon, A, B) = \frac{LCSS_{\delta,\varepsilon}(A, B)}{\min(n, m)}$$

This LCSS model allows stretching and displacement in time, so we can detect similarities in movements that happen at different speeds, or at different times.

5 Experimental Results

In order to implement and evaluate the proposed method for matching the 3D motion trajectories, we have used trajectories information of the Australian Sign Language (ASL) data set obtained from University of California at Irvine’s Knowledge Discovery in Databases archive [9]. The ASL trajectory dataset consists of 95 sign classes (words), and 27 samples were captured for each sign word. The coordinates x, y and z are extracted from the sign’s feature sets to calculate the trajectory signature. The length of the samples is not fixed. The details for the experimental setup are exactly the same as that described in [10], where the data set consists of sign words ‘Norway’, ‘alive’, and ‘crazy’. Each sign-word category has 69 trajectories.

Haft trajectories from each category were used for training, and the remains were used for testing. A test sample is classified by the nearest neighbor rule

Table 1 Sign-word recognition results using object motion-based trajectory

Method	Correction rate (%)
Signature of elbow points	84.76
PCA-based GMM	85.29
Global GMM	69.61
Pose normalization	52.38

($k = 5$). The experiment was repeated 40 times (each time with a randomly selected training and test datasets). The average result of recognition was 84.76 %. We also performed the experiment with pose normalization method [1]. Our proposed method was compared with other methods included PCA-based Gaussian mixture models (GMM) and global Gaussian mixture models [10], and the comparison result is reported in the Table 1.

Note that our proposed method used only a subset of trajectory data while other methods used the full length trajectory data. Even though the recognition result of our proposed method does not outperform the PCA-based GMM method, the number of data points for recognition process is much smaller, which implies less computational complexity. Therefore, our proposed method is more advantageous than PCA-based GMM in term of recognition speed.

6 Conclusion

In this paper, we proposed a new method for matching the 3D motion trajectories, and demonstrated experiments to show its effectiveness. It used only the features, named in *signature*, obtained from ‘*elbow points*’ which are the points that have the curvature value larger than a specific threshold.

In the first step, all trajectories are smoothed and then normalized by continuous PCA. By using continuous PCA, all trajectories are invariant in translation, rotation and scale. Once all the trajectories are normalized, a set of signature which contained both local features and global features of trajectory is computed from only the elbow points. LCSS matching algorithm was used to match the signatures from the elbow points in two trajectories. Comparison of one trajectory and another trajectory in a database, actually one set of signatures and another set of signatures in a database, is quite complicated if the database size is big and the length of the trajectory is long. Therefore, using only subset of full trajectory points is simple and fast in trajectory matching process.

Even though our method uses less information of the trajectory for matching, sign word recognition results showed that our proposed method can still maintains the recognition rate compared to the existing methods. This implies that the features from the elbow points are good enough to include the information for matching two trajectories. However, further works should include investment for

the sensitivity of the threshold value to recognition results, which affects the number of elbow points. Also, practical application study should be performed with large number of sign-words.

Acknowledgments This study was financially supported by Chonnam National University, 2011

References

1. Croitoru A, Agouris P, Stefanidis A (2005) 3D trajectory matching by pose normalization. In: Proceedings of the 13th annual ACM international workshop on geographic information systems, pp 153–162
2. Wu S, Li YF (2009) Flexible signature descriptions for adaptive motion trajectory representation, perception and recognition. *Pattern Recognit* 42:194–214
3. Yang JY, Li YF (2010) A new descriptor for 3D trajectory recognition. *Automation and logistics (ICAL)*, pp 37–42
4. Vlachos M, Hadjieleftheriou M, Gunopulos D, Keogh E (2003) Indexing multi-dimensional time-series with support for multiple distance measures. In: Proceedings of the ninth ACM SIGKDD international conference on knowledge discovery and data mining, pp 216–225
5. Vlachos M, Kollios G, Gunopulos D (2008) Discovering similar multidimensional trajectories. In: Proceedings 18th international conference on data engineering, pp 673–684
6. Aach J, Church G (2001) Aligning gene expression time series with time warping algorithms. *Bioinformatics* 17:495–508
7. Kehtarnavaz N, deFigueiredo JP (1988) A 3D contour segmentation scheme based on curvature and torsion. *IEEE Trans Pattern Anal Mach Intell* 10(5):707–713
8. Vranic D, Saupe D (2001) 3D shape descriptor based on 3D fourier transform. In: Proceedings of the EURASIP conference on digital signal processing for multimedia communications and services (ECMCS 2001), Budapest, Hungary, pp 271–274
9. Australian Sign Language Dataset (1999) <http://kdd.ics.uci.edu/databases/auslan/auslan.html>
10. Bashir F, Khokhar A, Schonfeld D (2005) Automatic object trajectory-based motion recognition using Gaussian mixture models. In: IEEE international conference on multimedia and expo, pp 1532–1535

Perceived Quality Model for Supporting Full Session Mobility in Multimedia Service Delivery Process

Dongjun Suh, Jinsul Kim and Seongju Chang

Abstract Context-awareness has empowered advanced multimedia computing technology to provide user-oriented services and environment. This research aims at achieving a seamless video delivery service concentrating on the user's mobility patterns during a multimedia streaming service session. Mobility supporting technology providing seamless services are categorized into host mobility and user mobility. The former corresponds to host-level handoff while the latter refers to user-level handoff. In host-level handoff, the influential parameters affecting the quality of video consumption are the total distance between the hosts, the distance for streaming to resume while user is in mobility mode and the screen size of the end host. The interdependency amongst the parameters is evaluated by a user subjective assessment and a befitting video quality model is developed, accordingly. Additionally, the optimal video delivery switching point to enable user mobility based video service is studied based on a quality analysis of host mobility services at varying distances among hosts. This quality model supporting a complete mobilization has a high correlation with the assessed quality and enables an adequate seamless mobility for multimedia service delivery.

D. Suh

Smart Environment Design Laboratory, KAIST, Gwahangno, Yuseong-gi,
Daejeon 305-701, South Korea
e-mail: djsuh@kaist.ac.kr

J. Kim

School of Electronics and Computer Engineering, Chonnam National University,
Gwangju 500-757, South Korea
e-mail: jsworld@jnu.ac.kr

S. Chang (✉)

Ubiquitous Space Research Center of KIUSS, KAIST, Gwahangno, Yuseong-gi,
Daejeon 305-701, South Korea
e-mail: schang@kaist.ac.kr

Keywords Session mobility · Seamless mobility · Multimedia handoff

1 Introduction

Multimedia mobility is largely classified into either host mobility [1] or user mobility [2, 3] in accordance with host-level and user-level handoff, respectively. Most of the precedent studies have separately considered each type of mobility.

A subjective assessment is carried out by laying impetus on the video consumption as well as the factors directly affecting its quality in both handoff schemes. The tests show the variance in subjective quality caused by the inter-related changes among the parameters such as overall distance among the hosts, the migration distance for streaming resuming and the screen size [4] of the end host. Concluding from the analysis of the effects of these parameters in subjective assessment, we propose an appropriate model of subjective quality which incorporates each parameter that is mainly related to the perceived quality. The proposed model would provide users with the highest quality multimedia services when adopted by the service providers or integrated multimedia servers.

2 Subjective Quality Assessment

Minimizing handoff delay is the chief factor to guarantee the quality of experience in multimedia service delivery. In this study, we demonstrated a quality model considering handoff delay between user's migration distances and streaming resuming location which varies in relation to the screen size of the end host.

2.1 Host-Level Handoff Test

The experimental conditions for our study were as follows: (1) Total distance between the two hosts- 3, 4, 6, 9 and 12M. (2) Display sizes- 46, 20 and 10 inch. (3) Video sequence- "Elephant's dream" [5] (encoded with H.264/MPEG-4 AVC 720p). (4) Evaluation method- 0–100 re-scaled Mean Opinion Score (MOS) method. (5) Subject- 18 subjects with the height between 170 and 178 cm (Mean: 174.5 cm, std.: 2.41 cm) having corrected visual acuity of 1.0 or better with normal color vision.

The subjects were allowed to displace their positions, which differentiates experiment procedures from the existing studies where the test subjects were held static. [6]. For the location tracking system, we adopted Cricket system [7] which has 3 cm tracking precision based on ultrasound and 433 MHz RF signals.

Our test procedures were as follows (see Fig. 1): (1) Before the test, all the subjects were asked to focus on the overall video quality including the degree of seamless continuity of video session, preferred viewing distance and perceived delay. (2) Video clip was shown to the subject at host A and then the subject was instructed to walk at a normal pace toward host B. (3) When the subject reached three fixed points, i.e. $d_{mg}\text{-}\#$, while migrating from host A to host B, the video session was shifted to and displayed by host B. Each point, i.e. $d_{mg}\text{-}\#$, was assessed for three displays with different screen sizes. (4) The evaluation was repeated for five different total distances.

The distance for streaming resuming while moving versus the subjective qualities in five total distance cases is shown in Fig. 2. From this analysis, we observe the followings: (1) Subjects prefer a larger visual screen over a smaller one and there is no clear relation between a fixed screen width of the hosts and the migration distance for streaming resuming, $d_{mg}\text{-}1, 2, 3$ at short distances (3M). (2) At further distances, the d_{mg} between a fixed screen size of the hosts become larger, and higher scores are acquired. Subjects are more tend to give low scores for high d_{mg} values at long distances (9–12M) as they experience a longer delay. (4) With an increase in total distance (D_{total}), there are dissimilarities in quality loss for each screen width of the hosts in the order of 46-inch <20-inch <10-inch. (5) A more significant loss of quality is found in 4–6M range compared to other ranges. From the analysis of streaming over 4–6M, we are able to consider a service transition for user-level handoff.

2.2 Mobility Transition with User-Level Handoff Test

Both host and user-level handoff schemes are analyzed in response to the screen width of the host and the migration distance subtracted three times screen width from total distance ($D_{total} - 3 \cdot W_{host}$) so as to fix the viewing distance of the end host by session transferring.

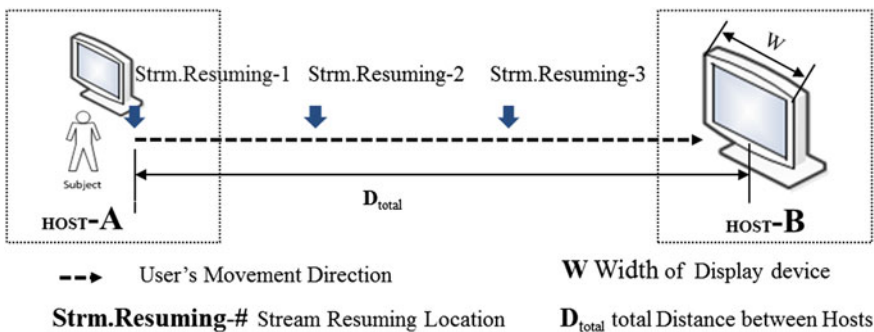


Fig. 1 Test scenario for host-level handoff

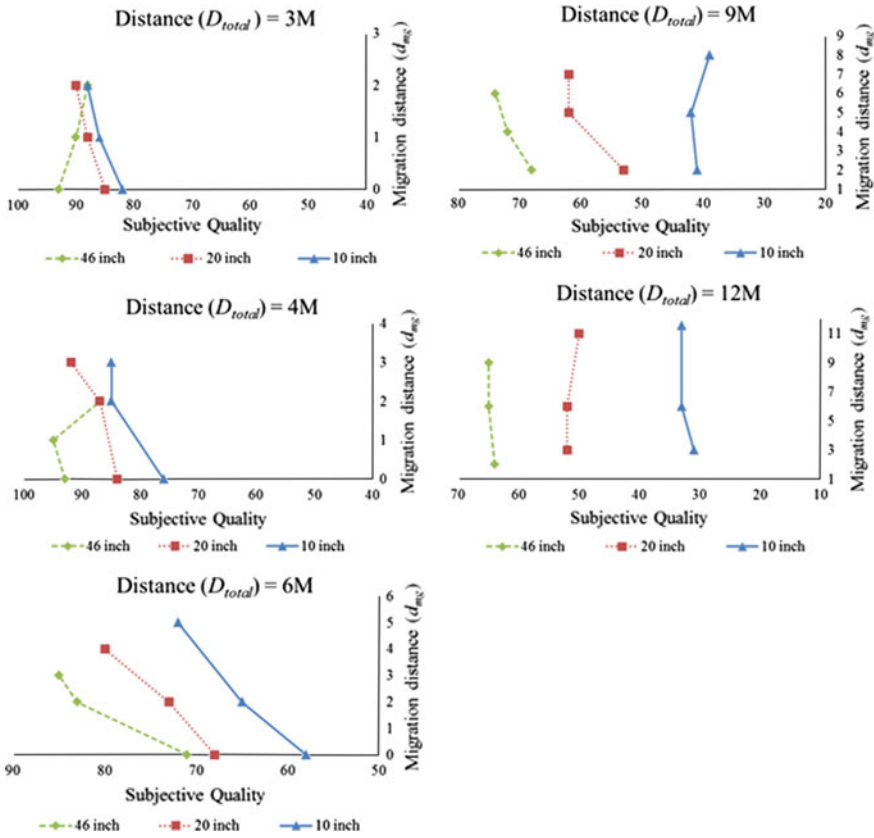


Fig. 2 Migration distance for streaming resuming with assessed subjective quality

The experiment follows the same methods used in Sect. 2.1. The 46-inch display as an end host is evaluated for ten fixed points from 0 to 9 m in 1 m interval. 20- and 10-inch displays are evaluated for 8 points at 0, 1.5, 3, 4.5, 6, 7.5, 9, 11 and 0 m and 1.5, 3, 4.5, 6, 7.5, 11 m, respectively.

From this analysis, we observe the followings: (1) As for user-level handoff, all host screen sizes show a similar sigmoidal increase in each distance case. On the contrary, for host-level handoff, the various sizes of host terminals result in differences in quality loss especially in 4–5 m range. (2) User-level handoff in a short distance between the two hosts causes annoyance due to a severe confusion in watching the screen. (3) The intersection point of the two methods occurs in 4–5 m which indicates a possibility of enhancing perceived service quality through the transition to user-level handoff.

3 New Quality Model

3.1 Host Mobility Modeling

An ideal quality metric should vary linearly with the subjective quality (SQ). The quality metric supporting host mobility is modeled as:

$$QM_{HostMobility} = (\alpha \cdot d_{mg} + \beta \cdot D_{total}) + (\gamma \cdot W_{host} \times \beta \cdot D_{total}) \tag{2}$$

where QM is the quality metric with host mobility, d_{mg} is the distance migrated by the user to continue streaming, W_{host} is the width of the end host display and D_{total} is an overall distance between the hosts while α , β and γ are constant coefficients. Through the linear regression analysis, we could obtain a new quality metric for host mobility as the following:

$$QM_{HostMobility} = (0.71 \cdot d_{mg} - 2.54 \cdot D_{total}) \left(1 + \frac{30}{W_{host}}\right) + 102.11 \tag{3}$$

The correlation coefficient between the SQ and (3) has high correlation of 0.93 (Fig. 3).

3.2 User Mobility Modeling

Figure 4 shows the results obtained from analyzing the distribution by distance using residual subjective qualities between the two handoff methods (see Fig. 3).

Fig. 3 Subjective quality profiles of the two handoff schemes based on the distance

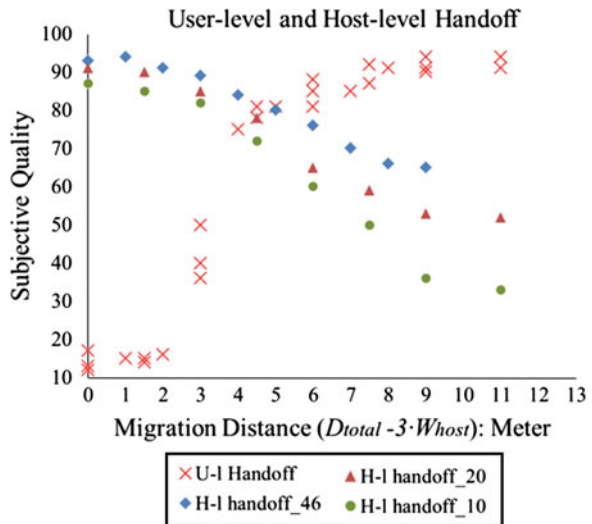
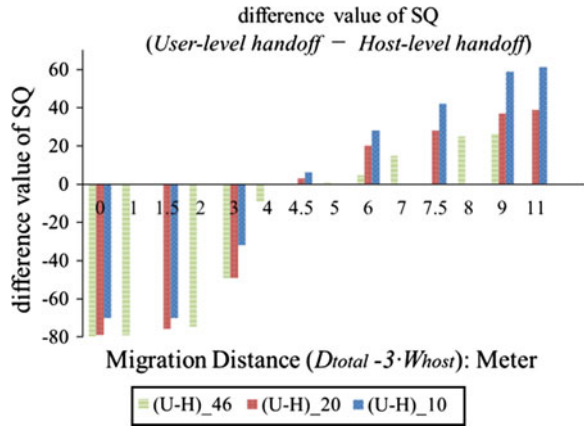


Fig. 4 Subjective quality differential between the two handoff schemes



User-level handoff achieves better performance when the total migration distance ($D_{total} - 3W_{host}$) between the two hosts is 4.5M or more. In addition, the tendency of preferences shows the form of sigmoidal function. The model considering user mobility derived by regression could be defined:

$$QM_{UserMobility} = Sc \cdot \left(\frac{\alpha \cdot W_{host}}{1 + e^{-\left(\frac{D_{total} - 3W_{host} - \beta}{\gamma}\right)}} + \delta \right) \tag{4}$$

From (4), coefficients α , β , γ and δ are obtained by regression method with differential MOS value between the two handoff methods as dependent variable and other parameters as independent variables. The value of α is $98.35/(W_{host})$, β is 3.73, γ is 1.09, δ is -84.43 and Sc is 0.21 for scaling to fit the 0–100 MOS scale. The quality resulting from the compensation of user-level handoff receives higher average value of 91.85, compared to the average value of (3) of 63.03 for the distance of 4.5 m or more.

3.3 New Quality Modeling

From the three parameters of (3) and (4), we derive a video quality metric that is modeled as is shown in (5):

$$QM_{FullMobility} = \alpha \cdot X1 + \beta \cdot X2 + \gamma \cdot X3 + \delta \tag{5}$$

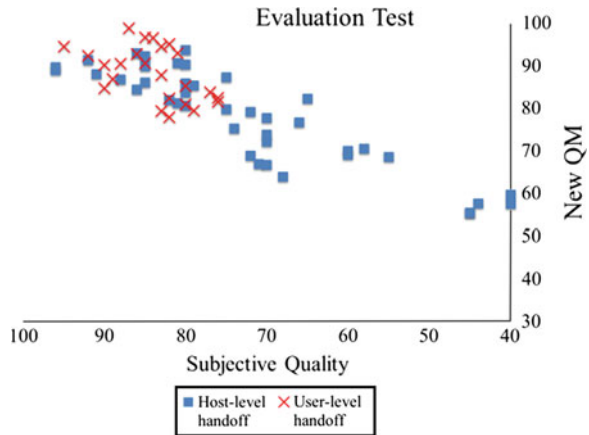
where, QM is the new proposed quality metric, $X1$ is d_{mg} , $X2$ is the term of the D_{total} and W_{host} , while $X3$ represents user-level handoff term. By (5), the new quality metric supporting full multimedia service mobility can be represented by the following equation:

$$\begin{aligned}
 &QM_{FullMobility} \\
 &= 0.71 \cdot d_{mg} - 2.54 \cdot D_{total} \left(1 + \frac{30}{W_{host}} \right) + \left(\frac{98.35}{(W_{host})^{1.29}} \right) \cdot \left(\frac{1}{1 + e^{-\frac{(D_{total} - 3 \cdot W_{host}) - 3.73}{1.09}}} \right) - 84.38 \\
 &\quad , \text{if } (D_{total} - 3 \cdot W_{host}) \geq 4.5
 \end{aligned}
 \tag{6}$$

4 Performance Evaluation

The experiment is carried out under the same conditions (Sect. 2) and procedures are taken by total of 10 subjects and “Big_buck_bunny [5]” is selected as the video sequence. Figure 5 shows the correlation between the assessed MOS and the proposed quality metric and the outcomes of new $QM_{FullMobility}$ for host-level handoff case are closely related to a linear shape when the correlation coefficient of the subjective quality is 0.89 in the evaluation set. The scatter plots of $QM_{FullMobility}$ for user-level handoff in Fig. 5 also shows a linear relationship with the assessed quality and demonstrates a more preferred quality than that of host-level handoff. In addition, new quality metric supporting full session mobility demonstrates a high average value of 92.83 compared to the average QM value with host mobility of 71.39 due to the compensation of user-level handoff for the distance of 4.5 m or more.

Fig. 5 New quality metric for full session mobility with subjective quality



5 Conclusion

Unlike the previous tests where the subjective assessment utilized either host level handoff or the user-level handoff, the subjective assessment of this study focused on both to correlate and contrast with a component analysis to provide users with the highest quality of multimedia consuming experience. We observed a high correlation, 0.89 and above with the scores for assessed quality. This indicates that our quality model is contributive for the evaluations of video quality supporting seamless mobility. The proposed quality metric would not only establish new standards of quality evaluation in video consumption in a mobility enhanced smart space but also offers valuable information to home network providers and multimedia service providers.

Acknowledgments This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No.2012R1A1A2007955) and financially supported by Korea Minister of Ministry of Land, Transport and Maritime affairs (MLTM) as U-City Master and Doctor Course Grant Program

References

1. Snoeren AC, Balakrishnan H (2000) An end-to-end approach to host mobility. In: International conference on mobile computing and networking, pp 155–166, 2000
2. Cui Y, Nahrstedt K, Xu D (2004) Seamless user-level handoff in ubiquitous multimedia service delivery. *Multimed Tools Appl* 22(2):137–170
3. Bellavista P, Corradi A, Foschini L (2007) Context-aware handoff middleware for transparent service continuity in wireless networks. *Pervasive Mobil Comput* 3(4)
4. Lee DS (2012) Preferred viewing distance of liquid crystal high-definition television. *Appl Ergon* 43:151–156
5. Xiph.org <http://media.xiph.org/video/derf>
6. ITU-R (2002) Methodology for the subjective assessment of the quality of television pictures Technical Report BT.500-11, ITU-R
7. Priyantha N, Chakraborty A, Balakrishnan H (2000) The cricket location-support system, *ACM MOBICOM*, pp 32–43

A Study of Stereoscopic 3D Technology Development Trends on Mobile

Cheong-Ghil Kim, Se-Hwan Park, Bong-Jin Back
and Taeg-Keun Whangbo

Abstract The rapid increase of mobile internet and digital technologies has shown an increased interest in demand for stereoscopic 3D to provide 3D digital contents on mobile. Furthermore, the fast growing LCD technology has already allowed 3D viewings even in Smartphone. As a result, the issues related with stereo imaging have been spotlighted greatly. This paper presents a brief overview of rapid developments in stereoscopic technologies for mobile devices to gain some perspective on the changes and progress with special emphasis on mobile 3D TV services.

Keywords Mobile 3DTV · Stereoscopic 3D · T-DMB · DVB-H · Smartphone

C.-G. Kim · B.-J. Back
Department of Computer Science, Namseoul University, Cheonan, South Korea
e-mail: cgkim@nsu.ac.kr

B.-J. Back
e-mail: genieker@naver.com

S.-H. Park
ReSEAT Program, Korea Institute of Science and Technology Information,
Seoul, Korea
e-mail: world00117@reseat.re.kr

T.-K. Whangbo (✉)
Department of Computer Science, Gachon University, Seongnam, Korea
e-mail: tkwhangbo@gachon.ac.kr

1 Introduction

3DTV and mobile TV are two emerging technologies in the area of audio–video entertainment and multimedia. In general, 3DTV assumes the content is to be viewed on large displays and simultaneously by multiple users with glasses-enabled stereoscopic display technologies or glasses-free autostereoscopic displays [1]. At the same time there have been many researches on various aspects of 3DTV content creation, coding, delivery, and system integration. As of mobile TV, standardization and legislation activities have lead to creation of similar yet content or country specific standards; for examples, the Korean 3D T-DMB [2], the European projects 3DPhone [3], and Mobile3DTV delivered through DVB-H [4].

Stereoscopic 3D utilizes the human vision system of feeling the depth of the scenes being viewed. It is the ability of our brain to fuse together the two images seen by the eyes (the stereo image pair) to form a single image, named the cyclopean image that contains embedded information about depth and an improved resolution of detail [5]. Stereoscopic 3D has its own distinct features, advantages and problems, together with other 3D viewing technologies, holograph and integrated 3D. Therefore, these technologies have been significantly increasing both in research and commercial communities. In the market, the first 3D Android Smartphone, LG Optimus 3D [6], recording, displaying, and shearing glasses-less 3D content, was introduced in 2011.

In this paper, a brief overview of rapid developments in stereoscopic technologies on mobile devices is introduced to gain some perspective on the changes and progress with special emphasis on mobile 3DTV. For this purpose, this paper is organized as followings. [Section 2](#) briefly presents the concept of stereoscopic 3D and overviews the overall structure of 3DTV system. [Section 3](#) briefly describes mobile 3DTV systems: DVH-T and T-DMB. [Section 4](#) concludes the recent status and points to the future research.

2 Background

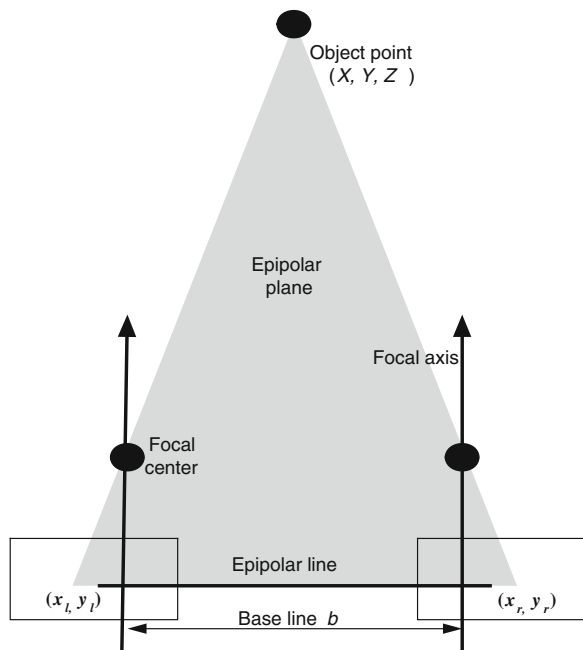
Stereoscopic 3D system can be devised using two cameras located at two different positions, which may imitate the human visual system known as binocular stereopsis that allows the visual sense to give an immediate perception of depth on the basis of the difference in points of view of the two eyes. It exists in those animals with overlapping optical fields, acting as a range finder for objects within reach. In stereo vision system, the geometry associated with solving this problem is simplified by assuming that the two cameras are coplanar with aligned image coordinate systems. [Figure 1](#) shows the basic structure for the stereo image formation and the stereo camera geometry. The center of the lens is called the camera focal center and the axis extending from the focal center is referred to as the focal axis. The line connecting the focal centers is called the baseline, b . The plane passing

through an object point and the focal centers is the epipolar plane. The intersection of two image planes with an epipolar plane makes the epipolar line. Let (X, Y, Z) denote the real world coordinates of a point. The point is projected onto two corresponding points, (x_l, y_l) and (x_r, y_r) , in the left and right images. The disparity is defined as the difference vector between two points in the stereo images, corresponding to the same point in an object, $v = (x_l - x_r, y_l - y_r)$ [7, 8].

Figure 2 [9] shows the block diagram which shows the content starts its life when it is produced and along the way to being displayed using a 3D-ready device with a number of formatting changes. Each stage may have a number of possible options.

The first stage is 3D content production section with three main approaches: live camera capture, computer generated imagery, and 2D-3D conversion. For distribution of contents, it may be a way of sending left and right views independently. However, it might be wasteful in terms of bandwidth and packaging for encoding. There are four methods for 3D content encoding: spatial compression, temporal interleaving, 2D + some form of metadata, and color shifting. There are a number of transmission platforms where 3D content may be deployed: terrestrial broadcast, cable, satellite, packaged material, IPTV, internet download, and mobile TV. In decoding, there are several options and they depend on the encoding, the delivery platform chosen and the display of choice. The following is a list of options as presented by a transcoder manufacturer: external hardware (SetTop Boxes, Blu-ray players, DVD players, gaming consoles, decoders), internal hardware (inside the TV or inside the Decoder), firmware update to

Fig. 1 Basic structure for stereo image formation and stereo camera geometry



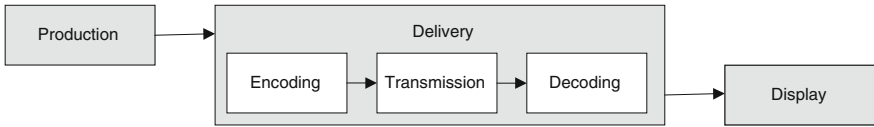


Fig. 2 Mobile 3D system

existing devices (Chipsets/STB/Decoders), new hardware (updated chipsets), and software update/download (PC, IPTV). The technology used to simulate the depth presence in a scene influences the type of display used to provide the 3D content. 3D display technologies include: anaglyph, stereoscopy, auto-stereoscopy, holography, and volumetric displays [9].

3 DVH-T and T-DMB

Mobile 3DTV system was developed by European consortium over DVB-H channel. It naturally consists of different components as shown in Fig. 3 [1], in which stereo video content is captured, effectively encoded, and then robustly transmitted over DVB-H to be received, decoded and played by a DVB-H enabled handheld [1].

Here, the stereo video framework was adopted while providing comfortable 3D experience to the user with acceptable spatial resolution and frame rate. At the stage of 3D content creation and coding, currently there is no single and generally adopted representation format for stereo video, taking specific mobile channel conditions into account. Most natural is to have two-channel stereo video. Capture of such video by synchronized cameras is relatively easy and the coding can be done efficiently, e.g. by the techniques of the emerging multi-view coding (MVC) amendment of the H.264 AVC standard. There are mainly two problems with two-channel video targeted for mobile platforms.

T-DMB, launched in Korea, is Multimedia Mobile Broadcasting (MMB), which delivers multimedia broadcasting services to mobile receivers, handheld receivers, and vehicular receivers even at high speeds matching at least IMT-2000 characteristics. T-DMB data service with video associated data service can provide static

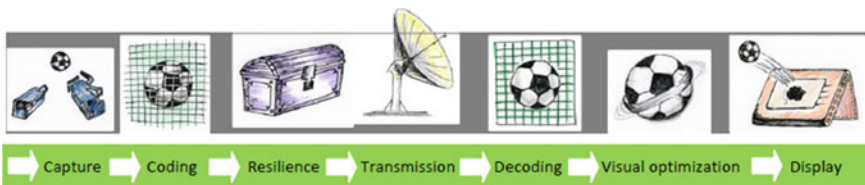


Fig. 3 Mobile 3D system

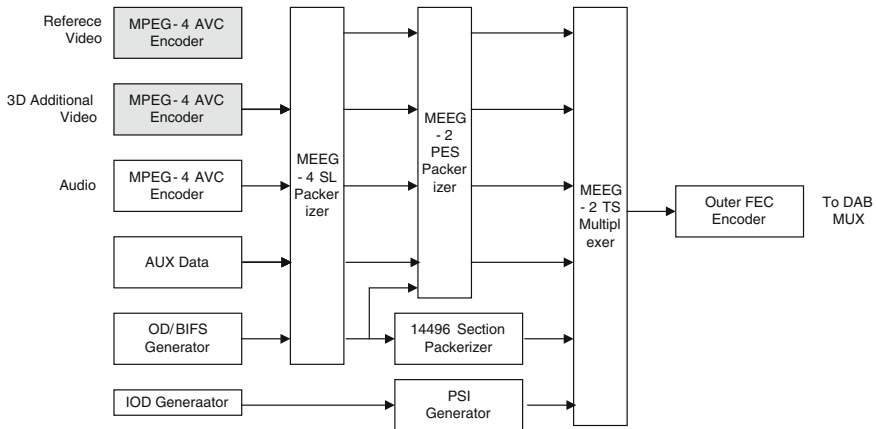


Fig. 4 3D T-DMB data service based on MPEG-4 BIFS

or dynamic image, 2D/3D graphics and text data associated to a specific video program using MPEG-4 BIFS, shown in Fig. 4 [10], which is a flexible scene description tool that allows synchronous representation of audio-visual objects in a scene. The BIFS includes information about visual properties of objects for rendering, spatial position of objects and relative time for rendering. Specifically, the proposed 3-D T-DMB receiver adopts a look-up table (LUT)-based simultaneous method to accomplish the real time implementation of DIBR algorithms, including warping, hole filling, and interleaving. Moreover, we establish the parameter values that are needed for generating the LUT based on theoretical analysis [10].

4 Conclusions

In this paper, we briefly overviewed the current stage of stereoscopic 3D technologies on mobile with examples of mobile 3DTV systems. Due to the rapid technical evolution on mobile industry, the optimal platform for 3D services is evolving continuously with various acceleration features both in software and hardware. Therefore, there will be more possibilities to expand stereoscopic 3D mobile application. Future work will cover more detailed works considering power-constrained mobile platform.

Acknowledgments This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research and Development Program 2012.

References

1. Gotchev A, Smolic A, Jumisko Pyykkö S, Strohmeier D, Akar GB, Merkle P, Daskalov N (2009) Mobile 3D television: development of core technological elements and user centered evaluation methods toward an optimized system. Proceedings of SPIE 7256, Multimedia on mobile devices 2009, vol 7256, pp 1–6, Jan 2009
2. Lee H, Cho S, Yun K, Hur N, Kim J (2008) A backward-compatible, mobile, personalized 3DTV broadcasting system based on T-DMB. In: Ozaktas H, Onural L (eds.) Three-dimensional television Springer, Berlin, pp 11–28
3. www.3dphone.org
4. www.mobile3dtv.eu
5. Edirisinghe EA, Jiang J, Edirisinghe EA, Jiang J (2000) Stereo imaging, An emerging technology. In: Proceedings of SSGRR 2000: international conference on advances in infrastructure for E-business, science, and education on the internet, Aug 2000
6. <http://www.lg.com/uk/mobile-phones/lg-P920-optimus-3d>
7. Kim CG, Sirni VP, Kim SD (2007) High performance coprocessor architecture for real time dense disparity map. J Korea Inf Proc Soc 14-A(5):301–308
8. Onural L, Sikora T, Ostermann J, Smolic A, Civanlar R, Watson J (2006) An assessment of 3DTV technologies. In: Proceedings of NAB 2006, Las Vegas, pp 456–467, Apr 2006
9. Piroddi R (2010) Stereoscopic 3D technologies. White Paper Snell Ltd., <http://www.snellgroup.com>
10. Yun K, Lee H, Hur N, Kim J (2008) Development of 3D video and 3D data services for T DMB. In: Woods AJ, Holliman NS, Merritt JO (eds) Stereoscopic displays and applications XIX. Proceedings of the SPIE, vol 6803, pp 68030Z–68030Z-12

Efficient Object Recognition Method for Adjacent Circular-Shape Objects

Sung-Jong Eun and Taeg-Keun Whangbo

Abstract The general object recognition method is based on the various area segmentation algorithms. However, there might be difficulties with segmenting the adjacent objects when their boundaries are not clear. In order to solve this problem, we propose an efficient method of dividing adjacent circular-shape objects into single object through three steps: detection of the region of interest (ROI), determination of the candidate segmentation points, and creation of a segmentation boundary. The simulation shows robust results of 6.5 % average difference ratio compared to the existing methods, even when SNR was severe.

Keywords Object recognition · Adjacent circular-shape objects · Local feature · Curve fitting

1 Introduction

Object recognition is a very important part of image processing. It can begin with area segmentation and image segmentation, which is crucial for image interpretation and an indispensable stage of image processing. As of the distribution of neighboring pixel values, non-segmentation or excessive segmentation occurs, which are common chronic problems with various image segmentation methods, and many studies have been conducted to resolve them.

S.-J. Eun · T.-K. Whangbo (✉)

Department of Computer Science, Gachon University, Sujung-Gu Seongnam,
Gyeonggi-Do, South Korea
e-mail: tkwhangbo@gachon.ac.kr

S.-J. Eun

e-mail: asclephios@hotmail.com

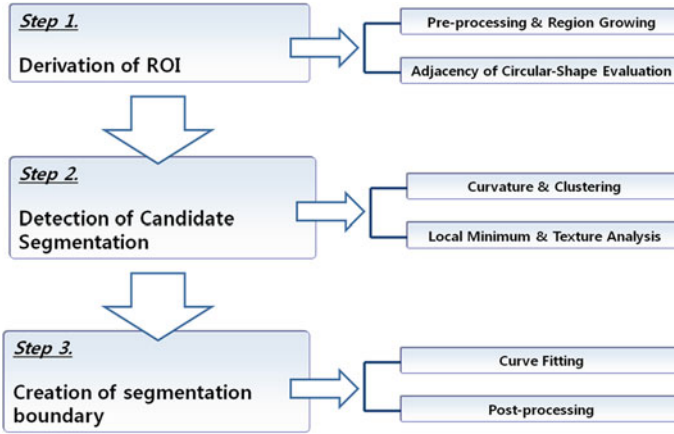


Fig. 1 Overview of the proposed method

Generally, image segmentation algorithms include the threshold value technique, the edge detection technique, region growing, and the technique of using texture characteristic values [1, 2]. So far, image segmentation has been utilized in many image processing algorithms such as Edge detection, Region growing [3], and Statistical and Structural method [4]. They have been extensively studied including multi-area segmentation methods [5]. For example, Graph Cut [6] method, GrabCut [7] method, Region Adaptive Algorithm method, and ACM or the snake method [8], and diverse snake methods have been studied extensively [9, 10].

However, the segmentation of the adjacent objects becomes inaccurate if the boundaries are vague. Therefore, segmentation algorithms for the accurate identification of two or more adjacent objects as single objects have been studied for a long time in spite of scarifying computing resources. This paper proposes a more accurate method of finding the boundaries of adjacent objects. Details will be introduced in the following chapters.

2 Segmentation Method of Adjacent Circular-Shape Objects

The proposed method consists of three steps, derivation of ROI, detection of candidate segmentation, and creation of segmentation boundary, as shown in Fig. 1 and details will be covered in subsections.

2.1 Derivation of Region of Interest (ROI)

As the proposed method presupposes the adjacency of circular-shape objects, the initial ROI are also determined based on the circular-shape objects. First,

smoothing filtering is applied to the inputted images to obtain a meaningful gradient image with a Gaussian filter; and then the gradient image was computed with the following Eq. 1.

$$|g(A) - g(B)| \leq \theta \quad (1)$$

In Eq. 1, $g(A)$ and $g(B)$ represent the gray level value in pixel A and B, θ does the threshold value. The empty areas are filled; the Compactness and Elongation are calculated; convergence is repeated until the step where it moves away from the true circle. The critical values are determined beforehand, and if the growth stops, the area is determined to be an ROI. This process is shown by the conditions of the following Eq. 2.

$$Compactness = \frac{P^2}{4\pi A} \quad Elongation = \frac{D_{\min}}{D_{\max}}. \quad (2)$$

According to Eq. 2, P represents the circumference of growing region, A the Area in growing region, D_{\min} the major axis in growing region, D_{\max} the minor axis in growing region, respectively. Thus, in this study, the ROI was determined based on the point where the object shape moved away from the circle.

2.2 Detection of Candidate Segmentation

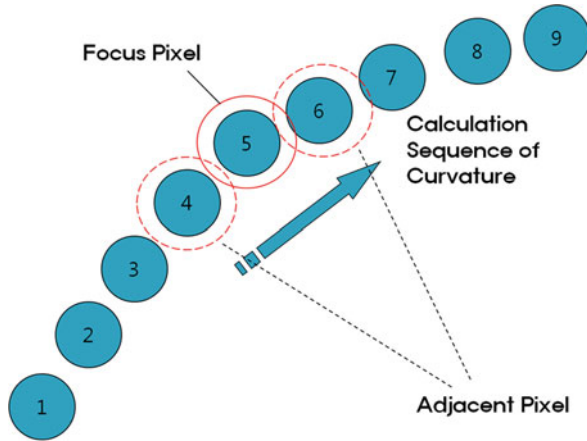
To find the area whose segmentation boundary must be corrected, the curvature [11] of the contour is computed based on the detected ROI. The curvature calculation process is illustrated in the figure below, together with all the pixels whose curvature values do not coincide with the critical value as the mean.

In Fig. 2, based on pixel no. 5, the gradient difference between the neighboring pixel nos. 6 and 4 is raised to the second power, and the resulting value is divided by the Euclidean distance difference between the two pixels. Then based on a total of eight neighboring pixels (nos. 6, 4, 7, 3, 8, 2, 9, and 1), pixel no. 5's curvature is calculated. Eight neighboring pixels are used to calculate this curvature because the curvature calculation experiment proved that this is the optimal number. The relevant curvature calculation can be defined again by the following Eq. 3:

$$\begin{aligned} Curvature(N) &= (g(N+1) - g(N-1))^2 / d((N+1) - (N-1)) \\ &+ (g(N+2) - g(N-2))^2 / d((N+2) - (N-2)) \\ &+ (g(N+3) - g(N-3))^2 / d((N+3) - (N-3)) \\ &+ (g(N+4) - g(N-4))^2 / d((N+4) - (N-4)) \end{aligned} \quad (3)$$

where g denotes the slope value and denotes the Euclidean distance. The next pixel's (no. 6's) curvature can likewise be calculated using pixel nos. 7, 5, 8, 4, 9,

Fig. 2 Concept of curvature calculation



3, 10, and 2. In this way, the curvatures of all the pixels on the peripheral line can be calculated.

Afterwards, the derived pixels are clustered with Mean-Shift clustering [12]. And then, the local minimum values of the pixels crossing at right angles are calculated. They become the first candidate segmentation points. The final candidate segmentation points are detected by considering the texture information of the adjacent areas of the detected first candidate points. The adjacent areas are the pixels crossing at right angles. Figure 3 below shows the method of determining the local minimum value, and Eq. 4 shows the calculation of the entropy value, which is the texture information used. Here, $P_{i,j}$ is used as the weight of each pixel. This calculation determines the peak of the entropy value as the final candidate segmentation point.

$$ENT : \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} P_{i,j} (-\ln P_{i,j}) \tag{4}$$

2.3 Creation of Segmentation Boundary

The final candidate segmentation points are the sampling results, and curve fitting is needed to connect them. For curve fitting, the boundary is created through the Catmul-Rom spline curve [13]. Equation 5 shows this curve-fitting method. Lastly, the boundary is smoothed using a Median filter.

$$Q(t) = 0.5 \times (1.0f, t, t^n) \times [Mat_{n \times n}] \times [P_{n-1}] \tag{5}$$

Given the control points $P_0, P_1, P_2,$ and P_3 and the value t , the location of the point can be calculated. P represents the control points, t the portion of the distance between the two nearest control points, Mat the $n \times n$ matrix, respectively.

3 Experiment

An experiment to assess the accuracy of the proposed algorithm was conducted with 60 medical images, including MR and cell images. The result is shown in Fig. 3. Then images with different SNRs were created through the assessment of the ground truth concept.

As shown in Fig. 3, more accurate results were obtained compared to the existing ACM and Region Growing algorithms. The results were compared with the reference image achieved by a specialist doctor. Thus, the accuracy of the method was evaluated quantitatively. Towards this end, the difference ratio between the reference image and the area from the proposed method was calculated, and can be expressed by the following Eq. 6.

$$R_{diff} = \frac{|R_{criteria} - R_{proposed}|}{R_{criteria}} \times 100 \tag{6}$$

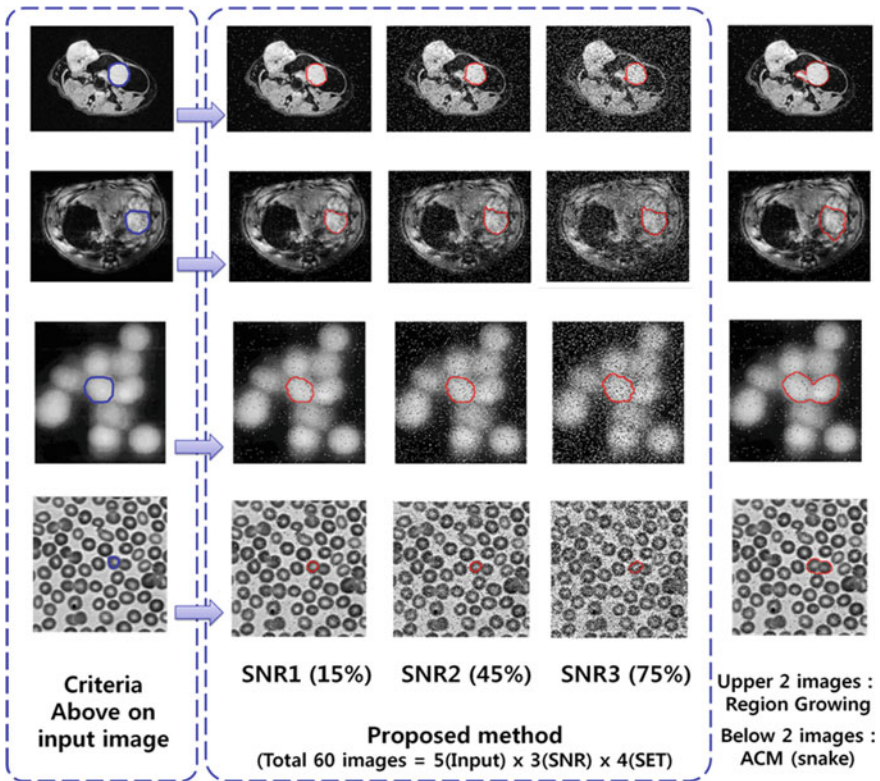


Fig. 3 Experiment results of the proposed method and exist method

Table 1 Results of comparison of the proposed method with the other methods

Method	Average area difference ratio (summation of the SNR1 to 3) (%)
Region growing	22.5
Snake	15.2
Proposed method	6.5

In Eq. 6, R_{diff} denotes the area difference ratio, $R_{criteria}$ denotes the area of the reference image, and $R_{proposed}$ represents the area created by the proposed method. For this experiment, a total of 60 medical images were processed, and the relevant image criteria were evaluated according to the results of the proposed method and of Eq 6, after a specialist doctor established the baseline using Adobe Photoshop CS. As a result, an average area difference ratio of 8.5 % was determined and shown in Table 1.

4 Conclusions

In this paper, an efficient method of dividing adjacent circular-shape objects into single objects is proposed. More robust results were derived in the experiment compared to the existing methods, even when SNR was severe. The future works will focus on improving the accessibility of the object identification method through the use of the characteristics of the objects to be identified.

Acknowledgments This research was supported by Gachon University in 2012; by the Ministry of Knowledge Economy of Korea under its Convergence Information Technology Research Center support program (NIPA-2012-H0401-12-1001) supervised by the National IT Industry Promotion Agency.

References

1. Kang CC, Wang WJ A novel edge detection method based on maximization of the objective function. *Pattern Recognit* 40(2):609–618
2. Norio B, Norihiko I, Toshiyuki T Image area extraction of biological objects from a thin section image by statistical texture analysis. *Electron Microsc* 45:298–306
3. Muerle JL, Allen DC (1968) Experimental evaluation of a technique for automatic segmentation of objects in complex scenes. *IPPR*, Thompson
4. Unser M (1995) Texture classification and segmentation for using wavelet frames. *IEEE Trans Image Process* 4(11):1549–1560
5. Li W, Zhou C, Zhang Z (2004) Segmentation of the body of the tongue based on the improved snake algorithm in traditional Chinese medicine. In: *Proceedings of the 5th world congress on intelligent control and automation*, pp 15–19
6. Zabih R, Kolmogorov V (2004) Spatially coherent clustering using graph cuts. In: *Proceedings of computer vision and pattern recognition*, vol 2, pp 437–444

7. Rother C, Kolmogorov V, Blake A (2004) GrabCut: interactive foreground extraction using iterated graph cuts. *ACM Trans Graphics* 23(3):309–314
8. Kass M, Witkin A (2004) Demetri Terzopoulos active contour models. *Int J Comput Vision* 1:321–331
9. Ng EYK, Chen Y (2006) Segmentation of the breast thermogram: improved boundary detection with the modified snake algorithm. *J Mech Med Biol* 6(2):123–136
10. Kang DJ, Kweon IS (1999) A fast and stable snake algorithm for medical images. *Pattern Recogn Lett* 20(10):1069
11. Murphy TM, Math M, Finke LH (2003) Curvature covariation as a factor of perceptual salience. In: *International IEEE EMBS CNECI*, pp 16–19
12. Comaniciu D, Meer P (1997) Mean shift analysis and application. In: *Seventh international conference computer vision and pattern recognition*, pp 750–755
13. Catmull E, Rom R (1974) A class of local interpolating splines. *Comput Aided Geom Des* 317–326

Part VIII
Convergence Data Mining
and Artificial Intelligence

Improved View Selection Algorithm in Data Warehouse

Jong-Soo Sohn, Jin-Hyuk Yang and In-Jeong Chung

Abstract In order to minimize the query processing time, a data warehouse maintains materialized views of aggregate data derived from a fact table. However, due to the expensive computing and space costs materializing the whole relations instead of part of the relations results in much worse performance. Consequently, proper selection of appropriate views to be materialized is very important to get a precise and fast response in the data warehouse. However, this view selection problem is NP-hard problem, and there have been many research works on the selection of materialized views. In this paper we propose an improved algorithm to overcome problems of existing view selection algorithms. In the presented algorithm, we first construct the reduced tables in the data warehouse using clustering method among data mining techniques, and then we consider the combination of reduced tables as the materialized views instead of combination of the original base relations. For the justification of the suggested idea, we show the experimental results in which time as well as space costs are about 1.7 times better than the conventional approaches which considered all the tuples in a relation to materialize.

Keywords Materialized views · Data warehouse · Clustering

J.-S. Sohn · I.-J. Chung (✉)
Department of Computer and Information Science, Korea University,
208 Seochangri, Sejong City, South Korea
e-mail: chung@korea.ac.kr

J.-S. Sohn
e-mail: mis026@korea.ac.kr

J.-H. Yang
Korea Institute of Planning and Evaluation for Technology in Food Republic
of Korea, An-yang, South Korea
e-mail: jhy@ipet.go.kr

1 Introduction

The relational database (RDB) is designed mainly for real time transaction processing such as On Line Transaction Processing (OLTP), it is improper for multi-dimensional data analysis such as On Line Analytic Processing (OLAP) or decision support system. In order to overcome this problem and to maximize the use of information from the huge amount of accumulated data effectively, data warehouse was introduced.

A view in a data warehouse is a virtual relation that is derived from a base relation or other view. Since we cannot maintain and materialize all possible views in a data warehouse due to the expensive computing time and space constraint, only a subset of views should be selected to be materialized. However, proper selection of materialized views in a data warehouse is NP-hard problem [1]. So far, there have been many research works on the selection of materialized views in a data warehouse [2–8], however these previous works have problems which we will describe in the following section.

In order to overcome these problems of the related previous works, we propose an efficient algorithm which uses clustering technique to select the materialized views, and thus can accelerate the response time as well as save the storage space. For the justification of the proposed algorithm, we show two independent experimental results: one is the ‘pubs’ database, used for educational purposes, and the other is much larger Enterprise Resource Planning (ERP) data base, currently being used in one of the leading enterprise in computer form design and manufacturing in South Korea. Both experimental results show that both space and time costs are approximately 1.8 times better than the conventional approaches.

2 Related Works

For the proper selection of materialized views, [2] proposed a greedy algorithm to minimize the query processing cost in the special case of the “data cubes”. However this paper does not mention the view maintenance cost and space constraint. In [3], an algorithm based on heuristic greedy method was proposed. However, this research has an inefficient evaluation tool. In research [4], the suggested HA_{mvpp} algorithm requires too much time to produce Multiple View Processing Plan (MVPP).

Algorithms in [1, 3] aim at minimization of the query processing cost. [1] is a variation of this algorithm and it aims at minimizing the total query processing time under the constraint of total view maintenance cost. Recently some artificial intelligence approaches such as genetic algorithms have been proposed to achieve the improved results in the view maintenance and query processing [5–8]. However these genetic algorithms have some problems due to the impractical solutions [9].

3 IVSA: Improved View Selection Algorithm

3.1 Improved View Selection Algorithm

In a different manner of conventional algorithms, we present an improved algorithm for selecting views to materialize using the clustering method among data mining techniques [10–13]. The proposed algorithm Improved View Selection Algorithm (IVSA), firstly finds high density clusters from the dimensions of the given tables, and secondly, produces the reduced tables using the found clusters. Next, the MVPP is produced using the reduced tables, and finally, materialized views are selected from the MVPP in accordance with cost estimation. The technique of materializing views is required to minimize the query response time in a data warehouse, which provides guidelines to enterprise managers through the analysis of market trends by supporting various OLAP capabilities.

The proposed IVSA has following four steps:

- Step 1: Find the high-density clusters from k-dimensional relations.
- Step 2: Produce the reduced tables using upper and lower bound values of the clusters found.
- Step 3: Establish the MVPP using reduced tables.
- Step 4: Select the materialized views while considering improvement of query response time and view maintenance cost.

```

IVSA( $\tau$ ,  $n$ ,  $T$ ,  $Q$ ,  $SC$ ,  $UDT$ ,  $UET$ ) {
/* : user's input threshold, n: number of queries or tables, T: set of target tables */
/* Q: set with n queries, SC: user's input space constraint */
/* UDT: user's input clustering dimensions which must be included */
/* UET: user's input clustering dimensions which must be excluded */
 $C = \emptyset$ ; /* set of clusters */  $RT = \emptyset$ ; /* set of reduced tables */
 $VP = \emptyset$ ; /* set of views used in query processing plan */
 $MV = \emptyset$ ; /* set of views to be materialized */
for ( $i = 0$ ;  $i < n$ ;  $i ++$ ) {  $C = C \cup \text{find\_cluster}(\tau, n, Ti, UDT, UET)$ ; }
for ( $i = 0$ ;  $i < n$ ;  $i ++$ ) {  $RT = RT \cup \text{generate\_reduced\_table}(Ci, Ti, RTi)$ ; }
make_mvpp( $n$ ,  $Q$ ,  $RT$ ); select\_view( $VP$ ); return  $MV$ ; }
/* step 1 */ find\_cluster(( $\tau$ ,  $n$ ,  $Ti$ ,  $UDT$ ,  $UET$ ) {
 $T = Ti$ ;  $target = 0$ ; /* variable for attributes' reflection density */
for ( $i = 0$ ;  $i < n$ ;  $i ++$ ) { for ( $j = 0$ ;  $j < n$ ;  $j ++$ ) {
/* primary key, foreign key, and user's input dimension of tables are excluded */
if ( $Ti.dj == \text{primary\_key} \parallel Ti.dj == \text{foreign key} \parallel Ti.dj == UETi.dj$ ) continue;
/* if a dimension is user's specified input dimension, it is included */
if ( $Ti.dj == UDTi.dj$ ) { for ( $k = 0$ ;  $Ti.di.low[k] \neq \text{NULL}$ ;  $k ++$ ) {
/* select a range of lower bound and upper bound for cluster */
 $C.i = Ti.di.low[k], Ti.di.high[k]$ ; } break; } /* move to the next table */

```



```

else if ( $\prod(Ti.di, Ti.dj) > \tau$  &&  $[C.i] > target$ ) {target =  $[C.i]$ ;
for ( $k = 0$ ;  $Ti.di.low[k] \neq \text{NULL}$ ;  $k++$ ) {C.i =  $Ti.di.low[k]$ ,
Ti.di.high[k];}} return C;
/* step 2 */generate_reduced_table(Ci, Ti) {/* operator  $\leftarrow$  returns index */
tmp  $\leftarrow Ti.Ci.low[0]$ ;
for ( $k = 0$ ;  $Ti.Ci.low[k] \neq \text{NULL}$ ;  $k++$ ) {
while ( $[tmp] \geq Ti.Ci.low[k]$  &&  $[tmp] \leq Ti.Ci.high[k]$ ) {Copy tuple from Ti
to RTi; tmp ++;}}
return RTi;}
/* step 3 */make_mvpp(n, Q, RT) {
for ( $i = 0$ ;  $i < n$ ;  $i++$ ) {
Make vpi using Q and RT as base relation instead of T;
Count the number of nodes in vpi and save into NNi;}
for ( $i = 0$ ;  $i < n$ ;  $i++$ ) {for ( $j = 0$ ;  $j < NNj$ ;  $j++$ ) {for ( $k = 0$ ;  $k < NNk$ ;
k ++)}
VP = VP  $\cup$  vpi; if ( $vpi.nodej == VPi.nodek$ ) VPi.nodek.fq ++;}} return VP;}
/* step 4 */select_view(VP) {
for ( $i = 0$ ;  $i < n$ ;  $i++$ ) {for ( $j = 0$ ;  $j < n$ ;  $j++$ ) {
VPi.Ca = VPi.Ca + VPi.nodej.Ca; VPi.Cm = VPi.Cm + VPi.nodej.Cm;
VPi.Cv = VPi.Cv + VPi.Ca + VPi.Cm;}
VP.Ca = VP.Ca + VPi.Ca; VP.Cm = VP.Cm + VPi.Cm;
VP.Cv = VP.Cv + VP.Ca + VP.Cm;}
/* sort the elements of VP in ascending order according to the value of Cv */
Sort(VP);/* select views within the bound of specified SC */
for ( $i = 0$ ;  $i < n$ ;  $i++$ ) {/* operator  $\Sigma$  returns storage space */
if ( $\Sigma TMV < SC$ ) {MV = MV  $\cup$  VPi; MV.Cv = MV.Cv + VPi.Cv;} else
break;}
return MV;}

```

3.2 Properties of Improved View Selection Algorithm

In the first step of the algorithm, the high-density cluster for target base relations is found using the clustering method of among data mining techniques. For each dimension of the table, the dimension with the maximum density value is selected, which exceeds the user's input threshold τ . As a novel approach which is not considered in conventional algorithms, this technique with clustering is crucial from the standpoint of providing an opportunity to implicitly utilize important information overlooked.

In the second step of the algorithm, reduced tables containing the only corresponding tuples are produced by using the lower and upper bound values of the selected dimension for each table. While traditional algorithms consider all the tuples of a base relation for materializing, the targets of materializing are restricted to the tuples of the reduced tables in the proposed algorithm IVSA. Therefore, it

can achieve the goals of improvement in query response time and saving of storage for views.

In the third step of the algorithm, we produce an MVPP using the reduced tables generated in the previous step. The existing algorithm [4] proposed the 0–1 integer programming method and HA_{mvpp} for establishing MVPP. While this 0–1 integer programming technique produces optimal MVPP, it takes too much time to implement. In our algorithm, we propose the off-line procedure for establishing MVPP using query frequency.

In the fourth step of the algorithm, the views which can derive benefits in the case of materialized ones were selected within the bounds of the user’s input space constraint, while considering view processing time cost and view maintenance cost in the produced MVPP. The conventional algorithms consider only the cost for join operation and restrict query frequency to the query itself. We argue that these cost estimation methods leave out some important factors in cost. In the IVSA, cost for the select operation is supplemented to cost estimation formulation.

4 Implementation Results and Analysis

4.1 Experimental Result of Materials Handling in the ERP

In this chapter, we present an experimental result on the large Enterprise Resource Planning (ERP) data base, currently being used in one of the leading enterprise in computer form design and manufacturing in South Korea. We use the keyword-based search method to accelerate the query response time. The ERP database of that company has altogether 981 tables where each table has rows from thousands to tens of thousands. In this paper we show some examples of making materialized view from the materials handling tables in the ERP database. Followings are the tables used in purchasing materials for production in the company.

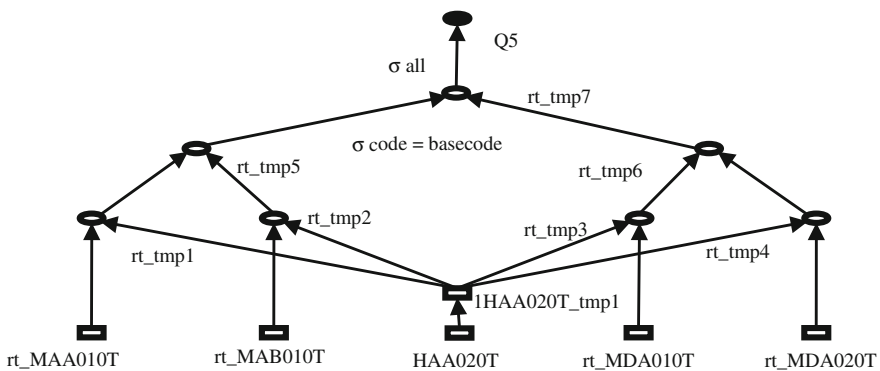


Fig. 1 MVPP for the query Q5

Table 1 Cost computation for the query Q1 with reduced tables

Table	<i>f_q</i>	#	<i>C_a</i> Q5	<i>C_m</i> Q5	<i>C_v</i> Q5
HAA020T	1	483	483	0	46,851
rt_MAA010t	1	1,372	133,267	0	133,267
rt_MAB010T	1	2,148	202,119	0	202,119
rt_MDA010T	1	8,567	803,262	0	803,262
rt_MDA020T	1	1,652	157,093	0	157,093
HAA020_tmp	1	93	1,428,179	0	1,428,179
rt_tmp1	1	194,333	204,513	390,869	595,382
rt_tmp2	1	15,891	16,437	31,681	48,118
rt_tmp3	1	912,767	91,466	984,148	1,075,614
rt_tmp4	1	192,778	196,452	381,445	577,897
rt_tmp5	1	137,445	142,483	274,329	416,812
rt_tmp6	1	1,449,420	1,445,282	2,836,808	4,282,090
rt_tmp7	1	27,843	27,843	4,801,296	4,829,139

We show the detailed steps to solve several query processing for purchasing items and placing order of materials in the company and then evaluate their performance. In order to improve response time of keyword based search, we consider the articles of the law containing the related notices as clustering targets. After generating the reduced table for the materials handling of ERP: ERP handling materials, we subsequently generated MVPP on the following query. Consider the following query Q1.

(Q1) Make a list for unit prices and enterprises on the orders

Query Q1 searches the table concerning clients and materials. User enters employee number and query Q1 displays the result to the user. Then query Q1 makes temporary table ‘HAA020T_tmp1’ of input employee number by searching HAA020T table. HAA020T_tmp1 is used to search specific information for clients and materials Fig. 1 displays MVPP of query Q1.

Table 1 is a result of time cost estimation in executing query Q1 when we insert the material purchase data into this algorithm—IVSA. Table 2 is a computation result of executing query Q1 with conventional method without using reduced table.

In this example, we set up user variables SC as 150,000. And then tuples rt_tmp1, rt_tmp3, rt_tmp4 and rt_tmp6 are selected one by one as materialized view. As displayed in Table 3, we can check that algorithm presented in this paper shows 1.62 times better storage space and 1.9 times better performance in terms of total cost.

Table 3 is a summarization of Tables 1 and 2. Table 3 shows that the query processing time is 1.77 times faster and storage space is 1.79 times smaller using the suggested method in this paper when user space constraint variables SC were not given.

Table 2 Cost computation for the query Q1 without reduced tables

Table	f_q	#	C_a Q5	C_m Q5	C_v Q5
HAA020T	1	483	483	0	483
MAA010t	1	2,332	227,759	0	227,759
MAB010T	1	3,652	356,579	0	356,579
MDA010T	1	14,564	1,422,165	0	1,422,165
MDA020T	1	2,808	274,240	0	274,240
HAA020_tmp	1	93	2,011,820	0	2,011,820
tmp1	1	330,366	350,188	666,942	1,017,130
tmp2	1	27,015	28,635	54,537	83,172
tmp3	1	1,369,151	1,451,300	2,764,041	4,215,341
tmp4	1	308,445	326,951	622,688	949,639
tmp5	1	206,168	218,538	416,211	634,748
tmp6	1	2,029,188	2,150,939	4,096,525	6,247,464
tmp7	1	28,170	28,170	8,448,525	8,476,695

Table 3 Performance comparison on the database of ERP system

		Conventional algorithms	IVSA
Partial materialization case	Materialized views	tmp1, tmp2, tmp5, tmp6	rt_tmp1, rt_tmp2, rt_tmp5, rt_tmp6
	Total cost	127,417	72,632
	Storage space	41,769	23,612
Full materialization case	Materialized views	ALL	ALL
	Total cost	593,591	332,180
	Storage space	84,713	47,201

5 Conclusions

As a technique of materialized views, this paper proposes IVSA algorithm which adopts the data mining clustering technique. In the proposed algorithm, the user can specify a dimension for mandatory clustering. This function excludes the possibility of leaving out the important information. The user can also specify the threshold value that indicates the compression strength of clusters. Finally, the user is able to input a space constraint value within which materialized views are selected. These kinds of user interfaces are not found in conventional algorithms.

The proposed algorithm IVSA, firstly, finds high density clusters from the dimensions of the given tables, and secondly, produces the reduced tables using the found clusters. Next, the MVPP is produced using the reduced tables, and finally, materialized views are selected from the MVPP in accordance with cost estimation.

As shown in the experimental results, the proposed algorithm achieves 1.76 times better on average performance in terms of both query response time and storage space of materialized views. Even in the case where the value of the space

constraint variable is not specified (i.e., when we assume there is no space constraint), our algorithm shows 1.78 times better on average performance database for ERP database, respectively.

Broadly, there lie two issues with the data warehouse. The first is selection of materialized views, and the other is maintenance of the views for consistency of a data warehouse. IVSA in this paper is in regards to the first issue. As future works, we will focus on how to update and maintain the reduced tables when there occurs any update in the source data.

References

1. Gupta H, Mumick I (1999) In: Selection of views to materialize under a maintenance cost constraint. Database theory—ICDT'99, pp 453–470
2. Harinarayan V, Rajaraman A, Ullman JD (1996) Implementing data cubes efficiently. SIGMOD Rec 25(2):205–216
3. Gupta H (1997) In: Selection of views to materialize in a data warehouse. Database theory—ICDT'97, pp 98–112
4. Yang J, Karlapalem K, Li Q (1997) Algorithms for materialized view design in data warehousing environment. 136–145
5. Zhang C, Yao X, Yang J (2001) An evolutionary approach to materialized views selection in a data warehouse environment. IEEE Trans Syst Man Cybern Part C: Appl Rev 31(3):282–294
6. Lee M, Hammer J (2001) Speeding up materialized view selection in data warehouses using a randomized algorithm. Int J Coop Inf Syst 10(3):327–354
7. Choi CH, Yu J, Gou G (2002) What difference heuristics make: Maintenance-cost view-selection revisited. In: Advances in web-age information management, 313–350
8. Yu JX, Yao X, Choi CH et al. (2003) Materialized view selection as constrained evolutionary optimization. IEEE Trans Syst Man Cybern Part C-Appl Reviews 33(4):458–467
9. Ashadevi B, Subramanian R (2009) Optimized cost effective approach for selection of materialized views in data warehousing. Int J Computer Sci Technol 9(1)
10. Chen MS, Han J, Yu PS (1996) Data mining: An overview from a database perspective. IEEE Trans Knowl Data Eng 8(6):866–883
11. Agrawal R, Imielinski T, Swami A (1993) Database mining: a performance perspective. IEEE Trans Knowl Data Eng 5(6):914–925
12. Berson A, Smith SJ (1997) Data warehousing, data mining, and OLAP. Computing McGraw-Hill, New York
13. Kalnis P, Mamoulis N, Papadias D (2002) View selection using randomized search. Data Knowl Eng 42(1):89–111

A Novel Weighting Technique for Mining Sequence Data Streams

Joong Hyuk Chang and Nam-Hun Park

Abstract Many of recent computer applications generate data as a form of data streams, so a study on mining data streams can give valuable results being widely used in the applications. In this paper, a novel weighting technique for mining interesting sequential patterns over a sequence data stream is proposed. Assuming that a sequence with small time-intervals between its data elements is more valuable than others with large time-intervals, the novel interesting sequential pattern is defined and found by analyzing the time-intervals of data elements in a sequence as well as their orders.

Keywords Weighted sequential pattern · Time-interval weight · Sequence data streams · Data stream mining

1 Introduction

Sequential pattern mining aims to discover interesting sequential patterns in a sequence database, and it is one of the essential data mining tasks widely used in various application fields such as Web access pattern analysis, customer purchase pattern analysis, and DNA sequence analysis. In many of the previous researches on sequential pattern mining problems, sequential patterns and items in a

J. H. Chang

Department of Computer and Information Technology, Daegu University, Naeri Jillyang Gyeongsan, Gyeongbuk 712-714, Republic of Korea
e-mail: jhchang@daegu.ac.kr

N.-H. Park (✉)

Department of Computer Science, Anyang University, 102 Samsungli Buleunmyun Ganghwagun, Incheon 417-833, Republic of Korea
e-mail: nmhnpark@anyang.ac.kr

sequential pattern have been considered uniformly. However, they have different weights in real world applications, and thus more interesting sequential patterns can be found when their different weights are considered in sequential pattern mining. Based on this observation, weighted sequential pattern mining [1–3] has recently been proposed and actively studied. In weighted sequential pattern mining, the weight of information is used in finding interesting sequential patterns, which is derived from its quantitative information and value in a real world application. For example, in a retail database, the quantity and price of an item being sold can be considered as its weight.

For a sequence or a sequential pattern, not only the generation order of data elements but also their generation times and time-intervals are important because they can help to get more valuable sequential patterns. In [4, 5], several sequential pattern mining algorithms have been presented which consider a time-interval between two successive items in a sequential pattern. However, they simply consider a time-interval between two successive data elements as an item. If the importance of sequences in a sequence database is differentiated based on the time-intervals in the sequences, more interesting sequential patterns can be found.

In general a sequence with small time-intervals between its data elements is more valuable than others with large time-intervals. Motivated by this observation, this paper proposes a new framework for mining novel interesting sequential patterns over time-interval sequence data streams and a mining method based on the new framework. First, a technique to get the weight of a sequence in a time-interval sequence data stream is presented, which is derived from the time-intervals of items in the sequence. Based on the weight of a sequence, a novel interesting sequential pattern of a time-interval weighted sequential pattern is defined, and a framework for finding the patterns in a time-interval sequence data stream is presented. In addition, adapting the proposed framework to the conventional method of mining sequential patterns over a data stream, this paper proposes a mining method of novel interesting sequential patterns over a time-interval sequence data stream, which can find time-interval sequential patterns over the data stream in a short time with a small memory.

The rest of this paper is organized as follows: The definition of a time-interval sequence data stream is described in Sect. 2. Section 3 presents a novel weighting technique for mining time-interval sequence data streams, which are based on time-intervals of data elements in a data stream. In Sect. 4, the effectiveness of the novel weighting technique is verified through a several experiments. Finally, Sect. 5 concludes this paper.

2 A Time-Interval Sequence Data Stream

Conventional sequential pattern mining considers the order of data elements of a sequence in general, so that a sequence in a sequential data stream is represented as an ordered list of data elements [6]. However, a time-interval sequence data

stream discussed in this paper has generation time information for each data element in the data stream, and is defined as follows: (i) Let $I = \{i_1, i_2, \dots, i_n\}$ be a set of current items, which have been used as a unit information of an application domain. (ii) A sequence S is an ordered list of items and its time stamp list $TSL(S)$ is an ordered list of corresponding time stamps of the items, which stand for the time when the items occur. They are denoted as $S = \langle s_1, s_2, \dots, s_l \rangle$ and $TSL(S) = \langle t_1, t_2, \dots, t_l \rangle$, respectively, where s_j is an item and t_j is the time stamp of s_j for $1 \leq j \leq l$. In addition, the relationship $t_{j-1} \leq t_j$ for $2 \leq j \leq l$ is satisfied. In a sequence, if items occur at the same time, they are ordered alphabetically. The length of S , $|S|$, is the number of items that form the sequence, and a sequence with n items is called an n -sequence. A sequence $\alpha = \langle a_1, a_2, \dots, a_n \rangle$ is called a *subsequence* of another sequence $\beta = \langle b_1, b_2, \dots, b_m \rangle$, and β is a *super-sequence* of α , if there exist integers $1 \leq j_1 < j_2 < \dots < j_n \leq m$ such that $a_1 = b_{j_1}$, $a_2 = b_{j_2}$, ..., $a_n = b_{j_n}$. Each sequence has a unique sequence identifier *SID*. A sequence generated at the k^{th} turn is denoted by S_k and its transaction identifier *SID* is k . (iii) When a new sequence S_k is generated, the current time-interval sequence data stream $TiDS_k$ is composed of all sequences that have ever been generated so far, i.e., $TiDS_k = \langle S_1, S_2, \dots, S_k \rangle$, and the *total number of sequences* in $TiDS_k$ is called its size and denoted by $|TiDS_k|$. In the rest of this paper, a sequence data stream means a time-interval sequence data stream, if not specified otherwise.

A sequence is represented as an ordered list of items in this paper, while it is represented as an ordered list of itemsets in practice. However, the new representation of sequences described herein is in fact a typical one. A sequence in the previous format can be transformed to the new format by sorting all the items first by time and then alphabetically. Likewise, a sequence in the new format can be transformed into the traditional format by first combining items that occur at the same time into an item set and then sorting these item sets by time [4]. Moreover, a sequence in the new format itself is capable of capturing some of the most important and popular sequences, such as Web-logs, DNA sequences, and documents [7].

3 A Novel Weighting Technique Based on Time-Intervals

For mining sequential patterns over a time-interval sequence data stream, the weight of a sequence in the data stream can be computed from the generation times of data elements in the sequence, which means the relative importance of the sequence in the sequence data stream. It is called the time-interval weight of the sequence.

To get the time-interval weight of a sequence in a sequence data stream, first the time-intervals in the sequence are found from the time stamps of items in the sequence. For a sequence $S = \langle s_1, s_2, \dots, s_l \rangle$ having its time stamp list $TSL(S) = \langle t_1, t_2, \dots, t_l \rangle$ in a sequence data stream, there exist $\frac{l \times (l-1)}{2}$ pairs of items

in the sequence because it consists of l items, and the *time-interval* between two items s_i and $s_j(1 \leq i < j \leq l)$ in the sequence, i.e., TI_{ij} , is defined as follows:

$$TI_{ij} = t_j - t_i$$

The time-interval between a pair of items is a positive value with no limitation. Therefore, to fairly enumerate the time-intervals of different pairs of items in a sequence data stream, they need to be normalized. For this purpose, the time-interval weight for each pair of items in a sequence is found on the basis of its time-interval, and defined as in Definition 1.

Definition 1 [*Time-interval weight*] Let u ($u > 0$) be the size of unit time and δ ($0 < \delta < 1$) be a base number to determine the amount of weight reduction per unit time u , for a sequence $S = \langle s_1, s_2, \dots, s_l \rangle$ and its time stamp list $TSL(S) = \langle t_1, t_2, \dots, t_l \rangle$, the time-interval weight of the time-interval TI_{ij} between two items s_i and s_j ($1 \leq i < j \leq l$), i.e., $w(TI_{ij})$, is defined as follows:

$$w(TI_{ij}) = \delta^{\lceil \frac{t_j}{u} \rceil} = \delta^{\lceil \frac{t_j - t_i}{u} \rceil}$$

The time-interval weight of a sequence is computed from the time-intervals of pairs of items in the sequence. For a sequence $S = \langle s_1, s_2, \dots, s_l \rangle$ and its time stamp list $TSL(S) = \langle t_1, t_2, \dots, t_l \rangle$, the time-interval weight of the sequence is found as in Definition 2 considering the time-intervals in the sequence.

Definition 2 [*Time-interval weight of a sequence*] For a sequence $S = \langle s_1, s_2, \dots, s_l \rangle$ and its time stamp list $TSL(S) = \langle t_1, t_2, \dots, t_l \rangle$, the time-interval weight of the sequence, i.e., $W(S)$, is defined as follows:

$$W(S) = \begin{cases} \frac{1}{N} \sum_{i=1}^{|S|-1} \sum_{j=i+1}^{|S|} w(TI_{ij}), & \text{where } N = \frac{|S|(|S|-1)}{2} \ (l \geq 2) \\ 1 & (l = 1) \end{cases}$$

The sequential pattern evaluation by support has been generally based on simple counting. Contrary to the classical sequential pattern mining, however, this paper proposes a novel interesting sequential pattern of a *time-interval weighted sequential pattern* which is based on the *time-interval weighted support* of a sequential pattern. The time-interval weighted support of a sequential pattern in a sequence data stream is found by using a time-interval weight of a sequence described in Sect. 3.1. For a sequence data stream $TiDS_k$ consisting k sequences, the *time-interval weighted support of a sequential pattern* X in the sequence data stream, i.e., $TW-Supp(X)$, is defined as follows:

$$TW - Supp(X) = \frac{\sum_{S: (X \subseteq S) \wedge (S \in TiDS_k)} W(S)}{\sum_{S: S \in TiDS_k} W(S)}$$

Accordingly, a novel interesting sequential pattern of a time-interval weighted sequential pattern can be defined. Given a support threshold $minSupp$ ($0 < minSupp \leq 1$), a sequential pattern X is a *time-interval weighted sequential pattern* if $TW-Supp(X)$ is no less than the threshold, i.e., $TW-Supp(X) \geq minSupp$.

4 Experimental Results

To evaluate the effectiveness and efficiency of the proposed time-interval weighting approach, a Time-interval Weighted sequential pattern mining over a sequence Data Stream (*TWDS*) method is used. It is based on the eISeq [6] method, and also has an additional operation to get the time-interval weight of a sequence in a sequence data stream from the time-intervals of data elements in the sequence.

For the experiments to evaluate the effectiveness and efficiency of the proposed method, two data sets *SDS_IM* and *SDS_AB* are used in this paper. Each data set is derived from a corresponding base data set generated by the IBM data generator [8], and the number of different items in each data set is 1000. The base data sets generated by the IBM data generator do not have any generation time information. Therefore, to use the data sets in the experiments for the proposed *TWDS* method, a corresponding generation time has to be assigned to each data element in the data sets. For this purpose, several approaches such as the approach using a probability distribution function and that using a randomization function can be considered. However, there is little relationship between the type of the approach and the performance of the proposed method, and the randomization function approach was used in this paper. For data set *SDS_IM*, the difference in generation time between two successive data elements in a sequence is in the range of 0–1000 ms. The data set *SDS_AB* is composed of two consecutive subparts *part_A* and *part_B*. *Part_A* is a set of sequences generated by a set of items *set_A*, and *part_B* is a set of sequences generated by a set of items *set_B*. The two subparts are generated by the same method described in [8], but there is no common item between *set_A* and *set_B*. The time-interval between two successive data elements in a sequence is in the range of 0–1000 ms in *part_A*, while it is in the range of 2000–3000 ms in *part_B*. That is, the sequences in *part_B* have relatively larger time-intervals than those in *part_A*.

Figures 1 and 2 show the number of sequential patterns for the data set *SDS_IM* to demonstrate the effectiveness of time-interval weighted sequential patterns. In this experiment, a support threshold was set to 0.001. The series of generated sequences is divided into five intervals, each of which consists of 200000 sequences. Figure 1 shows the number of sequential patterns in function of δ for each interval. The line named *NoWeight* shows the case of $\delta = 1.0$ which denotes the number of sequential patterns found by the eISeq method. In this case, all the sequences in a sequence data stream have the same weight regardless of the time-intervals of data elements in each sequence, so it denotes the number of sequential patterns found in mining sequential patterns based on simple support counting.

Fig. 1 Number of patterns in function of δ ($u = 500$, $minSupp = 0.001$)

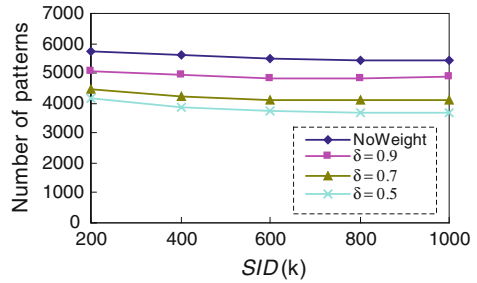
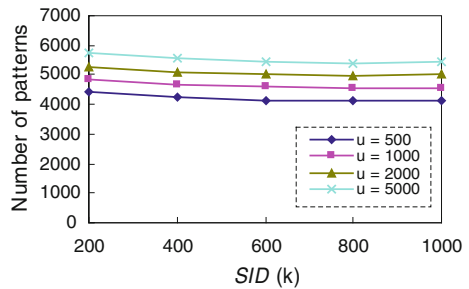


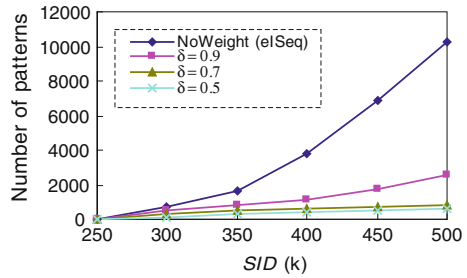
Fig. 2 Number of patterns in function of u ($\delta = 0.7$, $minSupp = 0.001$)



Among the sequential patterns found in mining sequential patterns based on simple support counting, several sequential patterns with relatively large time-intervals were not found in a resulting set found by the *TWDS* method. Therefore, the number of sequential patterns found in the case of $\delta < 1$ is less than that in the case of *NoWeight*. Moreover, the number of patterns decreases as the value of δ becomes smaller. Figure 2 shows the number of sequential patterns in function of u . Similarly in the case of δ in Fig. 1, the number of patterns decreases as the value of u becomes smaller because it is more sensitive to the increase of a time-interval as the values of δ or u become smaller.

To verify the adaptability of the *TWDS* method for the change of time-intervals in a sequence data stream, the data set *SDS_AB* is used which can simulate the change of time-intervals in a sequence data stream over time. In this experiment, the values of *minSupp* and u are set to 0.0005 and 500 ms, respectively. The series of generated sequences in *SDS_AB* is divided into 10 intervals, each of which consists of 50000 sequences. Figure 3 shows the number of sequential patterns derived from *part_B* of the data set whose time-intervals between two successive data elements are in the range of 2000–3000 ms, and it shows last five intervals. As shown in this figure, in the line named *NoWeight* which denotes the number of sequential patterns found by the eISeq method, the number of sequential patterns derived from *part_B* greatly increases as the sequences are continuously generated despite the large time-intervals in *part_B*. However, in the other cases when the values of δ are less than 1.0, the number of sequential patterns derived from *part_B* is much less than that in the case of $\delta = 1.0$, even though they are increased as the sequences are continuously generated. The sequences in *part_B*

Fig. 3 Number of patterns derived from *part_B* of *SDS_AB* ($u = 500$, $minSupp = 0.0005$)



have relatively large time-intervals between two successive data elements; therefore, the sequential patterns appearing in the sequences have a relatively smaller time-interval weighted support, and many of them cannot be found in the resulting set of time-interval weighted sequential patterns.

5 Conclusions

For a sequence or a sequential pattern, the generation times and time-intervals are as important as the generation order of data elements. In sequential pattern mining, therefore, the time-interval information of data elements can help to get more valuable sequential patterns. To obtain more valuable sequential patterns, this paper analyzed the weight of a sequence based on the time-intervals between its data elements, differentiating the importance, i.e., the interestingness, of a sequence as well as that of a sequential pattern. Through this mechanism, more interesting sequential patterns can be selectively found in mining sequence data stream.

To develop a novel interesting sequential pattern of a time-interval weighted sequential pattern for mining sequence data streams, this paper presented a new technique to get the weight of a sequence in a sequence data stream. The weight is computed from the time-intervals of the items in the sequence. After defining the novel interesting sequential pattern of a time-interval weighted sequential pattern based on the weight of a sequence, a new framework to find the patterns in a sequence data stream was presented. In addition, a mining method for finding time-interval weighted sequential patterns over a sequence data stream was developed which can find its up-to-date mining result in a short time with a small memory over the sequence data stream.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No. 2012R1A1B4000651)

References

1. Lo S (2005) Binary prediction based on weighted sequential mining method. In: Proceedings of the 2005 international conference on web intelligence, 755–761
2. Yun U (2007) WIS: Weighted interesting sequential pattern mining with a similar level of support and/or weight. *ETRI J* 29:336–352
3. Yun U (2008) A new framework for detecting weighted sequential patterns in large sequence databases. *Knowl-Based Syst* 21:110–122
4. Chen Y-L, Chiang, M-C, Ko M-T (2005) Discovering fuzzy time-interval sequential patterns in sequence databases. *IEEE Trans Syst Man Cybern-Part B: Cybern* 35:959–972
5. Chen Y-L, Huang TC-H (2003) Discovering time-interval sequential patterns in sequence databases. *Expert Syst Appl* 25:343–354
6. Chang JH, Lee WS (2005) Efficient mining method for retrieving sequential patterns over online data streams. *J Inf Sci* 31:420–432
7. Ji X, Bailey J, Dong G (2007) Mining minimal distinguishing subsequence patterns with gap constraints. *Knowl Inf Syst* 11:259–296
8. Agrawal R, Srikant R (1995) Mining sequential patterns. In: Proceedings of the 1995 international conference on data engineering, 3–14

Analyzing Efficient Algorithms of Frequent Pattern Mining

Unil Yun, Gangin Lee and Sung-Jin Kim

Abstract Frequent pattern mining has been playing an important role for analyzing data in various fields such as medical treatment, biology, finance, networks, and so on. Since Apriori algorithm was proposed, frequent pattern mining has rapidly developed due to active research activities, and numerous mining algorithms have been proposed, such as FP-growth, FP-growth*, LCM, AFORT, and MAFIA. In this paper, we analyze and compare a variety of frequent pattern mining approaches, and discuss advantages and disadvantages of their algorithms. For the comparison, we evaluate mining performance for each algorithm using real datasets. In addition, we also experiment scalability for the algorithms to analyze their characteristics exactly. In the experimental results, we can know that LCM guarantees the fastest runtime performance, and FP-growth* and AFORT show the most efficient memory usage. Using the characteristics analyzed from this paper, we can select and utilize the most appropriate algorithm with respect to numerous databases in the real world.

Keywords Frequent pattern mining · Performance analysis · Data mining

U. Yun · G. Lee · S.-J. Kim (✉)
Department of Computer Engineering, Chungbuk National University,
410 Gaesin-dong, Heungdeok-gu, Cheongju, Republic of Korea
e-mail: ksj@chungbuk.ac.kr

U. Yun
e-mail: yunei@chungbuk.ac.kr

G. Lee
e-mail: abcnarak@chungbuk.ac.kr

1 Introduction

Frequent pattern mining [3] is to find interesting patterns which occur in any database frequently, and the patterns obtained from mining process are used as important information that represents database's characteristics and finds hidden data from the database. Algorithms for mining frequent patterns have been studied actively since Apriori and FP-growth were proposed, and thereafter various techniques and methods have been researched so far. FP-growth stores a database as a compressed tree structure and mines frequent patterns by traversing the tree. In contrast to Apriori which needs multiple scans, FP-growth requires only two scans for mining patterns, and therefore the approach is more efficient than Apriori. After FP-growth, numerous techniques have been proposed, such as improving search speed for trees, reducing the number of local tree generations, sorting tree order dynamically, and so on. In this paper, we thoroughly analyze and compare outstanding and famous algorithms, FP-growth [4], FP-growth* [2], MAFIA [1], LCM [6], and AFORT [5], and evaluate mining performance for each algorithm in terms of runtime, memory usage, and scalability. The remainder in this paper is organized as follows: We analyze the target algorithms in Sect. 2 and compare them in Sect. 3. In Sect. 4, we experiment their runtime, memory usage, and scalability using real and synthetic datasets, and finally conclude this paper in Sect. 5.

2 Frequent Pattern Mining Algorithm

2.1 FP-Growth

FP-growth [4] conducts mining operations by scanning a database twice, and its procedure is performed as shown in Fig. 1. These steps are as follows. First, a database is scanned once, where the algorithm computes supports for each item in the database, and the items are sorted in their support descending order. In here, if there are certain items that do not satisfy a given minimum support, they are pruned according to Anti-monotone property. Second, the database is scanned again, where each transaction is sorted in the defined order and infrequent items are removed. Then, the sorted transactions are inserted into a global FP-tree starting from the item with the lowest support. After inserting all of the transactions in the same way, FP-tree is completed.

FP-tree consists of a header table and a tree structure, where they are connected as link-nodes and nodes with the same item are also concatenated through the link-nodes. After the tree construction, FP-growth performs mining operations selecting the bottom item in the header table first. Regarding the selected item, the algorithm confirms all of the node locations for the item through the link-node information, and searches the tree from the nodes with the item to tree's root, where the

<p>Algorithm: FP-growth</p> <p>Procedure <i>FP-growth</i>(<i>Tree</i>, <i>a</i>)</p> <p>Begin</p> <ol style="list-style-type: none"> 1. if <i>Tree</i> contains a single prefix path 2. let <i>P</i> be the single prefix-path part of <i>Tree</i>; 3. let <i>Q</i> be the multipath part with the top branching node replaced by a null root 4. for each combination (denoted as β) of the nodes in the path <i>P</i> 5. generate pattern $\beta \cup a$ with support = minimum support of nodes in β 6. let <i>freq pattern set</i>(<i>P</i>) be the set of patterns so generated 7. else let <i>Q</i> be <i>Tree</i>; 8. for each item <i>a</i>, in <i>Q</i> 9. generate pattern $\beta = a, \cup a$ with support = <i>a</i>.support 10. construct β's conditional pattern-base and then β's conditional FP-tree <i>Treeβ</i> 11. if <i>Treeβ</i> = \emptyset 12. then call <i>FP-growth</i>(<i>Treeβ</i>, β) 13. let <i>freq pattern set</i>(<i>Q</i>) be the set of patterns so generated 14. return(<i>freq pattern set</i>(<i>P</i>) \cup <i>freq pattern set</i>(<i>Q</i>) \cup (<i>freq pattern set</i>(<i>P</i>) \times <i>freq pattern set</i>(<i>Q</i>))) <p>End Procedure</p>
--

Fig. 1 Algorithm for FP-growth

searched paths become a conditional database. After that, using the conditional database, FP-growth generates a conditional FP-tree containing the selected item ‘i’ (i.e. i’s conditional FP-tree). In the tree, one item is selected again and the algorithm iterates expansion steps until a single-path is generated. If a single-path is discovered, the expansion is stopped and the algorithm extracts frequent patterns combining the items selected so far with all of the items in the current conditional FP-tree.

2.2 FP-Growth*

FP-growth* [2] is a frequent pattern mining algorithm based on FP-tree, and the detailed steps are presented in Fig. 2. The basic framework of FP-growth* follows FP-tree and FP-growth, but there is a difference. In contrast to FP-growth with 2 database scans, FP-growth* reduces the number of scans using a special array structure, FP-array when generating conditional FP-trees. FP-array is a two-dimensional array which can store item supports. For example, when we express the support of a pattern ‘ab’, FP-array stores the support in ‘a’-low ‘b’-column. This array is created in the second database scan. To generate a conditional FP-tree, FP-growth* confirms the corresponding FP-array in advance before scanning the current FP-tree. Through the confirmation, we can know all of the support information without the FP-tree scan. Based on the information, the corresponding conditional FP-tree is constructed by scanning the FP-tree only one time. If there are no items with supports greater than a minimum support, the algorithm stops

Algorithm: FP-growth*
Procedure <i>FP-growth</i> *(<i>T</i>)
Begin
1. <i>if</i> <i>T</i> only contains a single branch <i>B</i>
2. <i>for each</i> subset <i>Y</i> of the set of items in <i>B</i>
3. output itemset <i>Y</i> ∪ <i>T.base</i> with count = smallest count of nodes in <i>Y</i>
4. <i>else for each</i> <i>i</i> in <i>T.header</i> <i>do begin</i>
5. output <i>Y</i> = <i>T.base</i> ∪ { <i>i</i> } with <i>i.count</i>
6. <i>if</i> <i>T.FP-array</i> is defined
7. construct a new header table for <i>Y</i> 's FP-tree from <i>T.FP-array</i>
8. <i>else</i> construct a new header table from <i>T</i>
9. construct <i>Y</i> 's conditional FP-tree <i>T_Y</i> and possibly its FP-array <i>A_Y</i>
10. <i>if</i> <i>T_Y</i> ≠ ∅
11. call <i>FPgrowth</i> *(<i>T_Y</i>)
End Procedure

Fig. 2 Algorithm for FP-growth*

and terminates the current mining step. On the other hand, if there are 3 or more items with supports satisfying the minimum support, it scans the current FP-tree and generates a corresponding conditional FP-tree.

2.3 MAFIA

MAFIA [1] is a frequent pattern mining method based on BITMAP, and its algorithm is shown in Fig. 3. MAFIA conducts mining operations in the following sequence. First, the algorithm scans a database and then generates lattice tree in lexicographical order, where the lattice tree only has item names. MAFIA performs pattern expansions selecting upper nodes near to a root in the lattice tree

Algorithm: MAFIA
Procedure MAFIA(Current node <i>C</i> , <i>FI</i> , Boolean <i>isHUT</i>)
Begin
1. <i>HUT</i> = <i>C.head</i> ∪ <i>C.tail</i>
2. <i>if</i> <i>HUT</i> is in <i>FI</i>
3. Stop searching and return
4. Count all children, use PEP to trim the tail, and reorder by increasing support
5. <i>for each</i> item <i>i</i> in <i>C.trimmed_tail</i>
6. <i>isHUT</i> = whether <i>i</i> is the first item in the tail
7. <i>C_n</i> = <i>C</i> ∪ { <i>i</i> }
8. MAFIA(<i>C_n</i> , <i>FI</i> , <i>isHUT</i>)
End Procedure

Fig. 3 Algorithm for MAFIA

first, where the algorithm checks item supports using BITMAP to know whether any expanded pattern is really frequent one. BITMAP contains information that stores a database as a two-dimensional array structure. Assuming that n is the number of frequent items, and k is the number of transactions, then BITMAP generates storage space by $n \times k$. Using the structure, MAFIA can perform mining steps effectively since it does not traverse trees directly but calculates item supports more quickly through the AND operation of BITMAP.

2.4 AFOPT

Figure 4 shows AFOPT algorithm [5]. AFOPT sorts items from a database in support ascending order after scanning the database, and then stores each transaction depending on the sorted order. That is, any item with a lower support is stored in the location near to a root. Assuming that ‘ a ’, ‘ b ’, and ‘ c ’ are certain items with supports ($a < b < c$), then they are inserted in a tree in the sequence, ‘ a ’, ‘ b ’, and ‘ c ’, where the item, ‘ a ’ becomes a child node of the root and a ’s child tree, ‘ b , c ’ becomes a conditional database. Then, the algorithm combines the prefix, ‘ a ’ with the items in child tree, ‘ b , c ’ to mine patterns. Thereafter, AFOPT conducts the next step considering ‘ b ’ as a prefix, and then ‘ b ’ and b ’s child tree, ‘ c ’ are combined. In the next item, ‘ c ’, there is no item corresponding to c ’s child. Therefore, ‘ c ’ is only mined without any combination. Overall mining procedure of the algorithm is as follows.

In the first scan, AFOPT sorts items in a database in support ascending order, and makes a sorted list. In the second scan, this sorts transactions in the database, where infrequent items for each transaction are eliminated. The sorted transactions are inserted in a tree as follows: (1) the algorithm finds the location of the list

<p>Algorithm: AFOPT</p> <pre> Procedure <i>AFOPT</i> Begin 1. <i>if</i> there is only one branch in tree root then 2. Output patterns and return; 3. <i>for all</i> children c of root 4. Traversal subtree rooted at c and find the set of frequent items F^c, sort the items in F^c in ascending order of their frequencies 5. <i>if</i> $F^c > 1$ then 6. Traversal subtree rooted at c and build a new prefix tree <i>newroot</i> which contains only the items in F^c 7. <i>AFOPT</i>(<i>newroot</i>, <i>min_sup</i>); 8. <i>for all</i> children <i>subroot</i> of c 9. <i>sibroot</i> = the right sibling of c whose item equal to <i>subroot.item</i>; 10. Merge(<i>sibroot</i>, <i>subroot</i>); End Procedure </pre>
--

Fig. 4 Algorithm for AFOPT

which is matched with the left-most item with the lowest support in the current transaction. (2) The sorted transaction is added in the found location. After iterating these steps regarding all of the transactions, AFOP starts mining process using the completed list. The mining sequence starts from the left-most location, and all of the paths in the selected item of the list become a conditional database. Then, the algorithm constructs a corresponding conditional tree in the basis of the conditional database.

2.5 LCM

The next algorithm, LCM [6], as shown in Fig. 5, is *Back Tracking* based frequent pattern mining algorithm. When there is any frequent pattern, P , LCM adds any item, 'e' to P , and if the support of the combined pattern is still frequent, the algorithm expands the pattern again. In addition, LCM uses *database reduction* technique to check item supports more rapidly, where the technique removes infrequent items for each transaction and merges the same transactions if they exist, and then stores the reduced database into an array structure similar to FP-tree. Moreover, LCM utilizes *hypercube decomposition* and *occurrence deliver* techniques in order to decrease the number of item combinations for mining frequent patterns.

3 Analyzing Frequent Pattern Mining Methods

In this section, we compare and analyze the advantages and disadvantages for each frequent pattern mining method. FP-growth algorithm compresses a database using FP-tree structure, and FP-growth* and LCM is also based on the structure though LCM has a somewhat different form. Especially, FP-growth* conducts mining operations similar to FP-growth since it is an algorithm that improves the original FP-growth algorithm. In FP-growth*, the proposed structure, FP-array is used to increase efficiency for tree searches. Generally, tree-based mining algorithms consume lots of times in traversing trees. Therefore, if FP-array, which can reduce

<p>Algorithm: LCM Procedure $LCM(\text{representative } X, \text{itemset } S, \text{item } i)$ Begin 1. output all item sets $R, X \subseteq R \subseteq X \cup S$ 2. for each $j > i, j \notin X \cup S$ 3. if $X \cup \{j\}$ is frequent then 4. call $LCM(X \cup \{j\}, S \cup (X \setminus (X \cup \{j\})), j)$ End Procedure</p>

Fig. 5 Algorithm for LCM

the number of tree traversals, is applied in mining process, mining performance is surely increased. MAFIA also decreases the number of searching trees using its special structure, BITMAP. However, the two algorithms using the additional data structures have no choice but to require more memory space.

FP-growth* needs comparatively small storage space since FP-array only stores items and their supports. As a result, FP-growth* can prevent constructing useless conditional FP-trees in advance, thereby showing memory usage better than the original one in the end. However, in MAFIA, it stores all of the information for a database in BITMAP, therefore consuming enormous memory though the algorithm can quickly compute required patterns' supports. In contrast to FP-growth, LCM conducts pattern expansions through *Back Tracking* technique, and uses the outstanding data compression technique in FP-tree. Moreover, since LCM does not generate any local tree in expanding patterns, the algorithm spends less memory compared to FP-growth. AFOPT mines frequent patterns in the similar way of FP-growth though it uses the special list and tree structures different from FP-tree. In AFOPT, its tree has a property that any child node of the current node has always a greater support value. Thus, the algorithm has an advantage that generates smaller trees compared with FP-trees.

4 Performance Evaluation

In this section, we evaluate runtime performance with real datasets. The details of these datasets are denoted as shown in Table 1.

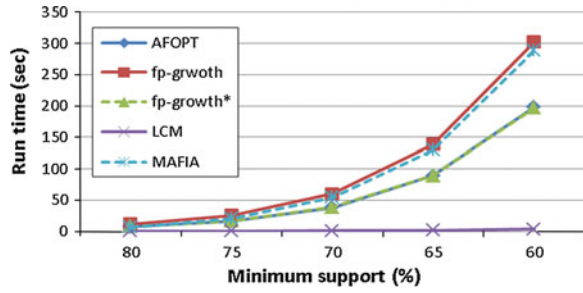
Connect dataset contains connection information in online networks, and Retail is a dataset for retail's product sales. These two datasets can be downloaded in <http://fimi.ua.ac.be/data/>. Chain_store dataset is composed of the product sales data from a major chain in California. The dataset is available at <http://cucis.ece.northwestern.edu/projects/DMS/MineBench.html>.

We experiment runtime performance for FP-growth, FP-growth*, MAFIA, AFOPT, and LCM in this section. For the comparison, we evaluate execution times for each algorithm as decreasing a minimum support. Figure 6 is the result for Connect dataset, where the minimum supports is changed from 80 to 60 %. Connect is dense, and its transaction has similar forms. As shown in the figure, MAFIA and FP-growth shows similarly poor performance since BITMAP in MAFIA does not reflect its advantage in the dataset. In contrast, we can observe

Table 1 Dataset characteristics

Dataset	#Items	Avg. Length	#Transactions
Connect	129	43	67557
Retail	16469	10.3	88162
Chain_store	46086	7.2	1112949

Fig. 6 Runtime result of connect



that AFOPT and FP-growth* has better runtimes compared to FP-growth due to their own enhancement techniques.

LCM guarantees the best performance in all of the cases since the algorithm uses the different mining approach. In Fig. 7, the graphs present each runtime result for Retail dataset. A range of the minimum support is set from 0.025 to 0.005 %. Retail is sparse and has the great number of items and large deviations in terms of transaction lengths. MAFIA also the worst runtime result as in Fig. 6. Since MAFIA generates BITMAP regarding the enormous number of items and transactions in the dataset, its runtime has no choice but to be slow. FP-growth* and AFOPT has faster runtimes compared to FP-growth, and LCM still shows the fastest performance. Next, we compare runtimes for Chain_store with a range of the minimum support from 0.05 to 0.01 %. The dataset has a sparse feature and the great number of items and transactions.

In Fig. 8, it is observed that LCM invariably presents the best result and AFOPT also shows good performance though it is slower than LCM. Moreover, these two algorithms represent stable runtime values regardless of the minimum support. In the test, MAFIA was excluded since this algorithm spent more than 2 GB memory and therefore was not executed correctly in our experimental environment. Furthermore, its runtime was also slowest among all of the algorithms.

Fig. 7 Runtime result of retail

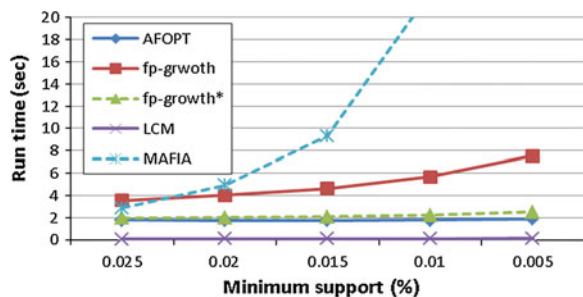
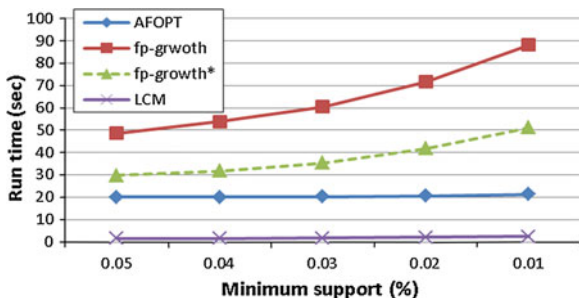


Fig. 8 Runtime result of chain_store



5 Conclusions

Through frequent pattern mining, we can obtain interesting and useful information which is hidden in various databases. In this paper, we compared and analyzed a variety of frequent pattern mining algorithms, and evaluated their mining performances in terms of runtime, memory usage, and scalability to discover characteristics for each algorithm. In the various experiments, we could know advantages and disadvantages for each algorithm, where LCM guaranteed the fastest runtime speed while it spent lots of memory space, and FP-growth* and AFOPT showed their outstanding performances with respect to memory tests. If we refer to the analyzed characteristics when selecting an algorithm suitable for any database and application environment in the real world, we will choose appropriate one more easily and exactly.

Acknowledgments This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 2012-0003740 and 2012-0000478).

References

1. Burdick D, Calimlim M, Flanick J, Gehrke J, Yiu T (2005) MAFIA: a maximal frequent itemset algorithm. *Trans Knowl Data Eng* 17(11):1490–1503
2. Han J, Cheng H, Xin D, Yan X (2007) Frequent pattern mining: current status and future directions. *Data Min Knowl Disc* 15(1):55–86
3. Han J, Pei J, Yin Y, Mao R (2004) Mining frequent patterns without candidate generation: a frequent pattern tree approach. *Data Min Knowl Disc* 8(1):53–87
4. Grahne G, Zhu J (2005) Fast algorithms for frequent itemset mining using FP-trees. *Trans Knowl Data Eng* 17(10):1347–1362
5. Liu G, Lu H, Lou W, Xu Y, Xu YJ (2004) Efficient mining of frequent patterns using ascending frequency ordered prefix-tree. *Data Min Data Min Knowl Disc* 9(3):249–274
6. Uno T, Kiyomi M, Arimura H (2004) LCM ver. 2: efficient mining algorithms for frequent/closed/maximal itemsets. In: *Workshop on frequent itemset mining implementations—FIMI*

Efficient Isomorphic Decision for Mining Sub Graphs with a Cyclic Form

Gangin Lee and Unil Yun

Abstract Graph mining means a series of processes for finding frequent sub-graphs in graph databases with complex structures. To obtain useful sub-graphs, isomorphic decision is needed since one graph data can contain lots of duplicated patterns. Therefore, we need to consider only patterns without duplications. However, these operations can cause enormous overheads due to knotty characteristics of graphs, which is called NP-hard problem. In addition, there also exists a problem that exponentially increases the number of unnecessary operations whenever any pattern size grows. In this paper, we propose a method that enhances efficiency of isomorphic decision in cyclic graphs based on a state-of-the-art algorithm, Gaston, which is called Egaston-CS (Efficient gaston for Cyclic-edge and Spanning-tree). In experiments, we compare our algorithm with previous algorithms, and thereby we demonstrate that Egaston-CS outperforms the others in terms of isomorphic decision.

Keywords Sub graph mining · Cyclic graph · Pattern expansion
Graph isomorphic decision

1 Introduction

In various fields of data mining studies, graph mining is one of the most interesting fields in recent years. Graph data structures are suitable for modeling numerous objects in the real world since they freely expand and variously express most of the

G. Lee · U. Yun (✉)

Department of Computer Science, Chungbuk National University, 410, Gaesin-dong, Heungdeok-gu, Heungdeok-gu, Cheongju, Republic of Korea
e-mail: yunei@chungbuk.ac.kr

G. Lee

e-mail: abcnarak@chungbuk.ac.kr

data while the other structures cannot do that. Sub graph data, which iteratively occur in graph database, can be very useful information in terms of data mining [4]. On the other hand, the advantages of graph data mean that data structure can become more complex. Thus, isomorphic decision of each graph is required to find valid patterns of which duplications are eliminated in complex graph data since certain graphs with isomorphism become only one pattern in the end. However, the number of operations for isomorphic decision needs exponential computing times (NP-hard problem) whenever a pattern size is increased. Therefore, reducing this number is one of the most important issues in terms of graph mining performance. In graph mining, pattern growth method [7, 8, 10] is one of the most well-known approaches proposed to reduce the number of unnecessary operations. In this paper, we propose a novel method that efficiently performs mining operations by improving extraction approaches of cyclic graph patterns. Our method is based on a state-of-the-art algorithm, Gaston [7, 8]. However, through the proposed method, we can obtain frequent sub-graph patterns more quickly than Gaston. The remainder of this paper is organized as follows. We introduce related work for our method in Sect. 2 and describe the details of the proposed algorithm in Sect. 3. We report experimental results by comparing Egaston-CS with the previous algorithm in Sect. 4 and finally conclude this paper.

2 Related Work

Graph mining methods applying Apriori have been studied in graph mining fields at first. Thereafter, methods with pattern growth have been researched, such as [5, 8, 9] to improve mining performance. These approaches decide graph isomorphism by defining graph data as canonical forms using Depth First Search (DFS). Thus, they could avoid the problem of pattern duplications effectively. In addition, various graph mining approaches are being actively studied such as [12] for mining approximate graphs, [13] applying constraints, [6] obtaining maximal sub-graphs, and [1] for stream graph patterns. In [3], the algorithm extracts sub-graphs in directed and weighted graphs. In applications of graph mining, there are the following algorithms, [2] finding graph patterns in time-evolving networks and [11] for mining meaningful information in regard to people interaction.

gSpan [10] is the first algorithm that applies pattern growth method in graph mining area. gSpan was designed to avoid duplicated pattern generations, where the algorithm efficiently mines graph patterns through constraints of extended order. Gaston [7, 8] is a state-of-the-art algorithm among pattern-growth based algorithms. A concept of Gaston algorithm is based on a free tree structure. This algorithm enhanced mining efficiency by avoiding repeated searches of a database using embedding list which stores all real positions of expanded edges on the database according to pattern growth method. Gaston presents good performance in path or free tree graphs while it cannot guarantee mining efficiency in cyclic graph. Accordingly, we propose efficient techniques for mining cyclic graphs in the next section.

3 Deciding Cyclic Graph Isomorphism

The original Gaston algorithm causes inefficient mining operations in extracting cyclic graphs. Therefore, we consider this problem as two aspects to raise the efficiency, where the former is an expanding order of cyclic edges and the latter is enumerating operations of spanning trees.

3.1 Preliminaries

A graph structure consists of vertices and edges. Then, the types of graphs are classified as directed graph or undirected graph according to edge's orientation and divided as simple graph or multi graph according to the number of edges for each vertex. In this paper, we consider only a simple undirected graph to help understand our proposal. However, it is trivial to apply the simple undirected graph to the directed graph and the unlabeled graph.

Definition 1 Given a graph, G , a set of vertices, V , and a set of edges, E , An edge, E is defined as the following formula.

$$E = \{(v1, v2) | v1, v2 \in V \text{ and } v1 \neq v2\} \quad (1)$$

Since edges do not have any direction, an edge $(v1, v2)$ is equal to $(v2, v1)$. Then, we judge whether certain two graphs are isomorphism or not referring to (1).

Definition 2 Let L be a set of labels and l be a function of returning node labels or edge labels. Given two graphs $G1 = (V1, E1, L1)$ and $G2 = (V2, E2, L2)$, if an embedding of $G1$ in $G2$ satisfies an injective function $f(V1) \diamond (V2)$, then this conditions are as follows.

$$\forall v \in V1 \Rightarrow l1(v) = l2(f(v)) \quad (2)$$

$$\forall (v1, v2) \in E1 \Rightarrow (f(v1), f(v2)) \in E2 \text{ and } l1(v1, v2) = l2(f(v1), f(v2)) \quad (3)$$

We define $G1 \subseteq G2$, if two graphs satisfy (2) and (3). In addition, if $G1 \supseteq G2$, $G1$ and $G2$ are considered as isomorphism.

3.2 Techniques for Expanding Cyclic Edges

Original Gaston algorithm utilizes minimum diameter spanning tree to decide whether created patterns are duplicated, where diameter means a path with maximum length in a graph. Here, we found out an improvement for isomorphic decision at first. Gaston algorithm allows expansions only when a spanning tree with a higher priority is not made or when one with a smaller diameter is not

discovered in cyclic graphs. The priority is decided considering what spanning tree is calculated in advance, where the calculating way is based on [8]. In summary, a cyclic pattern is extended only if the following inequality, either (4) or (5) is satisfied. Given a spanning tree in G , s_1 and a spanning tree in G' , s_2 , where G' is an extended graph in G , corresponding inequalities are

$$s_1.\text{priority} \geq s_2.\text{priority} \text{ and } s_1.\text{diameter} = s_2.\text{diameter} \tag{4}$$

$$s_1.\text{diameter} \leq s_2.\text{diameter} \tag{5}$$

where any graph without satisfying above formulas becomes a duplicated graph.

However, since the above manner cannot effectively solve NP-hard problem, it requires enormous and inefficient computing times. An expansion of cyclic edges can make spanning tree with a priority higher than that of spanning tree in the previous graph, but does not eliminate previous spanning tree. Using this previous spanning tree, we can derive the following important property in terms of avoidance of duplicated operations. Assume that all cyclic edges which can be expanded in a cyclic graph, G are $c_0, c_1, c_2, \dots, c_i, \dots, c_j, \dots, c_n$ ($0 > i > j > n$ in priority), and let X_0 be the number of duplicated operations if the expansion starts at c_0 and let X_n be the number of the operations if the expansion starts at c_n . Then the relation, $X_n < X_0$ is satisfied and thus we can reduce needless operations efficiently since starting expansion from c_n eliminates duplicated work.

Figure 1 shows a free tree and cyclic edges which can be expanded in the free tree, c_1, c_2, c_3 . Combinations of the cyclic edges are $\{\{c_1\}, \{c_1, c_2\}, \{c_1, c_2, c_3\}, \{c_1, c_3\}, \{c_2\}, \{c_2, c_3\}, \{c_3\}\}$. Here, a cyclic graph generated as the expansion of cyclic edge, c_2 creates a spanning tree less than the based free tree. Thus, the graph, f is deleted since it generates a useless pattern. c_2 also makes the same spanning tree in a cyclic graph combined with a cyclic edge, c_1 . Therefore, the graph, c is eliminated as above. If the algorithm conducts expansion search as general method, resulting order becomes b, c, d, e, f, g , and h . In this result, c_2 makes a useless spanning tree in the cyclic graph, c , but the algorithm cannot yet know whether c_2 makes the same result in the cyclic graph, f . Therefore, the algorithm have to compute enumeration of spanning tree again in the graph, f . That is, the algorithm must iterate the same work. However, if the graph, f is found earlier than the graph, c , then, an algorithm can omit the expansion operation of the graph, c because a priority of c_2 is lower than that of c_1 . If we in advance extend c_3 with a lowest priority, an order is changed such as h, f, g, b, e, c , and d . Then, the operation of c is deleted.

3.3 Techniques for Enumerating Spanning Trees

The second technique is enumerating spanning trees. To decide isomorphism in a cyclic graph, a spanning tree is needed. Previous algorithm searches for the spanning tree deleting cyclic edges temporarily. In addition, the algorithm only

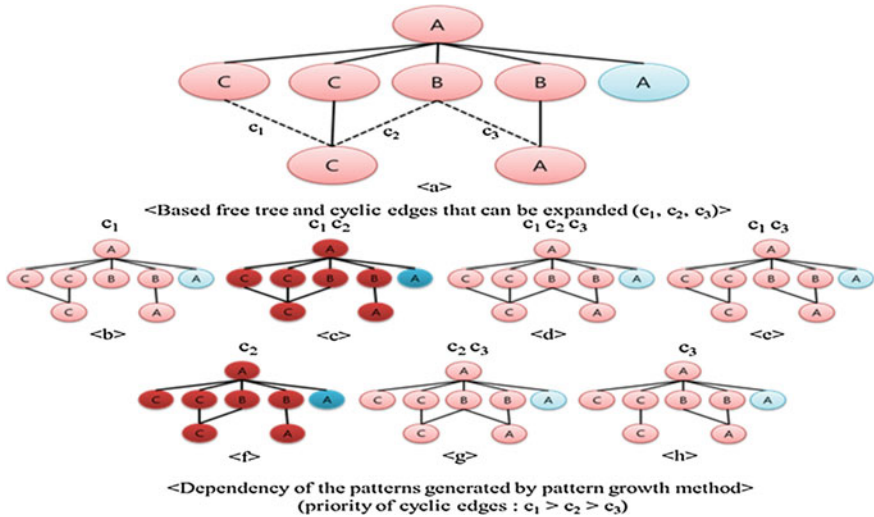


Fig. 1 Cyclic edges which can be expanded and dependency of the patterns

erases edges with a lower order not to make overlapped combinations. This is, from-nodes of next deleting edges must be less than or equal to those of the previous deleted edge, and to-nodes of these must be less than those of the previous deleted edge. However, this manner is very inefficient in terms of the following reason. If there exists a certain circulation which consists of cyclic edges with higher values than the last erased edge during an enumerating operation, this circulation is not deleted in a graph. That is, this implies that the algorithm does not make a tree in the end. Nevertheless, this algorithm will try to make all combinations of lower cyclic edges since it cannot know whether a tree is generated as a result. Eventually, this will not make a tree, and then will go back. For each step of this, an operation deciding all cyclic edges requires computing times proportional to the number of all remaining edges.

In addition, this operation causes enormous wastes since searching for the deleting edges is also proportional to the number of all edges. Figure 2 shows the above-mentioned example. Here, an edge order generated through combinations according to the created order of nodes is denoted as a value in each edge label. If 5th cyclic edge is first deleted in the graph, the circulation consisting of cyclic edges 6, 7, 12, 11, 9, and 10 is not eliminated eventually. Then, although lower values 1, 2, 3, and 4 are deleted depending on the above property, the algorithm cannot create a spanning tree since the above cyclic graph is not deleted. However, if a cyclic edge with a highest value is first deleted, the above case will not occur. The reason is that we can prevent remaining circulations. Therefore, we can solve the problem if cyclic edge, 12 is first deleted. In addition, we can reduce a checking operation of cyclic edges by in advance checking edges of nodes that do not participate in circulation such as v14 in Fig. 2.

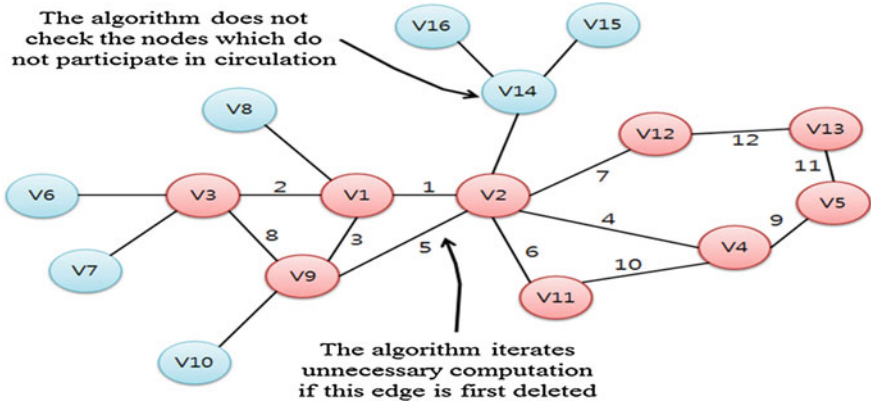


Fig. 2 Priority of cyclic edges for enumerating spanning tree

3.4 Egaston-CS Algorithm

Table 1 represents our algorithm Egaston-CS. Its procedure is as follows. Here, leg means a set which consists of two nodes and an edge (ex. leg = (v_1, e_1, v_2)) is a general leg if v_2 is a new node, and leg = (v_1, v_2, e_1) is a cyclic leg if both v_1 and v_2 are existent nodes). In the first function: extracting cyclic graph, inputting a graph, G and a set of legs, L , the function checks whether each leg is cyclic (line 1–2). If the condition is satisfied, this generates an extended graph, G' using G and L . Then, the function finds a spanning tree with a higher priority utilizing G' (line 4), and then mark true on L if the result of line 4 is true. Then it assigns the next L to current L , go to line 1, and then conduct the operation again. If line 4 is false, the function goes to line 7 and conducts join operations with L and each L' (line 8–9). Thereafter, this sorts L' in priority ascending order to expand a leg with a lowest priority first (line 10). For each sorted legs, L' , then one obtains valid cyclic graph patterns calling this function recursively (line 11–12). Terminating all recursive calls, we can gain a set of cyclic graphs, C as a result if C is not empty (line 13–15). The second function: finding spanning tree is performed when the line 4 is executed in the first function. Receiving a graph G , the function confirms that G is a spanning tree, and then returns false if G has a priority higher than the previous graph of G while it returns true (line 17–19). If G is not a spanning tree, this examines whether nodes and edges in G are cyclic and mark true or false on each node or edge (line 20–22). Then for each node and corresponding edges, the function generates a spanning tree extending each edge and calls this function recursively so as to know that the next expanding operation makes a spanning tree with a higher priority (line 23–26). Thereafter, the function returns true if there exists a corresponding result, or returns false (line 27–29).

Table 1 Egaston-CS algorithm

Egaston-CS	
Input: a graph, G, a set of legs, L	Output: a set of cyclic graphs, C
Extracting cyclic graph(a graph, G, a set of leg, L)	
1. for each leg, l in L do	
2. if l is a cyclic leg do	
3. $G' \leftarrow$ generate cyclic graph using G and l	
4. if finding spanning tree(G') do	
5. l.mark \leftarrow true and l \leftarrow next l	
6. go to the line 1	
7. else for each marked leg, l' in L do	
8. if l' is a cyclic leg & l'.priority \leq l.proirity do	
9. $L' \leftarrow L' \cup$ join l with l'	
10. sort L' in priority ascending order	
11. for each leg, l'' in L' do	
12. $C \leftarrow C \cup$ extracting cyclic graph(G', L')	
13. if C is empty do	
14. return \emptyset	
15. else return C	
Finding spanning tree(a graph G)	
16. if G is a spanning tree do	
17. if G has a higher priority do	
18. return false	
19. else return true	
20. else for each node, v and edge, e in G do	
21. if v is cyclic do v.mark \leftarrow true	
22. if e is cyclic do e.mark \leftarrow true	
23. for each marked node, v in G do	
//starting from a reverse of marked order in G	
24. for each marked edge, e in G do//same as above	
25. $G' \leftarrow$ generate spanning tree using G and e	
26. result \leftarrow finding spanning tree(G')	
27. if result is true do	
28. return result	
29. else return false	

4 Performance Evaluation

4.1 Experimental Environment

In this experiment, to evaluate a degree of improvement, we compare our algorithm with the previous algorithms in terms of the runtime. These algorithms ran with core 2×64 (3 GHz), 2 GB memory, and Opensolarise10, and were written in C++ language. Graph data used in this experiment are PTE and DTP respectively, and they are real datasets. PTE consists of 340 graph transactions while DTP

consists of 422 graph transactions. DTP dataset has cyclic graphs relatively more than those of PTE dataset. Accordingly, we can expect that our algorithm, Egaston-CS generally has a better effect on runtimes in DTP dataset rather than PTE dataset, and the algorithm really guarantees better performance as shown in experimental results though it also presents a good effect in PTE dataset.

4.2 Performance Analysis

For accurate performance evaluation, we compare not only Egaston-CS proposed in this paper with the original Gaston algorithm, but also parts of Egaston-CS, Egaston-C and Egaston-S in addition. Figure 3 represents runtimes of algorithms in PTE dataset. In the Fig. 3, Egaston-CS generally shows the best runtime ratio compared with other algorithms. This result demonstrates that our techniques of isomorphic decision have a good effect in the actual situation.

The next is with respect to performance evaluation in DTP dataset. Figure 4 illustrates that Egaston-CS outperforms the others except that minimum support is 4.30 and 5.00 %, which implies that our algorithm does not always enhance performance in all cases. When minimum support is 4.30 or 5.00 %, the performance of Egaston-C outperforms the others. However, Egaston-CS shows outstanding performance in most cases and is especially effective compared with the others when minimum support is lower such as 3.60 %. Egaston-CS also represents stable performance more than a part of our algorithm, Egaston-C or S. Through the graphs, we can know that Egaston-CS is effective in DTP more than in PTE. The reason is that DTP includes more cyclic graphs than that of PTE and therefore the advantage is reflected in DTP better than PTE, though Egaston-CS also guarantees outstanding performance in PTE.

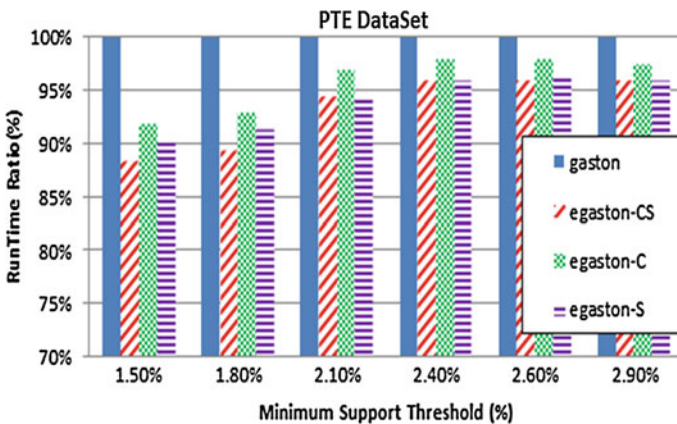


Fig. 3 Runtime result in PTE data set

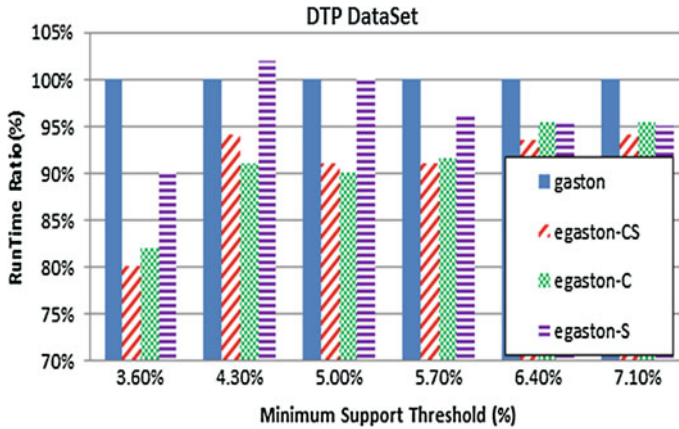


Fig. 4 Runtime result in DTP data set

5 Conclusions

In this paper, we proposed an improved algorithm that can eliminate needless duplications occurring in extraction process of cyclic graph patterns and that can enhance efficiency of deciding graph isomorphism. Then, we objectively analyzed performance of our proposed algorithm through the above performance evaluation. Especially, our Egaston-CS applied important improvements in terms of graph isomorphic decision and demonstrated effects using real data sets. In the future research, in addition to improvement of cyclic graph operation, we will develop novel approaches for join operations or edge expansions in free trees so as to increase operating efficiency of free trees. Moreover, we can improve mining performance by applying our techniques to the other approaches such as maximal pattern mining and noisy pattern mining methods.

Acknowledgments This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 2012-0003740 and 2012-0000478).

References

1. Bifet A, Holmes G, Pfahringer B, Gavaldà R (2011) Mining frequent closed graphs on evolving data streams. In: KDD'11 Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, pp 591–599
2. Bogdanov P, Mongiovi M, Singh AK (2011) Mining heavy subgraphs in time-evolving networks. ICDM, pp 81–90
3. Günemann S, Seidl T (2010) Subgraph mining on directed and weighted graphs. PAKDD 6119:133–146

4. Han J, Kamber M (2005) *Data mining: concepts and techniques*. Morgan Kaufmann, Publishers, San Francisco
5. Lahiri M, Berger TY (2010) Periodic subgraph mining in dynamic networks. *Knowl Inf Syst* 24(3):467–497
6. Lini T, Thomas SR, Valluri KK (2006) MARGIN: Maximal frequent Subgraph mining. *ICDM*, pp 1097–1101
7. Nijssen S, Kok JN (2005) The Gaston tool for frequent subgraph mining. *Electron Notes Theor Comput Sci* 127(1):77–87
8. Nijssen S, Kok JN (2004) A quickstart in frequent structure mining can make a difference. In: *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*, pp 647–652
9. Silva A, Meira W Jr, Zaki MJ (2012) Mining attribute-structure correlated patterns in large attributed graphs. *PVLDB* 5(5):466–477
10. Yan X, Han J (2002) gSpan: graph-based substructure pattern mining. In: *Proceedings of the 2002 IEEE international conference on data mining*, pp 721–724
11. Zhiwen Y, Zhiyong Y, Xingshe Z, Christian B, Yuichi N (2012) Tree-based mining for discovering patterns of human interaction in meetings. *IEEE Trans Knowl Data Eng* 24(4):759–768
12. Zou Z, Li J, Gao H, Zhang S (2010) Mining frequent subgraph patterns from uncertain graph data. *IEEE Trans Knowl Data Eng* 22(9):1203–1218
13. Zhu F, Yan X, Han J, Yu PS (2007) gPrune: a constraint pushing framework for graph pattern mining. In: *Proceeding 2007 Pacific-Asia conference on knowledge discovery and data mining (PAKDD'07)*, pp 388–400

Performance Evaluation of Approximate Pattern Mining Based on Probabilistic and Statistical Techniques

Unil Yun, Gwangbum Pyun and Sung-Jin Kim

Abstract Approximate frequent pattern mining is to find approximate patterns, not exact frequent patterns with tolerable variations for more efficiency. As the size of database increases, much faster mining techniques are needed to deal with huge databases. Moreover, it is more difficult to discover exact results of mining patterns due to inherent noise or data diversity. In these cases, by mining approximate frequent patterns, more efficient mining can be performed in terms of runtime, memory usage and scalability. In this paper, we benchmark efficient algorithms of mining approximate frequent patterns based on statistical and probabilistic methods. We study the characteristics of approximate mining algorithms, and perform performance evaluations of the state of the art approximate mining algorithms. Finally, we analyze the test results for more improvement.

Keywords Approximate frequent pattern mining · Lossy Counting technique · Chernoff technique · Probabilistic technique · Statistical technique · Performance evaluation

U. Yun · G. Pyun · S.-J. Kim (✉)
Department of Computer Engineering, Chungbuk National University,
410 Gaesin-dong, Heungdeok-gu, Cheongju, Republic of Korea
e-mail: ksj@chungbuk.ac.kr

U. Yun
e-mail: yunei@chungbuk.ac.kr

G. Pyun
e-mail: pyunb@chungbuk.ac.kr

1 Introduction

One of data mining techniques, approximate frequent pattern mining [2, 4, 7, 8] is a method for obtaining patterns near to really frequent ones from noisy databases. As the amount of data which have to be serviced has been increased exponentially, sizes of related databases are also rapidly increasing [6]. Approximate frequent pattern mining has been utilized in the various fields such as networks [9], bio analysis [10], and so on. FP-growth [1] is one of the previous frequent pattern mining algorithms. Though the algorithm can mine frequent patterns efficiently regarding general databases, it spends much more execution time and memory mining databases with huge sizes and very complex patterns. Due to the problem, most frequent pattern mining algorithms as well as FP-growth need enormous computer resources, thereby facing difficult situations to service results analyzed through mining process rapidly. The reason is that they use all of the data from databases to derive exact mining results. When there are some noises in databases, results after mining operations may be partially incorrect. However, there are no problems if we utilize the results to decision-making and analysis which do not require precise results. Therefore, we can use approximate frequent pattern mining in this situation.

In case a size of data is enormously large and fast analysis is needed, we can also apply approximate frequent pattern mining, thereby obtaining useful approximate results more quickly. That is, approximate frequent pattern mining is a technique suitable for situations that must analyze frequent patterns rapidly even though the final result is not completely accurate. With the advantages, numerous techniques related to this have been studied, and recent approximate pattern mining algorithms are guaranteeing faster runtime and less memory usage compared to previous ones as well as presenting very useful results which are greatly close to really frequent patterns. In this paper, we compare and analyze various approximate frequent pattern mining algorithms based on probabilistic and statistical techniques, and evaluate mining performance with respect to the algorithms. Thereafter, we finally conclude advantages and disadvantages for the algorithms.

2 Mining Approximate Frequent Patterns Based on Probabilistic and Statistical Techniques

2.1 Lossy Counting-Based Mining Technique

Approximate frequent pattern mining helps deduce and extract patterns which are most similar to really frequent patterns from databases by using statistical techniques. One of the mining methods, Lossy counting [3] decides an appropriate threshold through statistical techniques with a part of any database, where the

algorithm divides a database into *buckets* with each *bucket* size, ω . Then, the size of ω can be denoted as the following equation.

$$\omega = \left\lceil \frac{1}{\varepsilon} \right\rceil$$

where ε means a range of errors. Since the algorithm deduces correct answers through statistical methods, its mining result and performance are changed depending on the error range. Therefore, the higher ε becomes, the farther from correct answers the result is. On the other hand, the lower ε is, the closer to them the result becomes. Assuming that β is the number of *buckets*, we can know β by calculating the total number of transactions in a database and *bucket's* size. Let D be a list storing certain patterns discovered from the *bucket*. Then, D 's structure is denoted as $D = (set, f, \Delta)$, where *set* represents pattern information, f is a frequency (or support) computed as statistical methods, and Δ means an error estimated by f . Mining process of the algorithm is as follows. First, the algorithm divides a database by *bucket's* size, and then finds patterns in the first divided *bucket*. Thereafter, the algorithm decides a threshold using ω . Let $b_{current}$ be a threshold gained by the statistical method, and then $b_{current} = \lceil N/\omega \rceil$, where N means the current *bucket* number. Thus, the threshold value is increased whenever new *buckets* are mined. After setting the threshold, the algorithm computes pattern's support in the *bucket*. Δ is calculated as shown in the following formula.

$$\Delta = \frac{1}{\varepsilon} \log(\varepsilon \times N)$$

Before inserting the calculated f and Δ , pruning operations are performed in advance. If $f + \Delta < b_{current}$, the corresponding pattern is eliminated since the pattern is more likely to become an infrequent pattern according to statistical methods. The above steps are iterated regarding the remaining *buckets*, and then we can obtain a set of approximate frequent patterns. Figure 1 shows an overall mining procedure for Lossy counting algorithm.

2.2 Chernoff-Based Mining Technique

The next algorithm is based on chernoff [5], where the Chernoff is a formula that calculates errors between actual occurrences and results calculated by *Poisson's distribution* with respect to phenomena following *Poisson's distribution*. Chernoff computes probability based on *Bernoulli trial*. Assuming that trials generated by any phenomenon, X are denoted as $o_1, o_2, o_3, \dots, o_n$, then *Bernoulli trial* assigns $o_n = 1$ if o_n is equal to X or $o_n = 0$ if X does not occur. Let \hat{X} be the sum of X 's occurrence, and then the actual occurrence probability of X , x is defined as \hat{X}/n , and therefore chernoff's formula is denoted as follows.

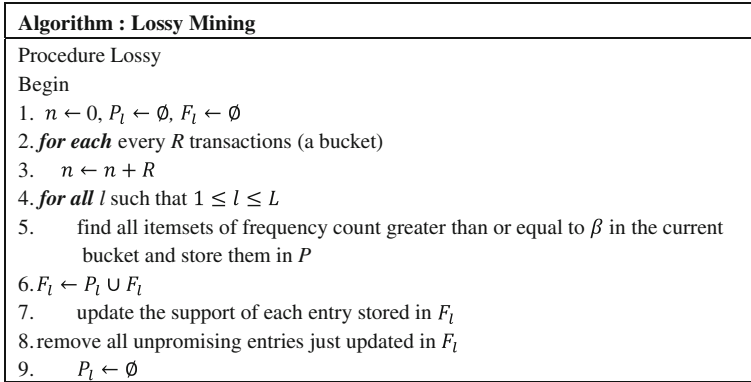


Fig. 1 Algorithm for lossy count mining

$$\text{pr}\{ |\hat{x} - x| \geq xr \} \leq 2e^{-\frac{nxr^2}{2}}$$

As we solve the formula regarding the error, xr , the following modified formula is generated.

$$xr = \sqrt{\frac{2x \ln(\frac{\lambda}{2})}{n}}$$

Using the result of the formula, we can calculate probability that any specific pattern occurs. Then, Chernoff computes the total probability for each pattern through *Poisson's distribution* as mining a part of the transactions, not all of them.

After that, Chernoff computes an error with respect to the calculated probability and then a support near to real pattern's support depending on the computed error. Through the support, we can early know whether any pattern is frequent or not and prune infrequent patterns. Using the above Chernoff formula, we can induce the following formula for calculating patterns' supports.

$$T_l = \frac{2[C_l^{pt} + 4 \ln(\frac{\lambda}{2})]}{p_k}$$

By the formula, we remove approximate infrequent patterns among the results computed from one batch (a part of total transactions). Iterating the same work for all batches, we can obtain the final result containing all of the possible approximate frequent patterns. Figure 2 is an algorithm for mining approximate patterns based on Chernoff.

Algorithm : Chernoff Mining
Procedure Chernoff Begin 1. $n \leftarrow 0, P_l \leftarrow \emptyset, F_l \leftarrow \emptyset$ 2. for each every R transactions 3. $n \leftarrow n + R$ 4. for all l such that $1 \leq l \leq L$ 5. find potential frequent itemset in terms of R in the current batch and store them in P_l 6. $F_l \leftarrow P_l \cup F_l$ 7. update the support of each entry stored in F_l 8. if the current batch is the first batch 9. calculate T_l in terms of n from F_l futher if $ F_l > T_l$ 10. $P_l \leftarrow \emptyset$

Fig. 2 Algorithm for chernoff mining

3 Performance Analysis for Approximate Frequent Pattern Mining

In this section, we compare and analyze performances for each algorithm, where the target algorithms are approximate frequent pattern mining approaches using probabilistic and statistical techniques, chernoff [5] and Lossy counting [3], and a general frequent pattern mining approach, FP-growth [1]. Regarding the algorithms, we evaluate runtime.

Table 1 shows real datasets for evaluating runtime and memory usage for each algorithm. Connect dataset includes connection information gathered from online networks, and Retail dataset consists of product sales data in Belgian retail store. Pumsb is a set of data for a census, and Kosarak dataset contains click stream data collected from hungarian on-line news portal.

These four datasets can be downloaded at <http://fimi.ua.ac.be/data>. Chain_store dataset consists of product sales information derived from a major chain in California. This is available at <http://cucis.ece.northwestern.edu/projects/DMS/MineBench.html>.

Table 1 Real dataset

Dataset	#Items	Avg. Length	#Transactions
Connect	129	43	67557
Retail	16469	10.3	88162
Pumsb	2113	74	49046
Kosarak	41270	8.1	990002
Chain_store	46086	7.2	1112949

3.1 Runtime Test

In these tests, we use real datasets in Table 1, and the results for each algorithm are times after all of the mining processes are completely terminated.

Figure 3 presents the runtime results for Connect dataset with a range of the minimum support from 80 to 60 %, where this dataset is composed of transactions having similar information. In this result, it is observed that Chernoff and Lossy are faster than FP-growth. Especially, Chernoff shows the best runtime since the algorithm conducts mining operations more quickly through pattern probability techniques, while Lossy spends longer runtime since procedures related to *bucket* need relatively long times though it is faster than FP-growth. Moreover, the lower the minimum support is, the larger the gap between Chernoff and Lossy is.

Figure 4 is regarding Retail dataset with the minimum support range from 0.025 to 0.005 %, where Retail has numerous transactions and items compared with the other datasets. In the figure, we can observe that the approximate frequent pattern mining algorithms, Chernoff and Lossy have outstanding performances compared to the general frequent pattern mining algorithm, FP-growth. Especially, FP-growth's runtime sharply increases as the minimum support becomes low, while the others shows stable runtime results, and Chernoff guarantees the best performance in common with Fig. 3.

In Fig. 5 with the minimum support between 0.05 and 0.01 %, Chernoff also presents the fastest runtime due to its pattern probability techniques. Chain_store used in this test has a large size and lots of items. FP-growth represents the worst speed since it mines all of the patterns. Next, we experiment runtimes for Pumsb dataset as decreasing the minimum support from 60 to 40 %, and the results are shown in Fig. 6.

In here, all algorithms show similar increasing rate until the minimum support is 45 %, but thereafter FP-growth represents a drastic change while the other algorithms has constant changes, and Chernoff provides the fastest speed in every case.

Figure 7 presenting the result of Kosark also shows the runtime result similar to Fig. 6 though its graphs have higher slopes compared with those of Fig. 6. In summary, FP-growth requires the longest runtime in all of the cases since the

Fig. 3 Runtime test (Connect)

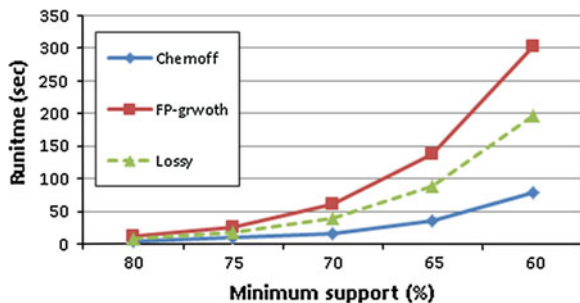


Fig. 4 Runtime test (Retail)

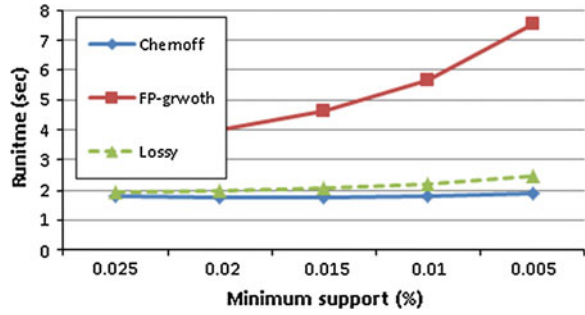


Fig. 5 Runtime test (Chain_store)

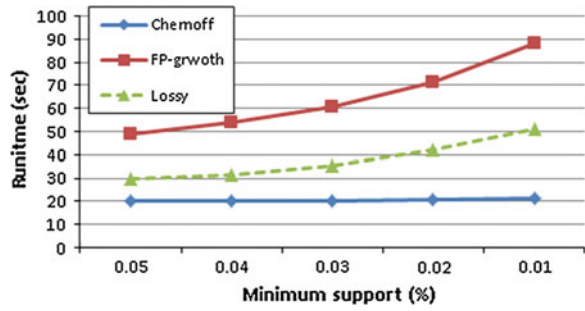


Fig. 6 Runtime test (Pumsb)

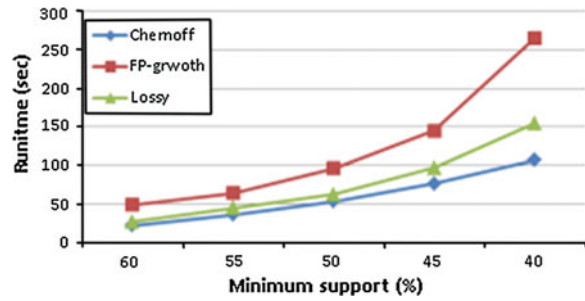
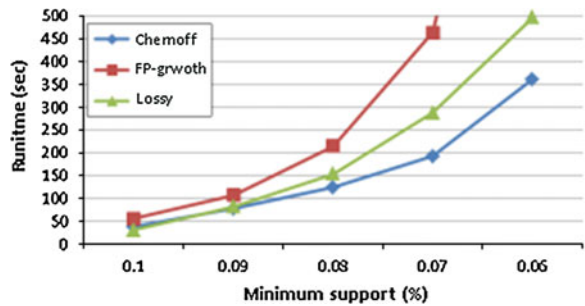


Fig. 7 Runtime test (Kosarak)



algorithm considers all of the frequent patterns. In contrast, Lossy and Chernoff shows outstanding runtime performances, and especially Chernoff has the fastest results in every case. In addition, as a dataset has more transactions and items, the advantage of approximate frequent pattern mining is maximized and Chernoff with probability techniques presents more strong performance.

4 Conclusion

In this paper, we analyzed approximate frequent pattern mining approaches based on probabilistic and statistical techniques and conducted performance evaluation regarding the approximate frequent pattern mining algorithms as comparing the general frequent pattern mining algorithm. For thorough analysis, we performed a variety of experiments, runtime, memory usage, and scalability. In the experiments, approximate pattern mining algorithms showed more outstanding runtime and memory performances compared to the general mining algorithm. In the scalability test, they presented stable runtime and memory increases as the number of transactions and items increased.

That is, approximate pattern mining algorithms guaranteed excellent scalability in terms of both runtime and memory usage. Since real-time systems require rapid mining speed, approximate frequent pattern mining is more suitable than general pattern mining. Furthermore, since sensor networks have to process large amounts of data with small resources, this approximate pattern mining based on probabilistic and statistical techniques is also more appropriate. If the method is appropriately applied in a variety of areas, we expect that this plays an important role in networks, high volume processing, and so on.

Acknowledgments This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 2012-0003740 and 2012-0000478).

References

1. Han J, Pei J, Yin Y, Mao R (2004) Mining frequent patterns without candidate generation: a frequent pattern tree approach. *Data Min Knowl Disc* 8:53–87
2. Wong RC, Fu AW (2006) Mining top-K frequent itemsets from data streams. *Data Min Knowl Discov* 13:193–217
3. Manku G, Motwani R (2002) Approximate frequency counts over data streams. *VLDB*
4. Chi R, Wai A (2006) Mining top-K frequent itemsets from data streams. *Data Min Knowl Discov* 13(2):197–217
5. Zhao Y, Zhang C, Zhang S (2006) Efficient frequent itemsets mining by sampling, advances in intelligent IT. *Active Media Technology*, pp 112–117
6. Han J, Cheng H, Xin D, Yan X (2007) Frequent pattern mining: current status and future directions. *Data Min Knowl Discov (DMKD)* 1.15(1):55–86

7. Yun U, Ryu K (2011) Approximate weight frequent pattern mining with/without noisy environments. *Knowl Based Syst* 24(1):73–82
8. Zhu F, Yan X, Han J, Yu PS (2007) Efficient discovery of frequent approximate sequential patterns. In: *International conference on data mining (ICDM)*, pp 751–756
9. Chen C, Yan X, Zhu F, Han J (2007) gApprox: mining frequent approximate patterns from a massive network. *ICDM*, pp 445–450
10. Wong P, Chan T, Wong MH, Leung K (2012) Predicting approximate protein-DNA binding cores using association rule mining, *ICDE* pp 965–976

Interactive Self-Diagnostic System Using Anatomical 3D Human Body

Sung-Ho Kim and Kyung-Yong Chung

Abstract Recently, due to the rapid increase in the elderly population, the interest in u-healthcare for personal and social needs is increasing. In addition, extensive medical information through various media services is one of the causes for the interest. However, the general public often doesn't have the time to visit a medical authority for u-healthcare in many cases. The absence of a system that can be easily and quickly accessible anytime or anywhere to monitor health is a sad reality, especially in light of the rapid development of IT convergence technology. In this study, we proposed an interactive self-diagnostic system that monitors the human body for u-healthcare. First of all, this study separates the human body of an adult male and female into skin, muscles, and skeleton. And these three categories are then modeled using 3DS MAX. The 3D human body structures are able to be viewed with a 3D viewer of the system. One of the key features of this system is using the picking technique. If you select certain part of the human body in the 3D viewer, the system provides a variety of detailed medical information about the diseases associated with the selected part. The 3D viewer has the advantage of being able to view the structure of the human body realistically and intuitively. Medical information about diseases is comprised of simple and clearly organized data about the causes, symptoms, treatment, prevention, recommended food, and related medical institutions (such as hospitals) that can deal with the disease. If you use our system anytime or anywhere for u-healthcare, it can prevent diseases in advance and provide answers to many questions about disease-related symptoms.

S.-H. Kim · K.-Y. Chung (✉)

School of Computer Information Engineering, Sangji University, 83 Sangjidaegil,
Wonju-si, Gangwon-do, South Korea
e-mail: dragonhci@hanmail.net

S.-H. Kim

e-mail: kimsh1204@sangji.ac.kr

Keywords Self-diagnostic system · Healthcare · 3D anatomy model · Medical information

1 Introduction

Generally, u-healthcare is the diagnosis, treatment, and prevention of disease, illness, injury, and other physical and mental impairment in humans [1]. U-healthcare is delivered by practitioners in medicine, chiropractic, dentistry, nursing, pharmacy, allied health, and other care providers. It refers to the work done in providing primary care, secondary care, and tertiary care, as well as public health. Humans are investing a lot of economic power and time in exercise or eating for u-healthcare. In recent years, public interest in u-healthcare is growing more and more with the release of IT convergence products related to u-healthcare. However, although we have many healthcare-related IT convergence products, as well as easy access to medical information, there still exist many problems such as wrong or insufficient medical information. In addition, there are disadvantages to u-healthcare, such as it being time-consuming to retrieve medical knowledge and information. In particular, humans have an inseparable relationship with disease in almost all parts of the human body. So a realistic and practical system that can provide medical knowledge and information related to the diseases that correspond to each part of the body will be very useful. Therefore, this paper proposes an interactive self-diagnostic system that can be provided with medical information related to disease by directly selecting a specific part after actually viewing a 3D anatomic model of the human body [2].

The rest of this paper is organized as follows. [Section 2](#) describes related works associated with the self-diagnostic system. [Section 3](#) describes the interactive self-diagnostic system using the anatomical 3D human body. The conclusions are given in [Sect. 4](#).

2 Related Works

In recent years, interests in health have been increased according to changes in lifestyle and environment. Also, interest in u-healthcare, which monitors one's health and provides specialized healthcare services whenever and wherever it is needed, has increased. The conventional u-healthcare systems show a lack of extensibility and device dependency, and still have some trouble in supporting customized information based on personal context information [3]. U-healthcare services provide medical and healthcare services continuously and generally for the healthy life of customers through active participation and cooperation in all members employed in industries based on IT convergence, through merging it to

other advanced technologies. In addition, personalization is a service that provides static and dynamic information from customers who have similar personalities in order to satisfy their requirements.

Therefore, u-healthcare service can be defined as a service that provides u-healthcare adapted to users by learning the situation, behavior, tendencies, preferences, and bio-information in a ubiquitous service environment. The previous research of [3, 4] is described in more detail.

In general, a 3D anatomic model of the human body has been mainly used for research and training in the medical field until now. Turbo Squid is a representative company for the production of a 3D anatomic model. They are trying detailed modeling of the skin, muscle, and skeleton of the human body. The previous research of [5, 6] is designed for educational 3D anatomical models. It is shown to students studying medicine. As a result, they are able to get very positive results. In particular, they provided 3D anatomical models to students in the form of Interactive 3D, and the reaction of the students was very good. However, the previous research of [5, 6, 7–9] and all studies associated with 3D anatomical models are intended for educational purposes. There were no research results to provide medical information to the general public via u-healthcare. Therefore, this paper provides an interactive self-diagnostic system by showing a realistic and practical 3D anatomical model to the general public who had a lot of an interest in u-healthcare.

3 Interactive Self-Diagnostic System Using Anatomical 3D Human Body

3.1 System Overview

U-healthcare means a system that can monitor bio-information in real-time using certain devices and mobile equipments in a home network and provide medical examination and treatment whenever and wherever it is needed through linking it to hospitals and doctors [3, 4]. The characteristics of the u-healthcare can be summarized as fast medical services, disease prevention, central processing of bio-data management, distribution of diagnosis, and management of aged, handicapped, and isolated persons. Fast medical services actively deal with bad conditions in patients by monitoring the state of patients [10].

In this paper, we propose a system that can view a 3D anatomic model of the human body using a 3D viewer. And when you select a specific part of the 3D model, you can see simple and clear medical information related to diseases that may involve the selected part. Therefore, database construction associated with the 3D anatomical model of the human body, medical knowledge and information is essential. In this paper, 3D anatomical model data of the human body used to construct the database was limited to adult men and women. And medical

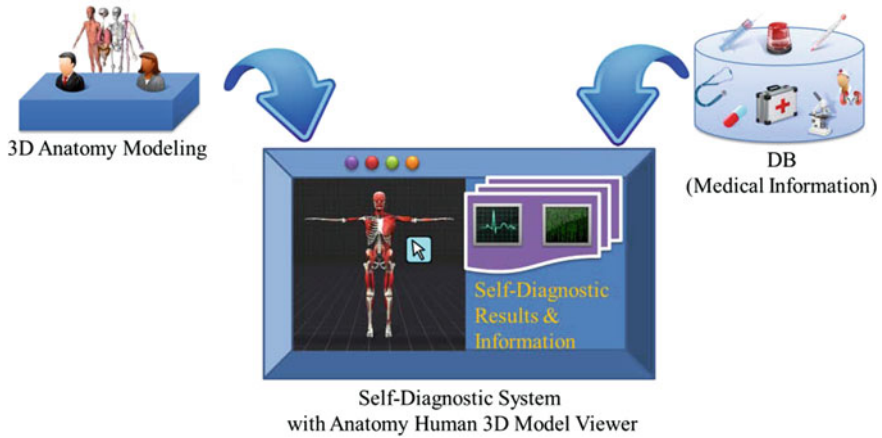


Fig. 1 Diagram for interactive self-diagnostic system

information was limited to a few items related to common diseases that may be frequently experienced at home. Figure 1 shows a schematic for the interactive self-diagnostic system. In the figure, the system can be called after the database is constructed from the 3D anatomical model data and medical knowledge and information. This system has been constructed from a 3D viewer that can view a 3D anatomical model. Also, this system has been constructed from medical knowledge and information about the diseases that correspond to the selected part of the 3D view by a user.

3.2 User Interface

In this paper, we propose a user interface for the interactive self-diagnostic system to view the 3D anatomical model as shown in Fig. 2.

This user interface consists of the 3D anatomical model viewer, controller for a 3D anatomical model viewer, controller for kind of model, gender of the 3D anatomical model, human anatomy part, kind of disease, description of disease, and buttons for reset, about, and OK. The 3D anatomical model viewer is driven in conjunction with the controller. The controller can perform translation, rotation, zoom, and select. In order to select the specific part of the 3D anatomical model, a picking technique is used. A controller for kind of model can select the skin, muscle, viscera, or skeleton. The Human anatomy part lists the name of each part of the body. The user may directly select the desired name from the human anatomy part. If you select a specific part of the 3D anatomical model viewer, the human anatomy part is automatically selected by name with the same location. The kind of disease is a list of the various types of diseases that a user can select.

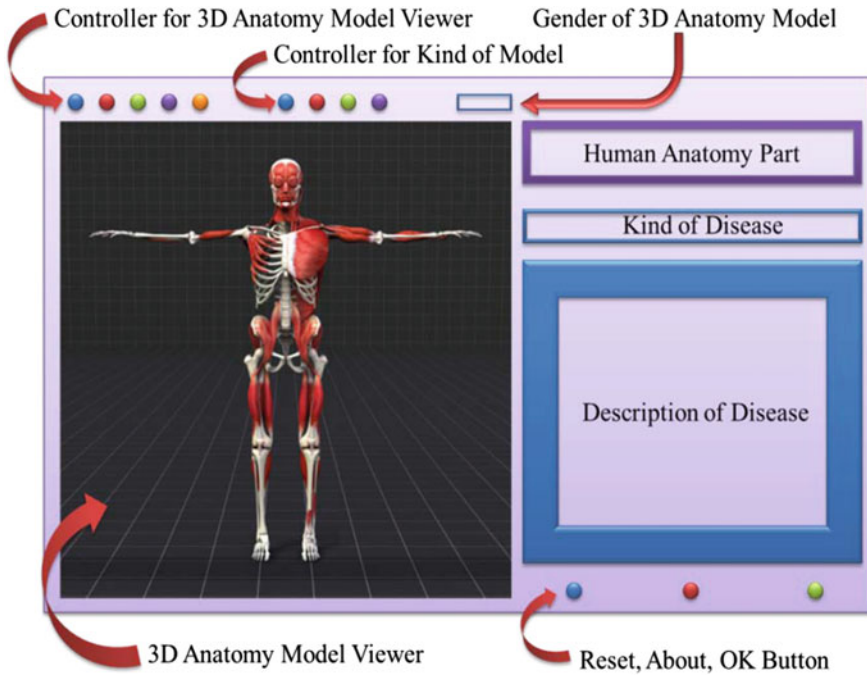


Fig. 2 User interface for interactive self-diagnostic system

The description of disease succinctly describes medical knowledge and information associated with the item selected from the kind of diseases.

3.3 3D Human Anatomy Modeling

3D anatomical model data used in this system was based on the [2]. And it was transformed into the OBJ file format created with 3DS MAX. And OBJ files

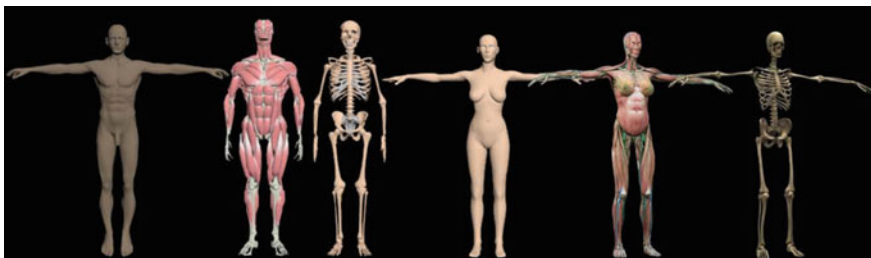


Fig. 3 3D male and female anatomy modeling (male body, male muscle, male skeleton, female body, female muscle, female skeleton)

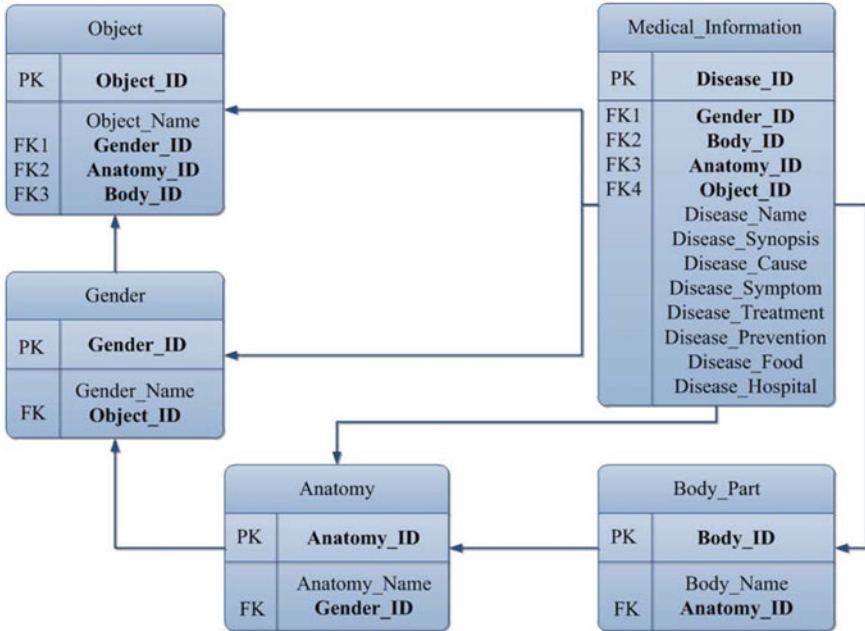


Fig. 4 Design of database table

perform a separate call to the MTL file [2]. OBJ files with MTL files were respectively modeled to have anatomy of three types such as skin, muscle and skeleton, as shown in Fig. 3. To do this, first, each part of human body was modeled to exist as a separate file. Then, it should look like a model by calling all of the files. A picking technique must be used in order to allow the user to choose specific part of human body. Also, a possible reason to look like model of the one is because each object has unique 3D coordinates in 3D space. This process should be performed according to the three types of skin, muscles, or skeleton.

3.4 Database Construction

The 3D anatomical model has been modeled separately from each other to be able to provide medical information about the diseases associated with the selected part by picking technique when you select certain part of the human body in the 3D viewer. Therefore, database construction uses each object of the 3D anatomical model [10]. Figure 4 is configuration of a table to design a database by applying the ID and name after the model is separated from each other for all parts of the human body.

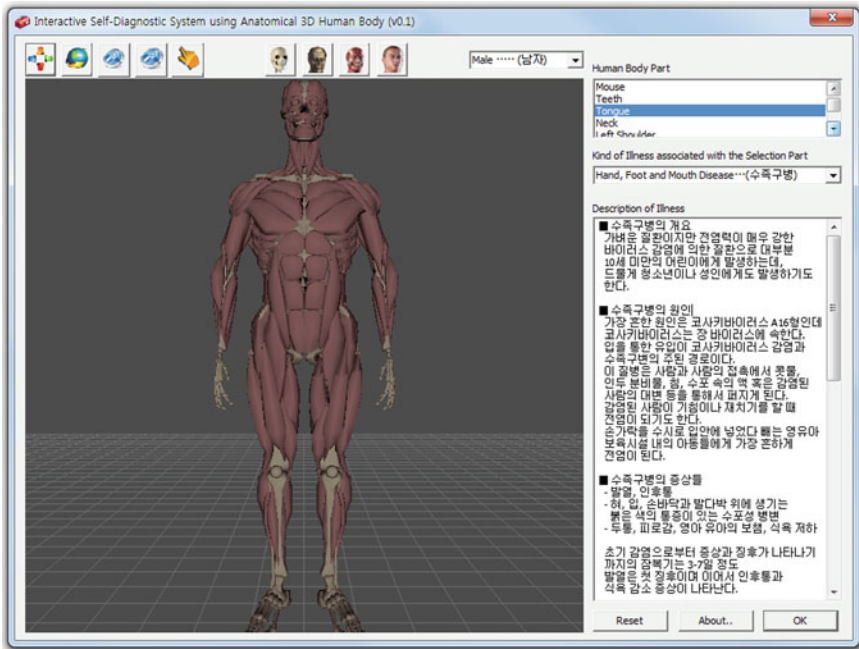


Fig. 5 Interactive self-diagnostic system

3.5 Result of Implementation

In this paper, a 3D anatomical model has been modeled in detail, which can be seen using a 3D anatomical model viewer. Also, when you select a specific part of the 3D anatomical model in the 3D anatomical model viewer of the interactive self-diagnostic system, you can see simple and clear medical information related to diseases that may occur with respect to the selected part. Figure 5 shows a male muscle model in the 3D anatomical model viewer. Also, it shows the description for the selected disease (hand, foot, and mouth diseases) and kinds of diseases that are applicable in the case of the tongue.

If you select a specific part of the 3D anatomical model, a list of the types of disease that occur in the selected part is shown. And when you select one of the diseases, this system provides simple and clearly-detailed medical information. The kind of disease is different depending on whether it is in the skin, muscle, or skeleton. Symptoms of disease are also different. Therefore, the proposed system can select the desired type of user of the 3D anatomical model such as skin, muscle, or skeleton, as well as gender. The 3D anatomical model viewer is also very intuitive to take full advantage of 3D. However, the proposed system has a disadvantage in that the 3D anatomical model is limited to adult men and woman. Therefore, 3D renderings of elderly men, elderly women, adult men, adult women,

boys, girls, infant boys, infant girls, baby boys, and baby girls must be included. The proposed system has been implemented using MFC based on Microsoft Visual Studio 2010(C++). In addition, the OpenGL API was used in order to implement the 3D viewer.

4 Conclusions

This paper is intended to develop an interactive self-diagnostic system to provide medical knowledge and information related to a disease corresponding to a specific part selected by the user using the 3D anatomical model. The 3D anatomical model has been modeled close to the actual human body by separating the skin, muscle, and skeleton. The 3D anatomical model viewer is both realistic and practical. In addition, the proposed system provides the necessary medical knowledge and information related to diseases of specific part of the 3D anatomical model selected by the user. For that reason, our proposed system is considered a very good system realistically to the general public for u-healthcare.

In the future, the interactive self-diagnostic system will provide a fine-grained 3D anatomical model depending on all ages and genders. So, the proposed system will be able to be used more efficiently. In addition, we will develop a system that can be used in a mobile environment as well as a PC environment. So we will perform an upgrade of the system so that it can be provided with medical knowledge and information anytime or anywhere for u-healthcare in a ubiquitous environment.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (No. 2012-0004478).

References

1. Healthcare. http://en.wikipedia.org/wiki/Health_care/
2. Male and Female Anatomy. <http://www.turbosquid.com/3d-models/essentials-male-female-anatomy-3d-model/671808/>
3. Ryu JK, Kim JH, Chung KY, Rim KW, Lee JH (2011) Ontology based context information model for U-healthcare service. In: Proceedings of the international conference on information science and applications 2011, IEEE CS, pp 116–121
4. Kim JH, Chung KY, Rim KW, Lee JH, Kang UG, Lee YH (2009) Design of U-healthcare service system based on personalized model in smart home. In: Proceedings of the international conference on u- and e- service, science and technology, CCIS 62, Springer, pp 104–111
5. Silén C, Wirell S, Kvist J, Nylander E, Smedby Ö (2008) Advanced 3D visualization in student-centered medical education. *Med Teach* 30(5):115–124
6. Battulga B, Konishi T, Tamura Y, Moriguchi H (2012) The effectiveness of an interactive 3-dimensional computer graphics model for medical education. doi: [10.2196/ijmr.2172](https://doi.org/10.2196/ijmr.2172)

7. Kapil S, Peter A, Ashish K (2010) The anatomy of anatomy: a review for its modernization. *Anat Sci Edu* 3(2):83–93 Wiley Online Library
8. Bharti T, Eric A, Ameya M, Sreeram V (2006) An interactive three-dimensional virtual body structures system for anatomical training over the internet. *Clin Anat* 19(3):267–274 Wiley Online Library
9. Bharti T, Eric A, Paul H, Erhan O, Alex T (2002) Web-based three-dimensional virtual body structures: W3D-VBS. *J Am Med Inform Assoc* 9(5):425–436
10. Kim JH, Kim JK, Lee D, Chung KY (2012) Ontology driven interactive healthcare with wearable sensors. *Multimedia Tools and Applications*, Published online

Rule-Based Naive Bayesian Filtering for Personalized Recommend Service

Jong-Hun Kim and Kyung-Yong Chung

Abstract The recommendation of u-Health personalized service in a semantic environment should be done only after evaluating individual physical health conditions and illnesses. The existing recommendation method of u-Health personalized service in a semantic environment had low user satisfaction because its recommendation was dependent on ontology for analyzing significance. Thus, this article suggests a personalized service recommendation method based on Naive Bayesian Classifier for u-Health service in a semantic environment. In accordance with the suggested method, the condition data are inferred by using ontology, and the transaction is saved. By applying a Naive Bayesian Classifier that uses preference information, the service is provided based on user preference information and transactions formed from ontology. The service based on the Naive Bayesian Classifier shows a higher accuracy and recall ratio of the contents recommendation than the existing method.

Keywords Naive bayesian filtering · Personalized service · Semantic service · Healthcare

J.-H. Kim

U-Healthcare Department, Bit Computer, 1327-33 Bitville Seocho-dong,
Seocho-gu, Seoul, South Korea
e-mail: kimjh@bit.kr

K.-Y. Chung (✉)

School of Computer Information Engineering, Sangji University, 83 Sangjidae-gil,
Wonju-si, Gangwon-do, Korea
e-mail: dragonhci@hanmail.net

1 Introduction

With the growing importance of the u-Health service, the kind of service that has been provided in the current web environment is now being developed as a multi-platform environment [1] with various devices applied, such as the smart phone and the smart TV. To provide the u-Health service, a service to consider a user's situation via private data needs to be offered. Today, in order to infer the context data, ontology is used, which has been proved to be highly efficient for the providing of the service [2]. The u-Health service in the semantic environment that has been offered for a personalized service is something to be inferred from the context data such as a person's disease, location, and weather, and this service is capable of being customized [3]. For the personalization recommendation of the u-Health service, the intellectual personalization recommendation service is conducted based on a user's vital information and a log of recommended information [4]. The current u-Health service in the semantic environment has used a rule-based filtering method and ontology and has conducted a semantic analysis of input data with a rule base applied to make inferences. This has made the current u-Health service capable of providing the intellectual service, but there have been limitations in its ability to satisfy user preferences. To date, not a single personalization service has succeeded [5]. Considering this, this thesis has proposed using the rule-based Naive Bayesian [6] filtering method so that the personalized recommendation service of the semantic environment would be realized. The proposed method would use ontology to infer the context data. For the service contents filtering, the Naive Bayesian method is adapted and, as the transaction is created by the ontology [7] and the user's preference information is filtered, a service is provided.

The rest of this paper is organized as follows. [Section 2](#) describes related. [Section 3](#) describes the rule-based Naive Bayesian filtering. [Section 4](#) describes the experiment and result. The conclusions are given in [Sect. 5](#).

2 Related Works

For the filtering methods for the personalization recommendation method are represented by the collaborative filtering method and the rule-based filtering method. Collaborative filtering [8] makes use of a secure database of users' service preferences, and new preference information will be similar to a new user's taste. For the rule-based filtering [9], a collaborative expert defines a user's preferences and makes a recommendation in association with the user's defined preferences and the features of a service. Much research on the personalization recommendation with the filtering methods mentioned above is being conducted. One study on the semantic web technique, which is a rule-based filtering method, would reduce research time, gradually improving the satisfaction level through feedback [10]. However, since it

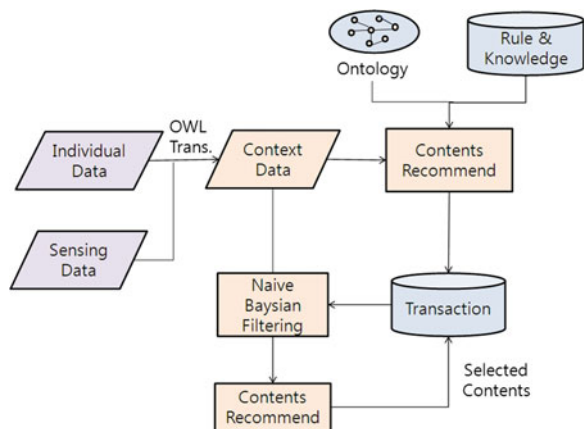
depends on rule-based filtering, an evaluation of the recommendation service is required in advance. Hence, as long as the rule-based filtering does not concern the preference information, a collaborative way to use preference information is necessary to enhance the user’s satisfaction level. For this reason, the Naive Bayesian-based filtering method is considered more convenient with less overhead compared with the high complexity filtering, for the former would reduce active analysis time [6].

3 Rule-Based Naive Bayesian Filtering

To provide the individualized service in the semantic environment, it is first necessary to convert the sensing data measured from the individual data and a sensor to Ontology Web Language (OWL) [11], which is possible to infer by the ontology, so that the sensing data are constructed to be context data. Second, as the ontology and the rule base are used based on the context data, the service contents are inferred. The created inference service contents information would be stored in the learning transaction. With the learning transaction, the current log data, applied, the Naive Bayesian inference is made not only to filter the entered context information and the service inference information but also to provide the customized recommendation service. When the user is recommended for some customized service contents and chooses one, the selected information is entered in the learning transaction. Figure 1 shows a rule-base Naive Bayesian filtering-based recommendation method for the personalization service recommendation.

To recommend the service contents by the Naive Bayesian filtering with the ontology-oriented learning and using the user’s preference information, the learned records are required and, with this, the conditional probability inference would be possible. The learning transaction is private record information to be filtered using the Naive Bayesian method. The structure of the learning transaction record and its

Fig. 1 Rule-based Naive Bayesian filtering method



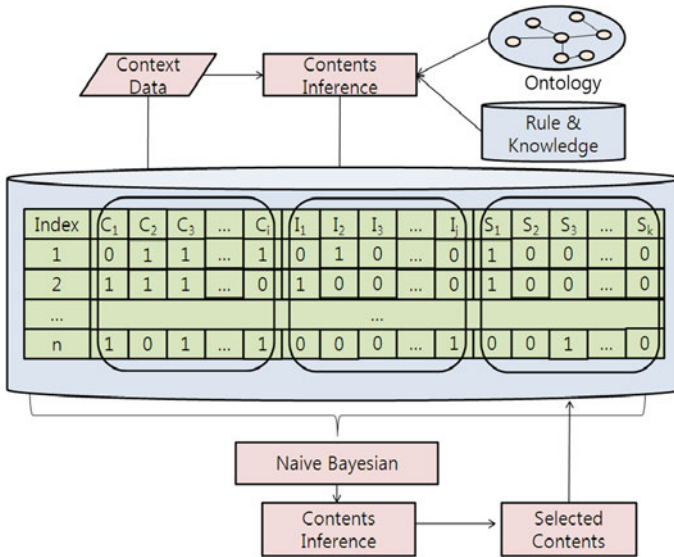


Fig. 2 Structure of learning transaction record

relation are described in Fig. 2. $C_1 \sim C_i$ comprise the context information entry section, $I_1 \sim I_j$ make up the contents information that has been inferred from the context information, and $S_1 \sim S_k$ are the preference information that the user has chosen. In other words, I_j and S_k indicate the same contents. In addition, each datum of information is composed of 0 (false) and 1 (true). The information gained from the context data is entered in the record as $C_1 \sim C_i$ and, when it is 0, it means that the context data have not yet been entered while, when it is 1, it should be confirmed that the context data have been entered. The information on the service contents inferred from the ontology based on the context information is entered in $I_1 \sim I_j$ and, when it is 0, the contents have not yet been inferred while, when it is 1, it indicates that the result has come out with a service inferred.

4 Experiment and Result

In this thesis, the precision and the recall of the proposed filtering method is evaluated. The comparative objects for the filtering performance evaluation are the Context Rule-based Filtering Method (CRFM) and the Rule-based Naive Bayesian Filtering Method (RNBM). The equation of the precision and the recall for the experimentation is presented as Eq. 1.

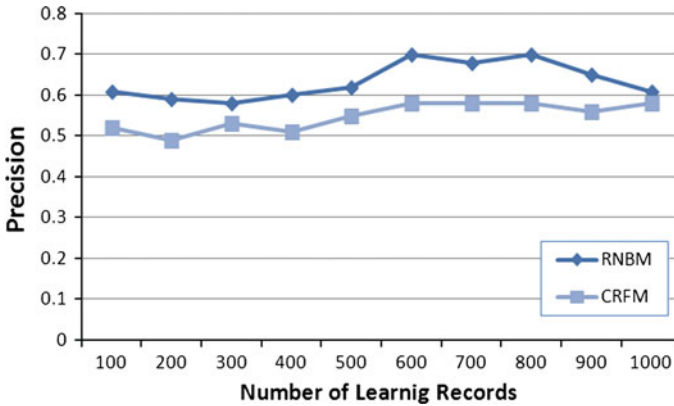


Fig. 3 Result of precision experiment

$$\begin{aligned}
 Precision &= \frac{|Selected Contents \cap Recommended Contents|}{|Recommended Contents|} \\
 Recall &= \frac{|Selected Contents \cap Recommended Contents|}{|Selected Contents|}
 \end{aligned}
 \tag{1}$$

For this scenario, the precision and the recall of the recommendation information are compared according to the number of records learned in the learning transaction. For the first measurement method, the precision is measured with a contents rate that is connected to the recommended contents. When the contents that have been associated with the input data and the preference are recommended, a high figure is presented. Here, the contents are provided as they are divided into what the user has actually selected and what have been recommended. The experimentation results are described in Fig. 3.

According to the precision analysis results, for the average precision of each test group, RNBN is 0.634 and CRFM is 0.548. Considering this, it is determined that RNBM has better precision. For RNBM, since the preference value and the inference value are simultaneously measured, the precision has appeared high. For CRFM, only the input data have been used. However, when the number of experimentation records exceeds 800, the precision of RNBM is found to be reduced. The reason for this is that the user has possibly selected other contents that would not go with the context data as not having selected the inferred service contents. For the second measurement method, since the recall is a figure that has been divided by the number of all the contents related to the recommended contents, the recall presents a high value as long as the contents that have been selected by the user are recommend. Here, the contents are figured as they are divided into what the user has collected and what has been actually recommended. The experimentation results are described in Fig. 4.

The recall test results said that, for the average recall of each test group, RNBM is 0.457 while CRFM is 0.311. In other words, RNBM is believed to have a higher

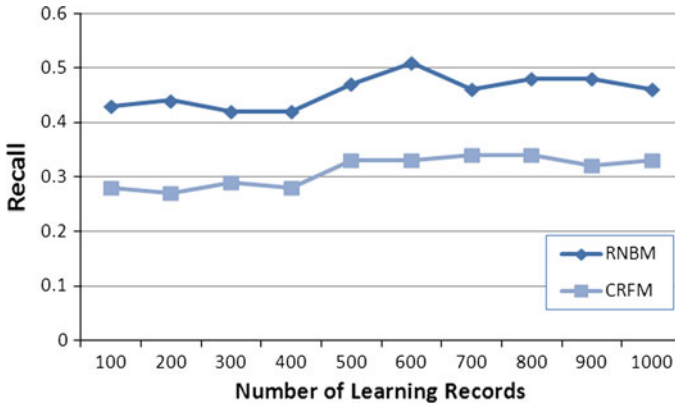


Fig. 4 Result of recall experiment

recall. Since RNBM has reflected the user’s preference, it has to have a higher recall than CFRM, which has not reflected the user’s preference.

5 Conclusion

In this thesis, a rule-based Naive Bayesian filtering-based personalization service recommendation method in the semantic environment has been proposed. The current semantic environment has not adapted to the user’s preference information and it is still the rule-based filtering. This has made the user’s satisfaction drop. For this reason, through the Naive Bayesian method, this thesis has realized that the context information, the rule-based inference information, and the contents selection information with the filtering concerns the rule-based filtering precision and the user’s preference information. According to the experimentation results, the filtering method proposed by this thesis is considered to have better precision and better recall than the current rule-based filtering. The proposed filtering method is expected to increase the contents recommendation satisfaction of the users of the u-Health service.

Acknowledgments This work was supported by the R&D Program of MKE/KEIT. Sincere thanks go to Mr. Jaekwon Kim who provided the idea for this thesis.

References

1. Park DK, Kim JH, Kim JK, Jung EY, Lee YH (2011) U-Health service model for managing health of chronic patients in multi-platform environment. *J Korea Cont Assoc* 11(8):23–32
2. Lee BM, Kim JK, Kim JH, Lee YH, Kang UG (2011) A customized exercise service model based on the context-awareness in u-health service. *J Korean Inst Info Technol* 9(2):141–152
3. Ryu JK, Kim JH, Kim JK, Lee JH, Chung KY (2011) Context-aware based u-health environment information service. *J Korean Inst Info Technol* 11(7):21–29
4. Kim JH, Lee DS, Chung KY (2011) Context-aware based item recommendation for personalized service. In: *Proceedings of the international conference on information science and applications*, pp 1–6
5. Ben Schafer J, Frankowski D, Herlocker J, Sen S (2007) Collaborative filtering recommender systems. *Lecture notes in computer science*, pp 291–324
6. Tan PN, Steinbach M, Kumar V (2007) *Introduction to data mining*, Addison Wesley, Upper Saddle River
7. Russell S (2010) *Artificial intelligence: a modern approach*, 3rd edn. Paperback, Pearson Education
8. Resnick P et al (1994) GroupLens: an open architecture for collaborative filtering of netnews. In: *Proceedings of ACM CSCW'94 conference on computer supported cooperative work*, pp 175–186
9. Breese JS, Heckerman D, Kadie C (1998) Empirical analysis of predictive algorithms for collaborative filtering. In: *Proceedings of the conference on uncertainty in artificial intelligence*, Madison
10. Eun CS, Cho DJJ, Jung KY, Lee JH (2007) Development of apparel coordination system using personalized preference on semantic web. *J Korean Inst Info Technol* 7(4):66–73
11. Owl Web Ontology Language Over View, <http://www.w3.org>

Design of an Actigraphy Based Architecture for Mental Health Evaluation

Mi-hwa Song, Jae-Sung Noh, Seung-Min Yoo and Young-Ho Lee

Abstract This paper introduces a decision support system architecture for continuous activity recognition and actigraphy, which are important for mental health evaluation; the architecture is based on triaxial accelerometer data. Recent developments in acceleration sensor device technologies have made it possible to precisely measure the acceleration of motor activity with a triaxial accelerometer for a lengthy period of time. We propose an AMD (Actigraphy based Mental health Decision support system) architecture for objectively measuring daily activity, recognizing continuous activities, and analyzing the behavior pattern of people with mental disorders, as well as the correlation between change in mood symptoms and mental disorders.

Keywords Actigraphy · Tri-axial accelerometer · Activities of daily living · Activity recognition

M. Song (✉)

U-Healthcare Institute, Gachon University, 191 Hambakmoeiro, Yeonsu-Gu, Incheon, South Korea
e-mail: mhsong@gachon.ac.kr

J.-S. Noh · S.-M. Yoo

Department of Psychiatry & Behavioral Sciences, School of Medicine, Ajou University, 5- Wonchon-Dong, Yeoungtong-gu, Suwon, South Korea
e-mail: jsnoh@ajou.ac.kr

S.-M. Yoo

e-mail: greatshyung@gmail.com

Y.-H. Lee

IT Department, Gachon University, 191 Hambakmoeiro, Yeonsu-Gu, Incheon, South Korea
e-mail: lyh@gachon.ac.kr

1 Introduction

The activity disorder among the mental disorder had been receiving the concern for a long time. According to Diagnostic and Statistical Manual of Mental Disorders, 4th Edition (DSM-IV), published by the American Psychiatric Association, several items of DSM-IV include the change of the activity in the diagnostic criteria of the disease [1]. Moreover, the depression patient shows the delay to the whole activity, body movement, language ignition and reaction time [2]. Even though the activity aspect was important, the study of the pattern of the movement of mental disorders relatively had been being behind.

Actigraphy means the value which is consecutively recorded by the body added type device for measuring the movement. And the measurement of the movement is usually made of the acceleration. Generally the sensor-device adheres to the waist. According to the device, it is characterized that information during several day or several months can be stored [3]. Recently, it is possible that objectively the long time the motoric activity is measured by the development of information technology. Thus, in this research, we propose AMD architecture that objectively measures daily activity, recognizes continuous activities and analyzes correlation between change of the mood symptom and the mental disorder and the difference of the behavior pattern for the people with mental disorder.

2 Related Work

Currently, it is not possible to check changes in the physical activity level of an inpatient. It is also difficult to determine the correlation between any changes in the activity level with the existing scale and the motoric component [4][5]. A previous study indicated that a considerable amount of movement in patients reflected a feeling of restlessness and agitation [6]. Another report has shown the possibility of objectively evaluating an increase in meaningless movements, which is observed at the beginning of anxiety, by actigraphy [7]. The number of studies on the potential use of exercise training for treating patients with melancholic and anxiety disorder is relatively small

Meanwhile, medical experts often have no way idea how their patients with depressive mood spend the time at home. Burns et al. [8] are developing a smartphone application that can not only monitor the frequency of phone calls, SMS, and emails but also use the phone's GPS and accelerometer to check if people have been lying down at home all day long. After determining their daily activity patterns, the application should be able to detect any changes in their behavior.

3 System Design

The goal of this study is to develop a model for objectively measuring the mental state of a patient. The information provided by such a model can be used to determine the course of treatment for mental disorders and to identify measures for their prevention. Figure 1 shows a flowchart that indicates how information from models can be used. We propose that the best way of solving the challenging problems discussed in Sect. 2 is through the development of a general architecture which is flexible enough that it enables us to prototype systems which are reconfigurable, and thus can address the problems in different ways.

3.1 Preprocessing

The AMD involves the following subordinate tasks for the management of input data collection and analysis. On the basis of functional and computational requirements, these tasks can be classified as follows:

- (1) The acceleration measurement range is -8 to 8 G and the sampling frequency range is 32 to 1/30 Hz. By using the very large built-in memory, data for

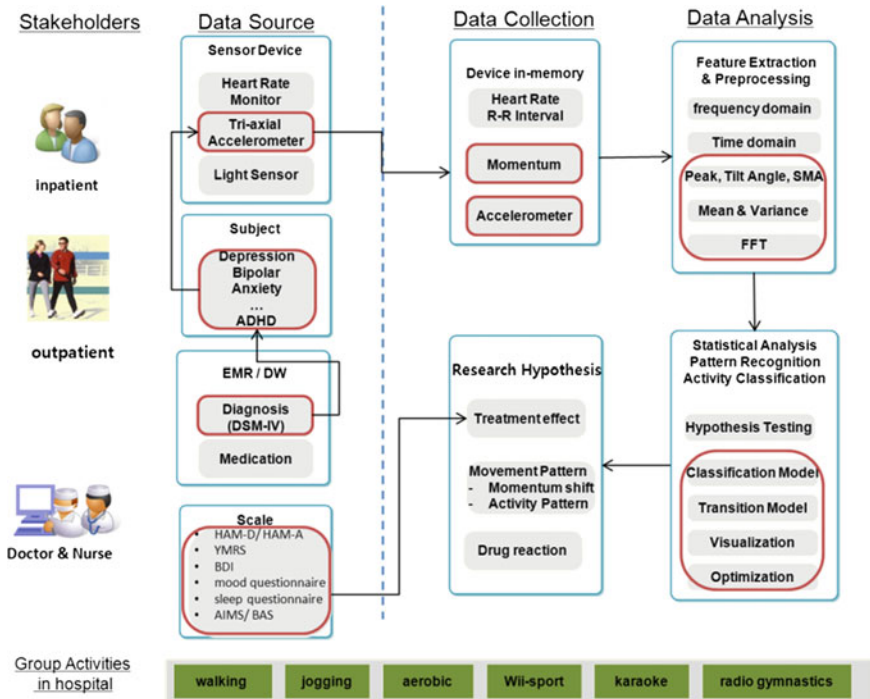


Fig. 1 Overview of the research. Red lines denote our research scope

periods up to 6 months can be stored. Data can be transmitted to a computer through the USB cable connection.

- (2) Basic processing of the input data is performed. For example, fast Fourier transform of data can be performed.
- (3) The work training activity data and maintaining the prior state. It means to do the inference mechanism like RBFN and the average of collected data stream the maintain from the transition model like HMM (Hidden markov model).

3.2 Inference Tasks

For our architecture implementation, the experimental hypothesis is based on the view that the task of labeling a series of class tags on a data object that represents the duration of an actigraphic event should be supported by a statistical model capable of robustly capturing the probabilistic context influencing the generation of a set of features at a given discrete time. In view of the sequential nature of event generation, building an inference model that encodes the state transition event structure was a very natural design step. That is, we needed a classifier, also known as class likelihood, linearly combined with a prior that is the probabilistic generalization of a state transition event in actigraphy. The idea of using a linear combination of likelihood and prior for building a pattern recognition system is not new. It has been well defined, especially in the field of statistical speech recognition and statistical machine translation, in which there is a wide range of signal processing techniques. Jelinek [9] presents an introduction to the parameters of Gaussian mixture and hidden Markov models. Our approach, however, does not build upon the Gaussian mixture learning scheme, but borrows from a variety of

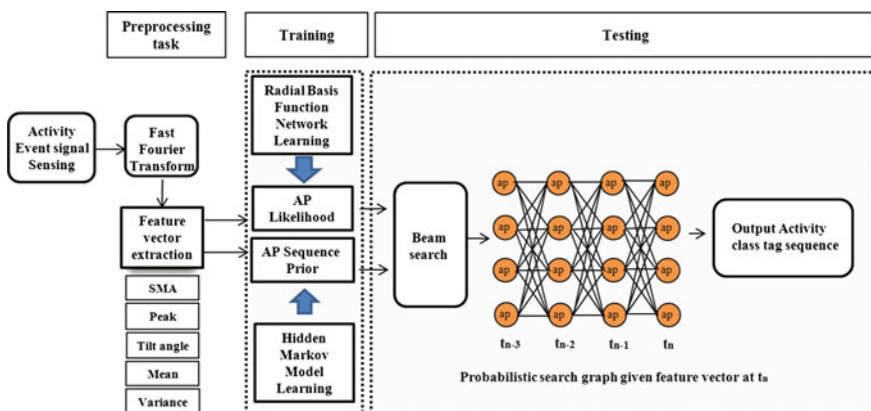


Fig. 2 Actigraphy based Inference architecture, ap: a node representing a likelihood of activity. AP denote activity or posture of daily living

methods, including transition event modeling. Figure 2 shows an actigraphy based inference architecture.

Since we hypothesized that embedding the memory of previous contexts into the inference process would improve the recognition capability of the system, the probabilistic generalization of the transition event in the human AP sequence was linearly combined with the classifier learned using a radial basis function network (RBFN) [10], where the memory structure was modeled on the basis of Markov chain learning. The rationale behind the probabilistic graphical model as an additional inference component is that human actigraphy reveals both a sequential and a recursive nature so that a class of an Activity object—a feature vector of a framed signal—at a given discrete time should become more recognizable by observing the class of the previous AP object. Using the prior model of such a stochastic tendency in both sequential and recursive event generation plays the role of a constraint, during classification at run time, on the range of the probable class subset that contributed to the generation of given actigraphic data.

4 Conclusions

AMD is an actigraphy-based architecture that can help in the diagnosis of mental disorders and in determining the course of treatment. The domain of activity event classification has recently attracted much interest from research groups. It seems to pose various challenges that require combined research efforts from many interdisciplinary fields including signal processing, machine learning, machine vision, and sensor networks. In addition, recently, devices that can measure physical activity by using in-built acceleration sensors have been developed. The information obtained from the accelerometer sensor should be pragmatically applied to analyze the physical activity pattern of inpatients and outpatients with mental disorders. The development of algorithms and decision support systems to process the information is also important.

Acknowledgments This research was supported by grant no. 10037283 from the Industrial Strategic Technology Development Program funded by the Ministry of Knowledge Economy.

References

1. Tryon WW (1986) Motor activity measurements and DSM-III. *Prog Behav Modif* 20:35–66
2. Sobin C, Sackeim H (1997) Psychomotor symptoms of depression. *Am J Psychiatry* 154:4–17
3. Tryon WW (1991) Activity measurement in psychology and medicine. Plenum Press, New York
4. Finazzi ME, Mesquita ME, Lopes JR, Fu LI, Oliveira MG, Del Porto JA (2010) Motor activity and depression severity in adolescent outpatients. *Neuropsychobiology* 61:33–40

5. Razavi N, Horn H, Koschorke P, Hugli S, Hofle O, Muller T, Strik W, Walther S (2011) Measuring motor activity in major depression: the association between the Hamilton depression rating Scale and actigraphy. *Psychiatry Res* 190:212–216
6. Grap MJ, Hamilton VA, McNallen A, Ketchum JM, Best AM, Arief NY, Wetzel PA (2011) Actigraphy: analyzing patient movement. *Heart Lung* 40:52–59
7. Mistraletti G, Taverna M, Sabbatini G, Carloni E, Bolgiaghi L, Pirrone M, Cigada M, Destrebecq AL, Carli F, Iapichino G (2009) Actigraphic monitoring in critically ill patients: preliminary results toward an “observation-guided sedation”. *J Crit Care* 24:563–567
8. Michelle NB, Mark B, Jennifer D, Darren G, Chris JK, Emily G, David CM (2011) Harnessing context sensing to develop a mobile intervention for depression. *J Med Int Res* 13(3):e55
9. Jelinek F (1997) *Statistical methods for speech recognition*. MIT Press, Cambridge
10. Rousseeuw PJ, Leroy AM (1987) *Robust regression and outlier detection*. Wiley, New York, pp 134–150

Efficient Detection of Content Polluters in Social Networks

Jin Seop Han and Byung Joon Park

Abstract A large number of Internet users are currently using social networking services (SNS) such as Twitter and Facebook. However, the SNS users are exposed to threats of malicious messages and spams from unwanted sources. It would be useful to have an effective method for detecting spammers or content polluters on social networks. In this paper, we present an efficient method for detecting content polluters on Twitter. Our approach needs only a few feature values for each Twitter user and hence requires a lot less time in the overall mining process. We demonstrate that our approach performs better than the previous approach in terms of the classification accuracy and the mining time.

Keywords Social Network · Content polluters · Detection scheme

1 Introduction

Social networking services (SNS) allow people to communicate and share information with their acquaintances as well as those they do not know in person. There are currently many social networking services such as Twitter and Facebook. Users can create their accounts easily with a social networking site through a simple input of email addresses and once they get in the network, they can

J. S. Han · B. J. Park (✉)
Department of Computer Science, Kwangwoon University,
447-1, Wolgye-dong, Nowon-gu, Seoul, South Korea
e-mail: bjark@kw.ac.kr

J. S. Han
e-mail: jshan74@kw.ac.kr

communicate with practically anyone on the network. However, because of this openness and accessibility of social networking services, the SNS users are exposed to threats of malicious messages and spams from unwanted sources. Thus, it would be useful to have an effective method for detecting spammers or content polluters on social networks. Recently, there has been some research work on detecting spammers or content polluters [1–6]. Most of them use some characteristics of spam messages and spammers to build classification models for the detection purpose. In this paper, we present an efficient method for detecting content polluters on Twitter. We also use, like other approaches, some behavioral patterns and characteristics of Twitter users to identify content polluters, but in a much more efficient way.

The rest of this paper is organized as follows: Sect. 2 describes some related work on detecting spammers or content polluters in social network services. Section 3 presents our approach that could efficiently detect content polluters on Twitter. In Sect. 4, we demonstrate the effectiveness and efficiency of the proposed approach by showing our experimental results with the real Twitter data set. Finally, we draw our conclusion in Sect. 5.

2 Related Work and Background

Lee et al. [6] discusses a content polluter detection method based on a wide variety of characteristics of each Twitter user, whether a legitimate one or a polluter. Their approach is directly related to ours since both methods aim to solve the same problem, i.e. to build a classifier model for content polluters in Twitter based on a set of characteristics of Twitter users. We will show that our approach will generate a classifier model that is at least as accurate as theirs, but in a much more efficient way in terms of data preparation and model building time. Since our approach and experiments use the same data set used by Lee et al. [6, 7], we describe their work in more detail in the next subsection. Hereafter, we will call their approach ‘Social-Honeypot’ for future reference in this paper.

To collect the characteristics data about content polluters in Twitter, the authors deployed special entities, i.e. honeypots, that tempt unwelcome automated Twitter accounts. Since these honeypots communicate only with other honeypot agents, not regular Twitter users, the authors assumed any users trying to follow these honeypot accounts to be content polluters. After various data such as their tweet contents, user demographics, and friendship structures had been collected from those content polluters, the authors began analyzing the data and finally identified 19 features to build a classification model for the content polluters and the legitimate users. Compared to other previous approaches, they used a much wider spectrum of characteristics of each user including the behavioral patterns and the change of the number of following over time. For each Twitter user, they obtained the following feature values to prepare the data set for the classification model [6].

- **User Demographics** (UD): the length of the screen name, the length of description, and the longevity of the account
- **User Friendship Networks** (UFN): the number of following, the number of followers, the ratio of the number of following and followers, the percentage of bidirectional friends, the percentage of bidirectional followers, the standard deviation of unique numerical IDs of following, and the standard deviation of unique numerical IDs of followers
- **User Content** (UC): the number of posted tweets, the number of posted tweets per day, the average number of links in a tweet, the average number of unique links in a tweet, the average number of @username's in a tweet, the average number of unique @username's in a tweet, the average content similarity over all pairs of tweets posted by a user, and the ratio of uncompressed size of tweets and their compressed size
- **User History** (UH): the change rate of number of following

3 The Proposed Approach for Detecting Content Polluters

Using 20 feature values (19 attributes described in subsection 2.2 and one classification attribute) for each Twitter user, the Social-HoneyPot approach successfully classified (with 98.42 % accuracy rate) content polluters from legitimate users. They reported that all those 20 features were required to achieve the best accuracy rate and that using other subsets of features resulted in less accurate classification models [6]. To achieve such a high accuracy rate, however, it has to spend a lot of time in extracting and computing those 19 feature values from the collected Twitter data. In addition, the time needed to build a classification model also gets larger as the number of features for each training instance grows. Thus, it will be beneficial to have an efficient detection method that needs less time in preparing the training data set and in building a classification model for content polluters. We propose an efficient approach which needs only several feature values and hence requires a lot less time in the overall mining process.

Our approach considers, among others, the change rate of the number of messages each user posted over a period of time. It is based on the intuition that regular legitimate users wishing to socialize or share information with others would post messages rather steadily over time than would the content polluters or spammers do. That is, our intuition is that, compared to legitimate users, there will be greater variations on the numbers of posted messages from the content polluters or spammers. We use the following formula (similar to the one in [6]) to compute the change rate of number of posted messages for each Twitter user:

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} |p_{i+1} - p_i|}$$

where n is the total number of days during the observation period and p_i is the number of messages posted by the user for the i th day.

To prepare a training data set for classification, we exploit the values for five other features which were also adopted by the Social-Honeypot approach, in addition to our proposed feature: the change rate of number of following, the number of following, the standard deviation of unique numerical IDs of following, the standard deviation of unique numerical IDs of followers, and the longevity of the account. All of these five features have higher discriminative power than other 14 features discussed in [6].

In the next section, we will demonstrate that our approach based a small set of observed features can build a classifier much more efficiently with even higher accuracy than the Social-Honeypot approach.

4 Experimental Evaluation

For the experimental evaluation of our approach, we have used the same Twitter data set as the one used by Social-Honeypot [6, 7]. The data set contains all the demographic data and tweets of 41,499 Twitter users, of which 46.4 % (19,276 users) are legitimate ones and 53.6 % (22,223 users) content polluters. The numbers of tweet messages by legitimate users and content polluters are 3,263,238 and 2,380,059, respectively. From this initial data set, we have conducted a series of experiments to compare the performance of two approaches in terms of both the classification accuracy and the time involved in data preparation and model creation.

For the experiments, we used the Weka toolkit [8] version 3.7.6 to create a classification model. To prepare a data file for the Weka system, we have extracted and computed the required feature values for each Twitter user from the initial data set and made an ARFF file from these computed values. Also, 10-fold cross-validation was employed to obtain various statistics. We used a PC with an Intel(R) Core(TM) i7-2600 k CPU of 3.40 GHz, 8 GB RAM running Windows 7 and C Programming Language for the experiments.

Our first set of experiments was conducted to evaluate the effect of the proposed single feature (i.e., the change rate of the number of posted messages) on the performance in terms of the accuracy of the created classification model. The extra time required to compute and add the proposed feature value to the existing ARFF file was 6 s. To see if the size of input data set affects the accuracy of the classifier, we have conducted several experiments with different data sizes: 1, 5, 10, 20, 50, and 100 % of the original data set respectively and all of them having the same proportion of legitimate users and content polluters as the original data set. As shown in Table 1, the greater the input size is, the higher, although slightly, the accuracy value we get. This indicates that we can get a better result from more training data. In all the experiments except the 1 % case, the inclusion of the

Table 1 Accuracy values with various data sizes

Size (Legit/Spammer) Model	1 % of total (193/ 223)	5 % of total (964/ 1,112)	10 % of total (1,928/ 2,223)	20 % of total (3,856/ 4,445)	50 % of total (9,638/ 11,112)	Total data of HoneyPot (19,276/ 22,223)
Social HoneyPot (%)	96.63	97.45	97.76	97.94	98.35	98.4216
Social HoneyPot + the proposed feature (%)	96.63	97.69	97.86	98.24	98.44	98.6048
5 features+the proposed feature (%)	96.88	97.06	98.05	97.96	98.42	98.4241

Table 2 Accuracy and processing time with various feature sets

Feature Set	Accuracy (%)	Data preparation time	Model-building time
Social HoneyPot (19 features)	98.4216	02:35:20	00:00:07.51
Social HoneyPot (19 features) + the proposed feature	98.6048	02:35:26	00:00:08.17
5 features + the proposed feature	98.4241	00:00:07	00:00:04.58

proposed feature values into the input ARFF file resulted in classifiers with accuracy values consistently higher than those of Social-HoneyPot. In case of the 1 % population, two accuracy values are the same. Thus, we can say that a slightly higher accuracy value could be obtained at the expense of extra few seconds.

We have conducted another set of experiments to compare the efficiency of our approach based on the proposed feature with Social-HoneyPot. With the original Social-HoneyPot data, it took 2 h 35 min 20 s to prepare the input ARFF file using the 19 features, 7 s using 6 features including the proposed feature, and 2 h 35 min 26 s using all of the 20 features including the proposed one. Because it took a lot of time to compute the average content similarity over all pairs of tweets posted by a user and the ratio between uncompressed size of tweets and their compressed size, there are huge differences between ours and the Social-HoneyPot-based approach. The experimental results are shown in Table 2. We can see that our approach based on a smaller set of features can build a better classifier with a higher accuracy value (although not the highest) in a drastically shorter time.

5 Conclusion

In this paper, we described an efficient detection method for content polluters operating on Twitter. We have proposed a single, yet important feature, the change rate of the number of posted messages for each user, and defined a small set of observed features that need to be computed from the raw Twitter data set. With a

series of experiments, we demonstrated that our approach could build a classifier with the higher accuracy values in much shorter time compared to the previous approaches.

Acknowledgments The present research has been supported by the Research Grant of Kwangwoon University (No. 60012007197)

References

1. Bosma M, Meij E, Weerkamp W (2012) A framework for unsupervised spam detection in social networking sites. In: European conference on information retrieval (ECIR), pp 364–375
2. Abu-Nimeh S, Chen T, Alzubi O (2011) Malicious and spam posts in online social networks. *IEEE Comput Soc* 44(9):23–28
3. Beck K (2011) Analyzing tweets to identify malicious messages. In: Proceedings of IEEE international conference on electro/information technology (EIT), pp 1–5
4. Stringhini G, Kruegel C, Vigna G (2010) Detecting spammers on social networks. In: Proceeding of annual computer security applications conference (ACSAC), pp 1–9
5. Wang A (2010) Don't follow me: spam detection in twitter. In: Proceeding of international conference on security and cryptography (SECRYPT), pp. 1–10
6. Lee K, Eoff B, Caverlee J (2011) Seven months with the devils: a long-term study of content polluters on twitter. In: Proceedings of AAAI international conference on weblogs and social media (ICWSM), pp. 185–192
7. http://infolab.tamu.edu/static/users/kyumin/social_honeypot_icwsm_2011.zip
8. <http://www.cs.waikato.ac.nz/ml/weka/index.html>

A Prototype Selection Algorithm Using Fuzzy k -Important Nearest Neighbor Method

Zhen-Xing Zhang, Xue-Wei Tian, Sang-Hong Lee and Joon S. Lim

Abstract The k -Nearest Neighbor (KNN) algorithm is widely used as a simple and effective classification algorithm. While its main advantage is its simplicity, its main shortcoming is its computational complexity for large training sets. A Prototype Selection (PS) method is used to optimize the efficiency of the algorithm so that the disadvantages can be overcome. This paper presents a new PS algorithm, namely Fuzzy k -Important Nearest Neighbor (FKINN) algorithm. In this algorithm, an important nearest neighbor selection rule is introduced. When classifying a data set with the FKINN algorithm, the most repeated selection sample is defined as an important nearest neighbor. To verify the performance of the algorithm, five UCI benchmarking databases are considered. Experiments show that the algorithm effectively deletes redundant or irrelevant prototypes while maintaining the same level of classification accuracy as that of the KNN algorithm.

Keywords k -nearest neighbor (KNN) · Prototype selection (PS) · Fuzzy k -important nearest neighbor (FKINN)

Z.-X. Zhang
School of Information and Electric Engineering, Ludong University,
Yantai, China
e-mail: billzhenxing@gmail.com

X.-W. Tian · S.-H. Lee · J. S. Lim (✉)
IT College, Gachon University, San 65 Bokjeong-dong Sujeong-gu,
Seongnam, Gyeonggi-do, South Korea
e-mail: jslim@gachon.ac.kr

X.-W. Tian
e-mail: tianxuemaog@gmail.com

S.-H. Lee
e-mail: shleedosa@gmail.com: shleedosa@gachon.ac.kr

1 Introduction

The k -Nearest Neighbor (KNN) classification based on the class of their nearest neighbors is quite straightforward [1]. However, KNN belongs to instance-based learning, which strongly depends on the training samples, and all computations can start only when classification is completed. The possibility of the memory requirement, time complexity, and accuracy of the classifier being affected is high [2–4]. Therefore, the shortcoming of KNN lies in the huge amount of calculation. In this regard, a solution is to select important samples prior to the training process and to remove in advance those samples that play a minor role in the classification process. Algorithms that implement this solution are collectively known as Prototype Selection (PS) algorithms [5–8].

This paper presents a new PS algorithm, namely Fuzzy k -Important Nearest Neighbor (FKINN) algorithm. In this algorithm, an important nearest neighbor selection rule is introduced. When classifying a data set with the FKINN algorithm, the most repeated selection sample is considered an important nearest neighbor. According to the important nearest neighbor selection rule, the deletion of redundant or irrelevant examples will reduce the computational time and the amount of memory needed to run the classifier.

To verify the performance of the FKINN algorithm, five UCI benchmarking databases are employed in this study. For classifying these databases with the algorithm, the 5 times two fold cross-validation tests are applied to operate the experiment. It is observed that the FKINN algorithm effectively deletes redundant or irrelevant prototypes while maintaining the same level of classification accuracy as that of the KNN.

2 Fuzzy k -Important Nearest Neighbor

This section describes the FKINN algorithm. The algorithm includes two processes, a training process and a test process, as shown in Fig. 1. The training process is divided into three steps. In step 1, a training data set is classified by a fuzzy KNN algorithm [9]. During this classification process, the value of k corresponding to the maximum recognition rate is selected. In step 2, the important nearest neighbors (INNs) are selected by using the INN selection rule, which is described in Sect. 3.2. In step 3, at the end of the training process, all the INNs are selected. The INNs are used as prototypes when executing fuzzy KNN, in the test process. Every sample of test data set finds out k -nearest neighbors which are from the INNs. Finally, classification is performed with the fuzzy KNN algorithm.

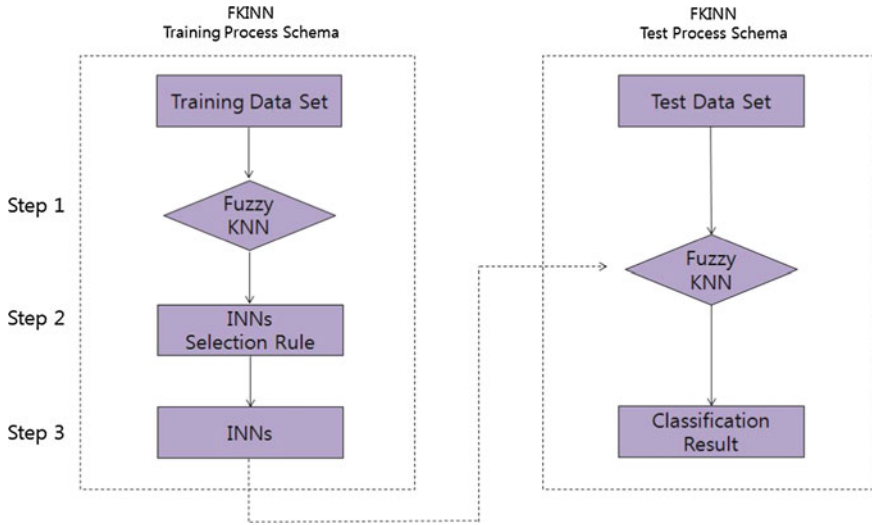


Fig. 1 FKINN algorithm schema

3 Experimental Result Evaluation

In this study, five real-world databases (DBs) were used. Table 1 shows the names and details of the DBs, such as the number of samples, features, and classes [10].

The FKINN algorithm is compared with the KNN algorithm by using a 5 times two-fold cross-validation method. In Table 2, the three factors used to compare the classification results of the FKINN and KNN algorithms are shown. These factors are the Mean Accuracy (MA), which is the average value of ten classification rates; the Mean Prototype (MP), obtained by averaging ten classification prototypes; and the Mean Test Time (MTT), which is the average of ten test times.

In the classification of a breast cancer DB, the MA, MP, and MTT in the case of the KNN algorithm are 96.28 %, 350, and 0.06 s, respectively; for the FKINN algorithm, the values are 96.6 %, 73, and 0.002 s, respectively. The statistical results of F-tests are shown in Table 2. For the classification of the breast cancer DB, the FKINN algorithm results in a lower MP ($p < 0.05$) and a shorter MTT ($p < 0.05$) compared to the KNN algorithm, while there is no significant difference

Table 1 Summary of benchmarking data set

Database	Number of samples	Number of features	Number of classes
Breast cancer	699	9	2
Pima Indians	768	8	2
Balance scale	625	4	3
Landsat	6435	36	6
Pendigits	10992	16	10

Table 2 5*2 Cross-Validation F-Test for MA, MP and MTT

Data sets	Analysis index	KNN	FKINN	F-Test (<i>p</i> value)
Breast cancer	MA	96.28 %	96.6 %	0.515
	MP	350	73	0.000
	MTT	0.06	0.025	0.002
Pima Indians	MA	72.32 %	73.95 %	0.258
	MP	384	185	0.00
	MTT	0.08(s)	0.035(s)	0.002
Balance scale	MA	86.82 %	87.91 %	0.225
	MP	313	117	0.00
	MTT	0.07(s)	0.04(s)	0.002
Landsat	MA	89.23 %	89.44 %	0.232
	MP	3218	1705	0.00
	MTT	6.29(s)	3.6(s)	0.00
Pendigits	MA	99.04 %	96.88 %	0.294
	MP	5496	2661	0.00
	MTT	15.61(s)	5.4(s)	0.00

in their MAs ($p > 0.05$). The same method is used for the other four DBs, that is, the corresponding F-tests are conducted on them. The results for all the DBs are the same as those for the breast cancer DB.

4 Conclusions

This paper presents a new PS algorithm known as FKINN algorithm. To verify its performance, five UCI benchmarking databases used in Antiya's [11] comparative research are considered. Cross-validation is used to compare the classification of these databases by the FKINN algorithm to the classification by the KNN algorithm, and three indicators—MA, MP, and MTT—are considered. Results of the comparison and the F-test show that the former algorithm performs better than the latter.

Acknowledgments This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the Convergence-ITRC (Convergence Information Technology Research Center) support program (NIPA-2012-H0401-12-1001) supervised by the NIPA (National IT Industry Promotion Agency).

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2044134).

References

1. Cover TM, Hart PE (1967) Nearest neighbor pattern classification. *IEEE Trans Inf Theory* 13:21–27
2. Wu Y, Ianakiev KG, Govindaraju V (2002) Improved K-nearest neighbor classification. *Pattern Recogn* 35(10):2311–2318
3. Sanchez JS, Barandela R, Marques AI, Alejo R (2003) Analysis of new techniques to obtain quality training sets. *Pattern Recogn Lett* 24(7):1015–1022
4. Garcia S, Derrac J, Cano JR, Herrera F (2012) Prototype selection for nearest neighbor classification: taxonomy and empirical study. *IEEE Trans Pattern Anal Mach Intell* 34(3):417–435
5. Amal MA, Riadh BA (2011) Survey of nearest neighbor condensing techniques. (*IJACSA*) *Int J Adv Comput Sci Appl* 2(11)
6. Chang CL (1974) Finding prototypes for nearest neighbor classifiers. *IEEE Trans Comput* 23(11):1179–1184
7. Cervantes A, Galvan IM, Isasi P (2009) AMPSO: a new particle swarm method for nearest neighborhood classification. *IEEE Trans Syst Man Cybern Part B Cybern* 39(5):1082–1091
8. Triguero I, Garca S, Herrera F (2010) IPADE: iterative prototype adjustment for nearest neighbor classification. *IEEE Trans Neural Netw* 21(12):1984–1990
9. Keller J, Gray MR, Givens JA (1985) A fuzzy k-nearest neighbor algorithm. *IEEE Trans Systems Man Cybern SMC-15*(4):406–410
10. Blake CL, Merz CJ (1998) UCI repository of machine learning database. Department of Information and Computer Science, University of California, Irvine
11. Fayed HA, Atiya AF (2009) A novel template reduction approach for the k-nearest neighbor method. *IEEE Trans Neural Netw* 20(5):890–896

Enhanced Reinforcement Learning by Recursive Updating of Q-values for Reward Propagation

Yunsick Sung, Eunyoung Ahn and Kyungeun Cho

Abstract In this paper, we propose a method to reduce the learning time of Q-learning by combining the method of updating even to Q-values of unexecuted actions with the method of adding a terminal reward to unvisited Q-values. To verify the method, its performance was compared to that of conventional Q-learning. The proposed approach showed the same performance as conventional Q-learning, with only 27 % of the learning episodes required for conventional Q-learning. Accordingly, we verified that the proposed method reduced learning time by updating more Q-values in the early stage of learning and distributing a terminal reward to more Q-values.

Keywords Q-learning · Terminal reward · Propagation · Q-value

1 Introduction

Virtual agents execute actions autonomously to achieve their goals in a virtual environment [1]. Q-learning enables learning by defining only a terminal reward without defining an environment model [2]. However, Q-learning requires iterative learning. Therefore the learning time increases.

Y. Sung

Department of Game Engineering, Graduate School, Dongguk University,
26, Pil-dong 3-ga Jung-gu, Seoul, Korea

E. Ahn

Department of Multimedia Engineering, Hanbat National University, San 16-1,
Duckmyeong-dong Yuseong-gu, Deajeon, South Korea

K. Cho (✉)

Department of Multimedia Engineering, Dongguk University,
26, Pil-dong 3-ga Jung-gu, Seoul, Korea
e-mail: cke@dongguk.edu

The following approaches have been used to address the learning time problem in Q-learning. First, approximation functions have been applied [3]. This approach, in which Q-values that are not learned are calculated by using approximation functions, provides methods to select actions without learning all Q-values. Thus, it can reduce the number of Q-values to learn. Another approach involves the participation of a human teacher in learning. For example, the teacher evaluates and evaluate the results of the actions executed [4]. Therefore, the learning time could be reduced. In another approach, the Q-values of both executed actions and similar actions are updated [5]. Given that similar actions have similar learning results, this approach can enhance the effects of learning. In yet another method, the effects of learning are increased by controlling the Q-learning process in reverse [6]. Here, an episode begins with the receipt of a terminal reward and the Q-values are updated in reverse. In this way, a terminal reward can be distributed to multiple Q-values by a single learning episode. A final approach involves the updating of Q-values by selecting those of interest [7]. In this case, the effects of learning are strengthened by selecting and updating in advance the Q-values that have an impact on the selection of actions.

As described above, various approaches have been implemented to reduce the learning time for Q-learning. However, identifying an approach to further reduce the learning time beyond existing approaches is necessary to enable autonomous virtual agents to execute optimum actions in consideration of larger state spaces.

In this paper, we propose a method to reduce learning time by extending the scope of terminal reward distribution when a virtual agent learns by Q-learning in a virtual environment. To increase the learning effect, a method of updating even on the Q-learning of unexecuted actions is applied. Accordingly, the learning time was further reduced by distributing a terminal reward to a more extensive scope of Q-values. The proposed method was verified to reduce the learning time relative to conventional Q-learning in a grid-based environment. As the method of terminal value distribution can be readily integrated with other approaches to improve Q-learning, it can be applied in various fields.

2 Related Work

In conventional Q-learning, one Q-value is updated with the maximum Q-value of the next state when an action is executed. Subsequently, the Q-value of the executed action and the state are updated upon receipt of a single terminal reward. To reduce Q-learning time, identification of an approach to update Q-values by distributing a terminal reward more extensively than conventional methods is necessary.

In $Q(\lambda)$ -learning, multiple Q-values are updated by eligibility [8, 9]. The eligibility of consecutive actions is evaluated when a terminal reward occurs. Then, a terminal reward proportional to the eligibility is added to each Q-value. The learning effect is enhanced by distributing a terminal reward to multiple Q-values depending

on eligibility. Ant-Q(λ) is an extension of Q(λ)-learning to the Ant System [10]. In QV-learning, Q(λ)-learning is integrated with a value function [11].

In a Multi-step action (MSA), multiple actions are bound into a single group and executed in sequence [12]. The Q-values of a group are updated at once when all actions included in the group are executed completely. Therefore, MSA reduces learning time by reducing the number of action selections.

A macro-action is similar to an MSA in that actions are grouped and executed but different in that it treats consecutive actions as a single action [13]. Thus, it selects the to-be-executed action from among the macro-actions defined by multiple actions and primitive actions defined as single actions. When a primitive action is chosen, only one action is executed. When a macro-action is selected, multiple actions included in the macro-action are executed sequentially. Q-values are updated only once when all actions included in a primitive action or macro-action are executed. Thus, the Q-values are updated as if a macro-action is one action.

In conventional Q-learning, the maximum Q-value of the next state and a reward are distributed. However, there is a method of distributing only the terminal reward [14]. In this method multiple adjacent Q-values are distributed when a terminal reward occurs. An adjacent Q-value is one that can receive a terminal reward by executing one action. It reduces learning time by distributing a terminal reward more than conventional Q-learning.

In previous studies, the learning time was reduced by updating more Q-values or distributing a terminal value greater than that of conventional Q-learning. However, as the learning time of Q-learning rapidly increases as a function of the dimension of state space, the learning time still remains a problem to be solved. In this paper, an approach is proposed to solve the learning time problem by integrating previous studies.

3 Bayesian Probability for Action Modeling

Investigation of an approach to distribute Q-values to wider scope is necessary for improving the existing method of distributing a terminal reward. In this section, a method to distribute a terminal reward and an approach to integrate methods distributing Q-values of similar actions are proposed.

The proposed Q-learning method is divided into two stages. The learning stage is identical to that of conventional Q-learning, which identifies a terminal reward. The purpose of the propagation stage is to distribute a terminal reward to more than one Q-value, unlike conventional Q-learning.

The propagation stage distributes a terminal reward in a step-by-step manner to adjacent Q-values without distributing it solely to the Q-value where an action is executed. The adjacent Q-values are those that can receive a terminal reward by executing the actions a limited number of times. During propagation, a terminal reward is first distributed among the Q-values that can receive the terminal reward

by the execution of one action. Next, it updates Q-values that receive a terminal reward by executing actions twice. It updates the δ th Q-value by repeating the update of Q-values. The *Propagation Scope* δ is defined before learning starts. The greater the value of δ , the more Q-values are updated although the more the calculation amount increases. However, lower values of δ reduce the amount of calculation as well as the number of to-be-updated Q-values. δ is determined experimentally. The reward propagation method (RPM) is defined as such a Q-value updating approach.

The learning results from the propagation stage should converge to states for a terminal reward to be received. Accordingly, the more the number of executed actions increases on the basis of the states to receive a terminal reward, the more the to-be-distributed values are reduced. Then, as shown in Eq. (1), the to-be-distributed terminal reward r is reduced by repeated multiplication of the discount factor γ , where α is a learning rate and r^m is the reward of the Q-value receiving the terminal reward by executing an action m times.

$$r^m \leftarrow \alpha \times \gamma^{m-1} \times r \quad (1)$$

If only a terminal reward is distributed without Q-values, the optimum action cannot be executed because of the difference in Q-values of the action executed. Then, as shown in Eq. (2), the difference between two Q-values distributed is also distributed. Q^m means the Q-value receiving the terminal reward by m times of action. β is the discount factor of Q-value.

$$r^m \leftarrow \alpha \times (\gamma^{m-1} \times r + \beta \times Q^{m-1} - Q^m) \quad (2)$$

Finally, a terminal reward cannot be distributed to all Q-values because of the amount of calculation required. The terminal reward distribution is determined by δ , as shown in Eq. (3).

$$\begin{aligned} &\text{IF } m \leq \delta \text{ THEN} \\ &\quad r^m \leftarrow \alpha \times (r^m + \beta \times Q^{m-1} - Q^m) \\ &\text{ELSE} \\ &\quad r^m \leftarrow 0 \\ &\text{END IF} \end{aligned} \quad (3)$$

4 Experiment

To validate the proposed method, RPM, a hunter-prey capture game was used where the size of the grid environment for experiments was a 12×12 cells without walls. In the experiments, we set α to 0.1, β to 0.9, ε to 0.001, and γ to 0.1.

We conducted experiments by controlling δ as described below. First, the success rate of actions in which a prey moved to the goal to avoid a hunter was compared by

Fig. 1 Success rate of conventional Q-learning and RPM

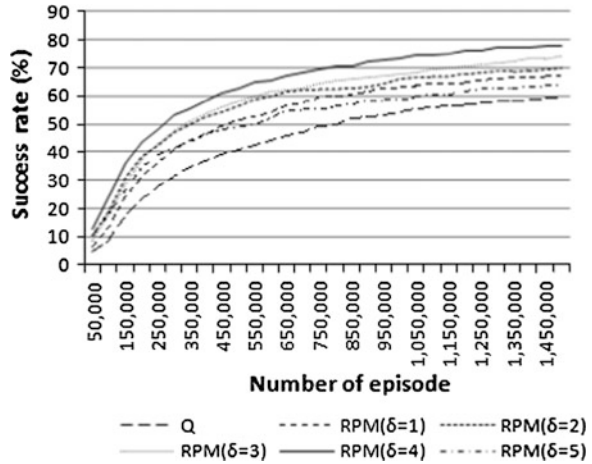
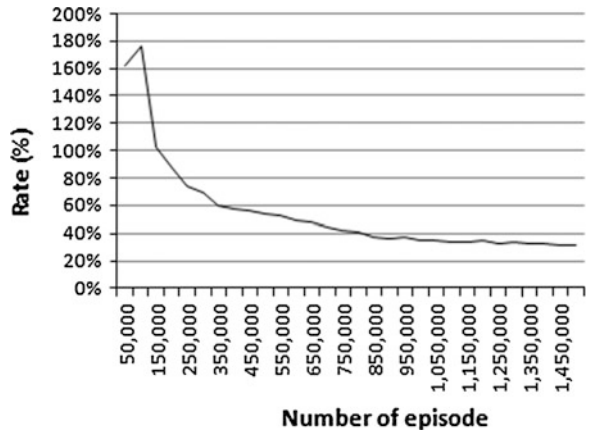


Fig. 2 Comparison between Q-learning and RPM



changing the propagation scope δ from 1 to 5. Figure 1 shows the experimental results. The best result was obtained when the propagation scope δ was four by with 1.5 million learning episodes. Next, conventional Q-learning was also examined with the same parameters. As shown in Fig. 1, the success rate was lower in conventional Q-learning than RPM.

Figure 2 compares the best result ($\delta = 4$) by the proposed method and the results of conventional Q-learning. RPM showed superior results up to 172 % of conventional Q-learning at the beginning of learning. Furthermore, the difference between two learning methods decreased as learning continued.

The experiment verified that better learning results can be obtained with the proposed approach than with conventional Q-learning by updating more Q-values. In particular, the difference in learning results at the beginning stage was significant.

5 Conclusion

In this paper, an approach to update multiple Q-values while reducing the learning time of conventional Q-learning is proposed. The proposed RPM increases the learning performance at the beginning by repeating the distribution of terminal rewards to adjacent Q-values. Furthermore, iterative learning is suitable for difficult conditions. The learning time is reduced by reducing the number of learning episodes.

In an experiment, we compared the learning times of a virtual agent trained with conventional Q-learning and with the proposed approach. The proposed approach achieved the same success rate as conventional Q-learning after only 27 % of the learning episodes required with conventional Q-learning. Accordingly, the reduction in learning time was experimentally verified.

Acknowledgments This work was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2011-0011266).

References

1. Sung Y, Cho K (2012) Collaborative programming by demonstration for human, robot, and software agent team members in a virtual environment. *IEEE Intell Syst* 27(2):674–679
2. Watkins CJCH, Dayan P (1992) Q-learning. *Mach Learn* 8:279–292
3. Melo FS, Ribeiro MI (2007) Q-learning with linear function approximation. In: *Learning theory: 20th annual conference on learning theory, Lecture notes in artificial intelligence (LNAI)*, vol 4539, pp 308–322
4. Thomaz AL, Hoffman G, Breazeal C (2006) Reinforcement learning with human teachers: understanding how people want to teach robots. In: *the 15th IEEE International Symposium on Robot Hum Interact Commun* pp 352–257
5. Jeong SI, Lee YJ (2001) Fuzzy Q-learning using distributed eligibility. *J Fuzzy Log Intell Syst* 11(5):388–394
6. Kormushev P, Nomoto K, Dong F, Hirota K (2008) Time manipulation technique for speeding up reinforcement learning in simulations. *Int J Cybern Inf Technol* 8(1):12–24
7. Moore AW, Atkeson CG (1993) Prioritized sweeping: reinforcement learning with less data and less time. *Mach Learn* 13:103–130
8. Jeong SI, Lee YJ (2001) Fuzzy Q-learning using distributed eligibility. *J Fuzzy Logic Intell Syst* 11(5):388–394
9. Singh S, Sutton RS, Kaelbling P (1996) Reinforcement learning with replacing eligibility traces. *Mach Learn* 22:123–158
10. Lee SG (2006) A cooperation online reinforcement learning approach in Ant-Q. *Lecture notes in computer science (LNCS)* 4232, pp 487–494
11. Wiering MA (2004) QV(λ)-learning: a new on-policy reinforcement learning algorithm. *Mach Learn* 55(1):5–29
12. Peng J, Williams RJ (1994) Incremental multi-step Q-learning. *Mach Learn* 226–232
13. McGovern A, Sutton RS, Fagg AH (1997) Roles of macro-actions in accelerating reinforcement learning. In: *Grace Hopper celebration of women in computing*, pp 13–18
14. Kim BC, Yun BJ (1999) Reinforcement learning using propagation of goal-state-value. *J Korea Inf Process* 6(5):1303–1311

Improved Method for Action Modeling Using Bayesian Probability Theory

Yunsick Sung, Kyhyun Um and Kyungeun Cho

Abstract The technical development of service robots has enhanced the variety of services provided by them to human beings. Service robots need to interact with human beings; hence, they require considerable learning time. The learning time can be reduced by adopting a learning approach in a virtual environment. To this end, it is necessary to describe a human being's movements in the virtual environment. In this paper, we propose a method to generate an action model of a virtual character by calculating the probability of human movements using Bayesian probability. The virtual character selects actions based on the action model, and it executes these actions. Using the proposed method, the path of a virtual character was decreased by around 74 %, as compared to related methods based on Bayesian probability.

Keywords Programming by demonstration · Bayesian probability · Service robot · Virtual environment

1 Introduction

The technical development of service robots can enable them to provide a wide variety of services. For example, in a kitchen, a robot can provide tools required for cooking via prior observation of the cooking process by a human being [1]. Further, a robot can guide visitors at an exhibition.

Y. Sung

Department of Game Engineering, Graduate School, Dongguk University,
26, Pil-dong 3-ga Jung-gu, Seoul, Korea

K. Um (✉) · K. Cho

Multimedia Engineering, Dongguk University, 26, Pil-dong 3-ga Jung-gu,
Seoul, Korea

e-mail: khum@dongguk.edu

A service robot is designed to directly interact with human beings; hence, it requires human interaction for interaction learning. Therefore, interaction learning is a time-consuming process. In order to reduce the learning time, it is necessary to find a substitute for human beings in the learning process of a service robot.

In this paper, we propose an approach whereby a virtual character learns human movements, and thus, facilitates the interaction learning process of a service robot in a virtual environment. The virtual character calculates Bayesian probability via programming by demonstration (PbD) [2]. Next, the action model is generated on the basis of the calculated probability. An experiment is conducted to verify the process that enables a predecessor to control the virtual character in a virtual environment. Thus, the virtual character learns and executes actions by modeling the daily activities of a human being.

The remainder of this paper is organized as follows. In Sect. 2, we provide an overview of previous studies on action modeling using Bayesian probability. In Sect. 3, we propose a method for modeling human movement using Bayesian probability. In Sect. 4, we describe an experiment for verifying the modeling process in a virtual environment. Finally, in Sect. 5, we present the conclusions.

2 Related Work

In PbD studies, virtual characters execute actions by observing those executed by a predecessor. One example is a method for selecting actions on the basis of Bayesian probability [3]. However, such a method does not allow a virtual character to execute consecutive actions like a human being because the virtual character considers only current state. Thus, the executed consecutive actions do not appear natural like those of human beings. In order to overcome this problem, the Bayesian probability is calculated on the basis of the previous action [4] and previous states [5]. However, such a method has a limited scope because it considers only the actions executed immediately before the selected actions. Therefore, it is necessary to develop a method for selecting actions on the basis of more states and actions, specifically, consecutive states and actions.

In this paper, we propose a method for generating the action model by considering more states and actions than those considered in other methods based on Bayesian probability. In our method, a virtual character learns by classifying states and actions in greater detail, as compared to other methods, and then it executes the actions with the highest probability calculated on the basis of the learning.

3 Bayesian Probability for Action Modeling

In this section, we describe the action modeling of a virtual character in a virtual environment, which serves as substitute for a human being. A virtual character that interacts with a service robot in a virtual environment is referred to as a virtual

human. The virtual human learns the daily activities of a human being from a predecessor. Then, under the control of the predecessor, the virtual human generates the probability model required to execute the actions based on the observed consecutive actions and states.

$$P(a_t = a \mid s_t = s, \dots, s_{t-m} = s, g_t = g, a_{t-1} = a, \dots, a_{t-n} = a), \quad (1)$$

$$0 \leq m \leq t, 0 \leq n \leq t - 1.$$

where action a_t is to be executed at time t . The virtual human selects and executes the action with the highest probability from among the actions executed from time $t-n$ to $t-1$, by considering a goal time t and states from time $t-n$ to t . Under the control of the predecessor, the virtual human observes the executed actions more precisely by considering consecutive actions and states, and it selects the actions on the basis of the observations. The probability is calculated by considering only consecutive states, only consecutive actions or both consecutive states and consecutive actions in the virtual learning environment. As compared to other related methods, the proposed method can select actions more flexibly by calculating the Bayesian probability in a variety of ways.

4 Experiment

A virtual house was configured for the experiment. The virtual human was controlled by a human being, after which the Bayesian probability was calculated. The approach proposed in this paper is compared with Thureau's approach [4]. Thureau considered the previous action a_{t-1} as well as the state s_t and goal g_t for executing consecutive actions of a human being. The approach proposed in this paper can calculate the probability in various ways. In order to make a clear distinction between our approach and Thureau's approach, the Bayesian probability for the proposed approach is calculated as shown in the equation (2). Thus, when selecting a_t , the previous state s_{t-1} is considered without taking the previously executed actions into account.

$$P(a_t = a \mid s_t = s, s_{t-1} = s, g_t = g). \quad (2)$$

A virtual human executes actions after learning the daily activities of a human being, as it moves around the entire space of a virtual environment. Figure 1 shows a part of the virtual house and the movement of a virtual human in each approach. First, the virtual human starts from the location P, as shown in Fig. 1a. A human being teaches the virtual human by moving it as follows: left (\rightarrow), left (\rightarrow), left (\rightarrow), and down (\downarrow). The virtual human fails to move in certain directions because of the presence of a wall. Second, the virtual human learns according to Thureau's approach and the approach proposed in this paper, as shown in Fig. 1b and c, respectively.

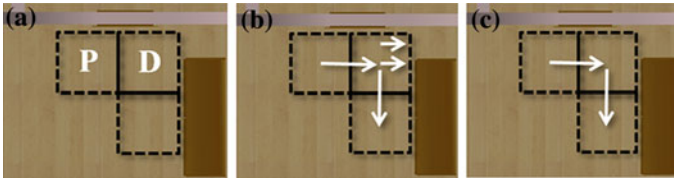


Fig. 1 Limitations of Thureau's approach. The method proposed in this paper can shorten the path of a virtual human by executing actions based only on the state. **a** Location D, P **b** Thureau's approach **c** Approach proposed in this paper

The learning results show that the same actions were executed by a human being, whereas the learning results differed according to the learning approach. Under Thureau's approach, the virtual human moves same as a human being. Although its movements are not as efficient as those of a human being, it tries to move toward the wall. Thus, the virtual human executes four actions in Thureau's approach. In contrast, the approach proposed in this paper enables the virtual human to move by executing only two actions. Accordingly, the number of actions of the virtual human were optimized, and hence, reduced.

5 Conclusion

In this paper, we proposed a method for improving related methods that generate an action model of a virtual human on the basis of Bayesian probability. The action selection probability could be calculated according to the configuration of the virtual environment. The actions executed by a predecessor were observed in detail, and previously executed consecutive actions and consecutive states were considered.

We conducted an experiment to teach the daily activities of human beings to a virtual human as a substitute for a human being for interaction learning. The distance covered by the virtual human was around 74 % less, as compared to related Bayesian probability methods. We believe that the interaction learning method of a virtual robot requires further investigation.

Acknowledgments This work was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A2009148).

References

1. Fukuda T, Nakauchi Y, Noguchi K, Matsubara T (2005) Sequential human behavior recognition for cooking-support robots. *J Robotics Mechatron* 17(6):717–724
2. Cypher A (1993) *Watch what I do: programming by demonstration*. MIT Press, Cambridge

3. Rao RPN, Shon AP, Meltzoff AN (2004) A Bayesian model of imitation in infants and robots, imitation and social learning in robots. Humans, and animals. Cambridge University Press, Cambridge, pp 217–247
4. Thureau C, Paczian T, Bauckhage C (2005) Is Bayesian imitation learning the route to believable gamebots. In: Proceedings of GAME-ON North America, pp 3–9
5. Gorman B, Thureau C, Bauckhage C, Humphrys M (2006) Bayesian imitation of human behavior in interactive computer games. In: Proceedings of 18th international conference on pattern recognition, vol 1, pp 1244–1247

Decision Tree Driven Rule Induction for Heart Disease Prediction Model: Korean National Health and Nutrition Examinations Survey V-1

Jae-Kwon Kim, Eun-Ji Son, Young-Ho Lee and Dong-Kyun Park

Abstract Heart disease has the highest rates of death in non-communicable disease and there have been much research on heart disease. Even though there is recognition for importance of heart disease prediction, related studies are insufficient. Therefore, to develop heart disease prediction model for Korean, we suggest data mining driven rule induction for heart disease prediction in this paper. Proposed method suggest heart disease prediction model by applying decision tree driven rule induction based on data set from Korean National Health and Nutrition Examinations Survey V-1 (KNHANES V-1). The prediction model is expected contribute to Korea's heart disease prediction.

Keywords Data mining · Heart disease prediction · Decision tree · Rule induction · KNHANES V-1

J.-K. Kim

School of Computer Science and Information Engineering, Inha University,
253, Yonghyun-dong, Nam-gu, Incheon, South Korea
e-mail: jaekwonkorea@naver.com

E.-J. Son · Y.-H. Lee

School of Information Technology, Gachon University, 191, Hambakmoero,
Yeonsu-gu, Incheon, South Korea
e-mail: a_projangA@naver.com

Y.-H. Lee

e-mail: lyh@gachon.ac.kr

D.-K. Park (✉)

u-Healthcare Center, Gachon University Gil Hospital, 1198, Guwold-dong,
Namdong-gu, Incheon, South Korea
e-mail: pdk66@gilhospital.com

1 Introduction

Heart disease is continually a major cause for death in spite of costly and steady management. Prediction of heart disease can reduce health care costs and required for future national health promotion [1]. The most typical guideline predicting heart risk factor is Framingham Risk Score [2]. However, this guideline is not suitable for Korean. There are aware of the importance for the prediction of heart disease in Korea but the studies are lacking. Therefore, development of heart disease prediction model for Korean is necessary [3].

Data mining explores hidden rule or pattern and is analytical method for discovering new knowledge. Also heart disease prediction model using artificial neural network, fuzzy method and so on is showing high accuracy [4].

In this paper, we suggest data mining driven rule induction model for heart disease prediction of Korean. Proposed method suggest Korean's heart disease prediction model using decision tree driven rule induction. Decision tree is an analytical method of classification and prediction which is frequently used for clinical disease prediction [5]. In this paper, to develop heart disease prediction model for Korean, we used 5th Korean National Health and Nutrition Examinations Survey (KNHANES V-1) [6], and suggest prediction model using decision tree method, C 5.0.

2 Related Works

Data mining is widely used for clinical medical data analysis and there are many studies in progress to predict heart disease. Vahid Khatibi [4] generated guideline of FRS and PROCAM, suggesting rulebase using Dempster-Shafer evidence theory, and suggests heart disease prediction model using fuzzy method. Sanz [4] predicted heart diseases using genetic algorithm and fuzzy method which referred to guideline of FRS. Erstwhile prediction model is not suitable for Korean because the model refers to FRS guideline. Therefore, for Korean's heart disease prediction, mining model based on Korean clinical data is required.

3 Rule Induction

3.1 Data Set

In this paper, we used 5th Korean National Health and Nutrition Examinations Survey (KNHANES V-1, 2010) from Korea Centers for Disease Control and Prevention. KNHANES V-1 researched people's health and nutrition levels by National Organization statistics and provided a basis for policy formulation and

Table 1 General characteristics

Attributes	Descriptions	Types	Values
Sex	Sex	Flag	1 = male; 2 = female
Age	Age	Range	[20,82]
marri_1	Married	Flag	1 = married; 2 = Single
DE1_lt	Diabetes status	Flag	1 = Diabetes; 2 = absent
BS3_1	Smoking status	Set	0 = not Smoke; 1 = Smoking; 2 = Sometime; 3 = Past
HE_sbp_tr	Systolic blood pressure	Range	[89,178]
HE_dbp_tr	Diastolic blood pressure	Range	[56,117]
HE_ht	Height	Range	[142,186]
HE_wt	Weight	Range	[42,130]
HE_chol	Total Cholesterol	Range	[109,356]
HE_HDL	HDL cholesterol	Range	[28,102]
HE_LDL_drct	LDL cholesterol	Range	[39,243]
D11_pr; D12_pr; D13_pr; D15_pr; D16_pr	CHD diagnosis (Output)	Flag	0 = absent; 1 = present

evaluation such as National Health Promotion Plan [6]. Among 8019 total respondents, 6,336 uncertain respondents, 23 of subjects under 20 years of age, 1,228 respondents who did not respond heart disease related survey were excluded.

Finally, 432 cases were selected, input variables were total 12, and output variables were total 5. Output variable was the prevalence of hypertension, hyperlipidemia, stroke, myocardial infarction and angina. If there were more than one disease prevalence, we defined as heart disease (Table 1).

3.2 Decision Tree Based Rule Induction

Decision tree is one of classification method consist of root node, internal node, leaf of terminal node and expressed by type of If A, then B, Else C. C 5.0 is an inductive learning method based on C 4.5 categorized by Entropy. C 5.0 is useful to conduct pruning if decision tree have too many levels and leaf node which improves generalization ability of learning decision tree. Analysis procedures of decision tree follow four steps.

- Step (1) Decision tree creation: Shape a decision tree after considering appropriate separation standards and stopping rule, in accordance with the purpose of analysis and the structure of data.
- Step (2) Pruning: Remove branches having improper inference rules or risk of classification error.

- Step (3) Feasibility Evaluation: Evaluate decision tree through risk chart, profit chart, or cross-feasibility evaluation.
- Step (4) Model Analysis and Prediction: Interpret the decision tree model result and set up the prediction model.

In this paper, prediction model for Korean’s heart disease were created using rule induction of C 5.0. C 5.0 is easily interpreted and analyzed by propositional knowledge creating If–Then Rule. Therefore, the rule was created by applying rule induction algorithm to KNHNES-V1).

4 Experiment and Result

The scenario of the experiment is shown in Fig. 1. Data of 432 people for experiment consist of training data 216(50 %), testing data 216(50 %). Also, C 5.0 induction rule algorithm was applied using Clementine 11.1(SPSS, INC., Chicago, USA). C 5.0 induction rule algorithm option was set-up as Pruning Severity = 75 %, Minimum records per child branch = 2. To reduce error rate of rule, boosting was designed for ten circuits. Figure 2 shows generated rule through training data.

The rule predicting heart disease was total 15, each of the rules indicates status if disease prevalence. The rule evaluated accuracy; data set of proposed model and FRS guideline were compared [2]. The evaluation results are shown in Table 2.

FRS had low accuracy because it is not based on Korean’s heart disease prediction. On the other hand, proposed model had high accuracy because it used Korean data set. Therefore, proposed model can be a significant model on Korean’s heart disease prediction.

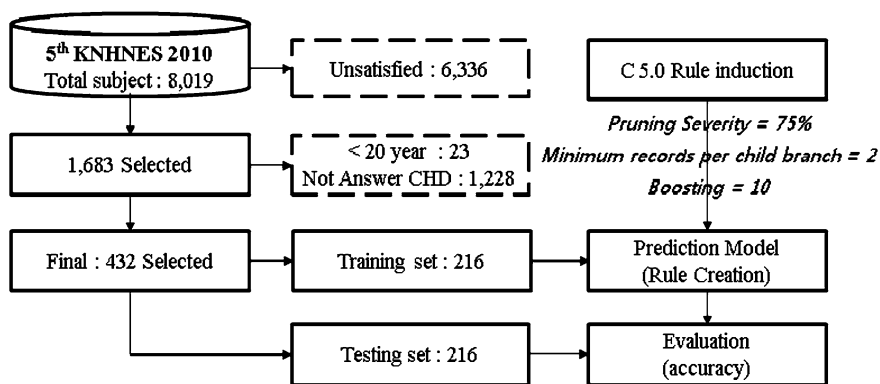


Fig. 1 Experiment scenario

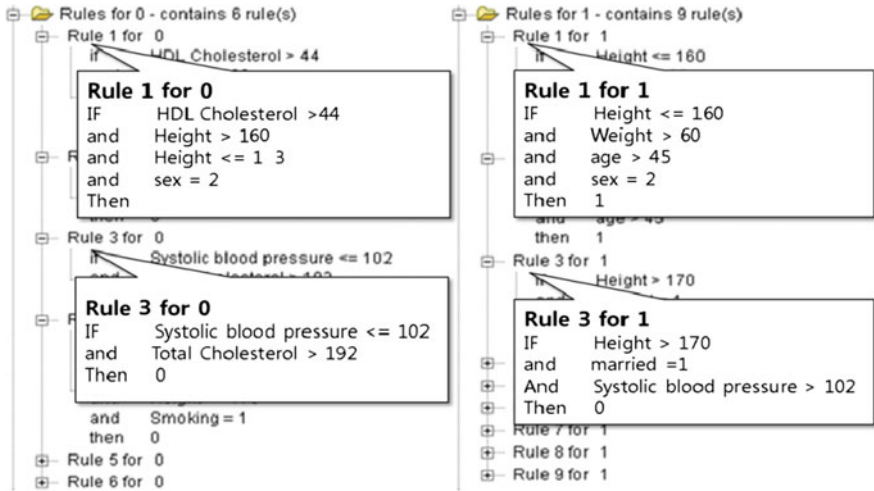


Fig. 2 The result of the generate decision tree detailed rules

Table 2 Experiment result

	FRS guideline		Propose (testing set)	
	Male	Female	Male	Female
Accuracy (%)	53.5	31.9	82.2	83.5

5 Conclusion

Heart disease prediction is very important, and requires prediction guideline for Korean. In this paper, we suggested data mining driven heart disease prediction. Proposed model used data set of Korean National Health and Nutrition Examinations Survey and decision tree driven C 5.0 rule induction algorithm. As the experimental results, proposed model had higher accuracy than abroad guideline. Proposed prediction model is expected to contribute to Korea heart disease prediction.

Acknowledgments This work was supported by the R&D Program of MKE/KEIT [10032115, Development of Digital TV based u-Health System using AI].

References

1. World Health Organization (2010) The world health report 2008. http://www.who.int/whr/2008/whr08_en.pdf. Accessed Nov 2010
2. Wilson P, D'Agostino R, Levy D, Belanger A, Silbershatz H, Kannel W (1998) Prediction of coronary heart disease using risk factor categories. *Circulation* 97:1837–1874

3. Do Young L, Eun Jung R, Eun Suk C, Ji Hoon K, Jong Chul W, Cheol Young P, Won Young L, Ki Won O, Sung Woo P, Sun Woo K (2008) Comparison of the predictability of cardiovascular disease risk according to different metabolic syndrome criteria of American heart association/national heart, lung, and blood institute and international diabetes federation in Korean men. *J Diabetes Metab* 32(4):317–327
4. Vahid K, Gholam AM (2010) A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment. *Int J Expert Syst Appl* 37(12):8536–8542
5. Anooj PK (2012) Clinical decision support system: risk level prediction of heart disease using decision tree fuzzy rules. *Int J IJRRCS* 3(3):1659–1667
6. Korea Centers for Disease Control and Prevention (2010) 5th Korean national health and nutrition examinations survey (KNHANES V-1). Centers for Disease Control and Prevention, Seoul
7. Pagola M, Bustince H, Brugos A, Fernandez A, Herrera F (2011) A case study on medical diagnosis of cardiovascular diseases using a genetic algorithm for tuning fuzzy rule-based classification systems with interval-valued fuzzy sets. In: *Proceedings of 2011 IEEE symposium on advances in T2FUZZ*, pp 9–15

Data Mining-Driven Chronic Heart Disease for Clinical Decision Support System Architecture in Korea

Eun-Ji Son, Jae-Kwon Kim, Young-Ho Lee and Eun-Young Jung

Abstract We present Clinical Decision Support System (CDSS) architecture to implement extensible and interoperable clinical decision support service in perspective of heart study using data mining. In our architecture, intelligence agent engine is critical component for implementing intelligent service using data mining. In this paper, we suggested Fuzzy logic driven Heart risk factor Prediction Model (FHPM) architecture in CDSS. In this CDSS architecture, components for intelligent service with missing value processing logic, Fuzzy linguistic and rule induction method are consisted. FHPM can create chronic heart disease guideline using Korean Data set. FHPM can provide clinical decision support services for the heart disease prediction for Korean.

Keywords CDSS · Data mining · Fuzzy logic · Heart disease · FHPM

E.-J. Son · Y.-H. Lee

School of Information Technology, Gachon University, 191, Hambakmoero, Yeonsu-gu, Incheon, South Korea

e-mail: a_projangA@naver.com

Y.-H. Lee

e-mail: lyh@gachon.ac.kr

J.-K. Kim

School of Computer Science & Information Engineering, Inha University, 253, Yonghyun-dong, Nam-gu, Incheon, South Korea

e-mail: jaekwonkorea@naver.com

E.-Y. Jung (✉)

U-Healthcare Center, Gachon University Gil Hospital, 1198, Guwold-dong, Namdong-gu, Incheon, South Korea

e-mail: eyjung@gilhospital.com

1 Introduction

Globally, chronic heart disease has the highest rates among all death rates in non-communicable disease and it is accelerating [1]. Effort are continued to prevent heart disease. Recently, there are many researches that cover management and prediction of heart disease by combining ICT technology [2].

CDSS, a system to provide necessary knowledge and support correct decision making when diagnosing or planning treatment, also able to predict heart disease risk by taking advantage of the existing hospital information system and ICT (Information Communication Technology) technology [3].

Even though there are needs for Korea cardiovascular prediction guideline, foreign guidelines are used because there is no domestic guideline. Therefore, cardiovascular disease prediction guideline for Korean is requested. In this paper, we suggested FHPM (Fuzzy logic driven Heart risk factor Prediction Model) architecture in CDSS. In our architecture, intelligence engine is critical component for implementing intelligent service using data mining. In this engine, components for intelligence service with missing value processing logic, Fuzzy linguistic and rule induction method are consisted. FHPM create heart disease guideline using Korean data set.

2 Relative Work

Heart disease is critical illness that directly affects patients' life. Typical method that predicts heart risk factor are Framingham Risk Score (FRS) and Prospective Cardiovascular Munster (PROCAM) guideline. However, these are unsuitable for heart risk prediction in Korea because these studies did not consider Korean patients. A guideline to predict Koreans' heart disease is needed because currently it is insufficient [4]. Therefore, the study using ICT technology such as a statistic technique, data mining, artificial intelligence is needed for predicting Korea's heart disease. Anooj [1] suggested Fuzzy Rule through rule induction using Decision Tree in order to offer CDSS and risk level prediction. Khatib [5] used FRS and PROCAM guidelines by Dempster-sharfer evidence theory, designing fuzzy-evidential hybrid inference engine, and suggested CHD risk prediction model. Therefore, in this paper, we designed model to predict Korea's heart risk and CDSS to provide service between physician and patient such case-by-case basis.

3 FHPM Architecture

3.1 Data Model

Typical guidelines detecting heart disease are FRS, PROCAM and ATP III that allow searching 10 years heart risk factor. Therefore, factors from the guidelines are explained in the following Table 1.

Because the guidelines didn't consider Asian's prescription, Fifth Korean National Health And Nutrition Examination Survey (KNHANES V-1) [6] in order to suggest Korean prediction model. Also, this paper suggests the rule using induction for Korean guideline.

The input variables are divided into two groups, separated into categorical data set and continuous data set. Prediction model was learned through learning set of output variables as data set related cardiovascular of KNHANES V-1. Result set is output of finished prediction model from the process.

3.2 CDSS Architecture

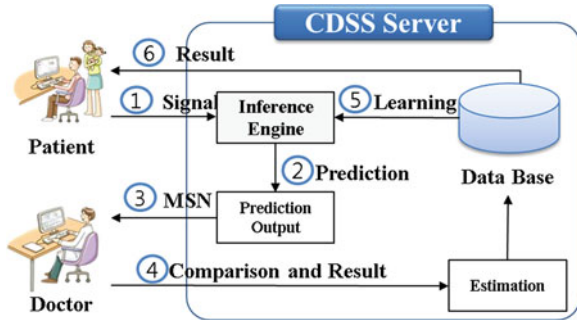
CDSS architecture which can measure patients' heart risk using FHPM data set, consist of three steps as follows (Fig. 1).

- Step 1 Measure own physiological signal such as personal age, HDL, LDL, blood pressure, and smoking status every day
- Step 2 Inference engine evaluate the prediction result through inferring physiological signal
- Step 3 Prediction evaluation results transmitted to physician's mobile device
- Step 4 Physician evaluates the prediction result in 5 steps (healthy, exercise needed, caution, danger, very dangerous)
- Step 5 Evaluated data are stored in a database, and learned to an intelligent inference engine
- Step 6 Patients confirm the prediction results

Table 1 Data set

Group	Data	
Input	Categorical Data set (5)	Sex, Smoking, Diabetes, Education, Married
	Continuous Data set (8)	Age, Weight, Height, Total Chol., LDL Chol., HDL Chol., Systolic BP, Diastolic BP
Output	Training set (5)	Hypertension, Hyperlipidemia, Stroke, Myocardial Infarction, Angina pectoris
	Result set (1)	Heart Risk Level

Fig. 1 CDSS Architecture



Inferred risk factors from inference engine inform 10 years heart risk factor through the current health status. Therefore, it evaluates patient’s health status and how it’s going to affect after 10 years.

3.3 Inference Engine Architecture

Intelligent inference engine is suggested referred to data model. FHPM architecture is shown in Fig. 2.

Fifth KNHNES data set is classified into training set and testing set. Rule induction is defined by training set in the form of IF-THEN. Defined rule was stored in fuzzy rule. Missing value processing module is step detecting effectiveness of input data, detecting a faulty data or incorrectly entered data. Continuous data is transmitted in fuzzy inference engine and produces results using fuzzy rule and fuzzy linguistic function. Fuzzy linguistic function consists of function based on evidence. After Inference of continuous data completed, defuzzification method is used with categorical data. Defuzzification use mamdani fuzzy inference. After inference is completed, Heart Risk Level is shown.

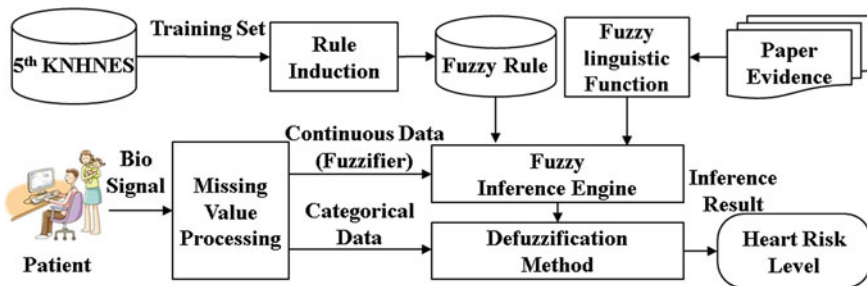


Fig. 2 Fuzzy logic driven heart risk factor prediction architecture

4 Conclusion

In this paper, we designed CDSS architecture for predicting heart disease. The proposed CDSS architecture is data mining driven intelligent systems through diagnosis and prediction. The architecture evaluates heart risk index not only relying on computer, but 5 steps evaluation by clinician. Proposed CDSS architecture presents Fuzzy logic driven Heart risk factor Prediction (FHPM) architecture for intelligent service, which enables to evaluate heart risk level of patients. FHPM uses Korean data and suggest fuzzy rule by rule induction and provide basis to Korea guide line.

Also, when patient enter incorrect value, it is able to detect through processing of missing value. Finally the architecture is able to measure patient's heart risk level by data inference using heart risk data of fuzziness through fuzzy inference engine. The architecture, which is an intelligence system combining ICT technology, is able to provide clinical decision support service on Korean heart disease patients.

Acknowledgments This study was supported by a grant of the Korean Health Technology R&D Project, Ministry of Health & Welfare, Republic of Korea (A11202).

References

1. World Health Organization (2010) The world health report 2010. http://www.who.int/whr/2010/whr10_en.pdf. Retrieved Nov 2010
2. Anooj PK (2012) Clinical decision support system: risk level prediction of heart disease using decision tree fuzzy rules. *Int J IJRRCS* 3(3):1659–1667
3. Garg AX, Adhikari NKJ, McDonald H et al (2005) Effects of computerized clinical decision support systems on practitioner performance and patient outcomes. *J Am Med Assoc* 293(10):1223–1238
4. Lee DY, Rhee EJ, Choi E et al (2008) Comparison of the predictability of cardiovascular disease risk according to different metabolic syndrome criteria of American heart association/national heart, Lung, and Blood institute and international Diabetes federation in Korean men. *J Korean Diabetes* 32(4):317–327
5. Vahid K, Gholam AM (2010) A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment. *Int J Expert Syst Appl* 37(12):8536–8542
6. Korea Centers for Disease Control and Prevention (2010) 5th Korean national health and nutrition examinations survey (KNHANES V-1). Korea Centers for Disease Control and Prevention

A Study on the Occurrence of Crimes Due to Climate Changes Using Decision Tree

Jong-Min Kim, Hwang-Kwon Ahn and Dong-Hwi Lee

Abstract In this study, we figured out what relationship the elements (weather, temperature, precipitation, wind speed, humidity) of meteorological changes have with the incidence of the five violent crimes through data mining. For the data used in this study, the number of meteorological occurrences from January 1, 2011 to March 30, 2012 through portal sites and the elements of meteorological changes of the day recorded in the Korea Meteorological Administration were used as materials. In this study, an analysis was made using the C4.5 algorithm of decision tree to verify what crimes occur according to the elements of the climate change. As a result of such an analysis, most of the crimes were high in the incidence in the following meteorological conditions: when the weather is cloudy; when the temperature is more than 9 °C; when the precipitation is less than 10 mm; when the wind speed is less than 4 m/s; and the humidity is more than 50 %. Given these meteorological conditions, cloudy weather showed the highest rate of crime incidence.

Keywords Decision Tree · J48 Algorithm · C4.5 Algorithm · Crime · WEKA

J.-M. Kim (✉) · D.-H. Lee (✉)

Department of Industry Security, Kyonggi University, Chungjeongno 2-ga,
Seodaemun-gu, Seoul, South Korea
e-mail: dyuo1004@gmail.com

D.-H. Lee

e-mail: dhclub@naver.com

H.-K. Ahn (✉)

Department of Protection & Security Management, Kyonggi University, Iui-dong,
Yeongtong-gu, Suwon-si, Gyeonggi-do, South Korea
e-mail: ahk@kyonggi.ac.kr

1 Introduction

Korea enjoys a comfortable life with the development of the society but the incidence of crime increases steadily each year according to the rapid development. Furthermore, a method employed in violent crimes such as rape, murder, etc. is also becoming intelligent and cruel. In order to prevent these crimes, researches on crime have been continuously conducted but most of the researches have focused on demographic and economic variables. It was not until recent years that researches were performed to see what influence physical conditions have an influence on crimes. These studies are becoming an issue in explaining the cause of crime as a situational crime prevention theory that describes the relationship between the crime and the physical environments such as weather, temperature, precipitation, etc. as well as suggesting crime prevention measures [1].

In this study, using the C4.5 algorithm, the WEKA's decision tree, we attempt to review what influence these physical environments (weather, temperature, precipitation, etc.) have on crime depending on changes in meteorological elements.

In [Chap. 2](#), we take a look at the relationship between weather and crime and the C4.5 algorithm as decision tree and quantified the data in the materials and variables presented in [Chap. 3](#). In [Chap. 4](#), we apply the C4.5 algorithm of the decision tree to make conclusions in [Chap. 5](#).

2 Related Study

2.1 *The Relationship between Weather and Crime*

Early criminologists reported that weather has an psychological influence on crime. Thus, the people who live in tropical climates have a mild personality, whereas those who live in temperate climates are aggressive. In other words, they claimed that the high-temperature season affects individuals directly, which leads them to aggressive behavior from time to time [2].

They thought that the reason why the southern region of the United States shows higher homicide rates is because of this. With regards to this, their studies reported that some rapid climatic determinism is high in blacks who came from the high temperatures of the African region. The studies indicated they also show high homicide rates, as African-Americans has taken over such ancestors' temper that is aggressive and impulse—controlled.

Table 1 Researches

Date	Author	Subject
2006	McLean, Iain	Climatic effects on incidence of sexual assault [3]
2002	McCleary, Richard & Chew Kenneth S. Y	Winter is the infanticide season: seasonal risk for child homicide [4]
2000	Rotton, James & Cohn, Ellen G	Weather, Disorderly conduct, and Assaults: from social contact to social avoidance [5]

2.2 Weather Effects and Crime

Weather effects may affect the incidence of crime while weather elements (temperature, humidity, precipitation, etc.) having an impact on personal feelings. The findings of analyzing the relationship between meteorological elements and crime are obtained in various ways depending on researches and another results from the study between meteorological elements and crime are derived showing that temperature and crime has high correlation between them. Especially, it is reported that if the discomfort index is higher, the incidence of the violent 1 five crimes increase. Table 1 shows researches using weather effects and crime.

2.3 Decision Tree

The decision tree is a scheme used for data mining classification, which is to create the pattern that exists between the records by analyzing the previously stored data, which means to make the form of the classification model tree to represent specific attributes by classification. The created classification model is used to classify new records and to predict the value of the class. As the algorithm to make the decision tree, we use the C4.5 algorithm [6].

2.4 C4.5 Algorithm

The C4.5 algorithm is a decision tree revised and developed by Ross Quinlan. The initial version of this algorithm, ID3 (Interactive Dichotomizer 3) algorithm, had a large impact on learning machinery. It is the case that one class belongs to all the inferior-sets so that the training set input to form the decision tree in C4.5. The tree is formed till composed. If 'p' is the probability of a message, the information delivered

to the message is measured using $-\log_2 p$. In S , a set of cases, when a case is randomly selected, the probability which this case belongs to is as follows [7]:

$$\frac{freq(C_i, S)}{|S|} \quad (1)$$

In here, $|S|$ is the number of cases that belong to S , and $freq(C_i, S)$ means the number of cases that belong to C_i in Set S . Thus, the information delivered by this case is as follows:

$$-\log_2\left(\frac{freq(C_i, S)}{|S|}\right) \quad (2)$$

To get the expected Information in Set S , it is recommended to apply a weighted average of information conveyed by each case.

$$info(S) = \sum_i^k \left(\frac{freq(C_i, S)}{|S|} \times \log_2\left(\frac{freq(C_i, S)}{|S|}\right) \right) \quad (3)$$

3 Proposed Method

In this chapter, we describe research targets and variables to apply the decision tree as a data mining technique to data on the incidence of crime according to the changes in the weather.

3.1 Research Targets and Variables

The data used in the decision tree of this study are the elements based on the number of crime incidents from January 1, 2011 to March 1, 2012 and the elements of meteorological changes on the days recorded in the Korea Meteorological Administration.

The use variables included type of crime, weather, temperature, precipitation, wind speed, and humidity. Figure 1 shows these variables arranged in an EXCEL form.

Number	Crime	Weather	temperature	rainfall	wind_speed	humidity
1	murder	rainy	-1.5	6	3.2	71.4
2	murder	overcast	0.8	0	2.9	59.9
3	murder	overcast	1.8	0	2.1	55.3
4	larceny	overcast	5.1	0	5.7	61
5	robbery	overcast	-3.2	0	1.9	73.4
6	robbery	snow	-3.9	0	0.8	81
7	robbery	sunny	0.2	0	2.2	46.1
8	gewalt	snow	-7.7	0	2.0	52
9	larceny	sunny	-1.8	0	3.0	38.8
10	larceny	overcast	-6.2	0	1.9	61.5
11	robbery	overcast	1.1	0	1.8	72.9
12	rape	rainy	4.6	1.5	1.3	72.5
13	rape	sunny	5.4	0	2.0	53.4
14	gewalt	sunny	8.2	0	1.1	63.9
15	murder	sunny	8.2	0	1.1	63.9
16	gewalt	sunny	3.9	0	1.7	54.4
17	gewalt	rainy	5.4	1	2.4	83.4
18	gewalt	snow	0.5	2.3	3.3	74.5
19	gewalt	sunny	0.1	0	2.8	50.6
20	rape	sunny	12.1	0	2.0	40.6
21	rape	rainy	12.3	0	2.3	93.1
22	rape	rainy	11.4	5	2.9	74.3
23	murder	overcast	16.1	2.5	4.5	69.9
24	murder	overcast	9.8	0	4.5	77.9
25	rape	rainy	11.3	0.5	1.1	55.3
26	rape	overcast	17.3	0	1.9	56.1
27	rape	rainy	15.2	2.5	1.1	90.8
28	rape	overcast	18.2	0	2.3	66
29	murder	overcast	16.9	0	1.1	69.1
30	murder	overcast	20.1	0	2.0	53.5

Fig. 1 Variables Data

In order to analyze the above files from WEKA, users need to change them into an ARFF format. Figure 2 shows the data converted into an ARFF format.

3.2 Distribution Form

If users call the file of ARFF from WEKA, the distribution of the variables used in this paper is shown as Fig. 3.

```

@relation Weather_of_crime
@attribute crime {murder, robbery, rape, larceny, , gewalt}
@attribute Weather {sunny, overcast, rainy, snow}
@attribute temperature real
@attribute rainfall real
@attribute wind_speed real
@attribute humidity real
@data
murder,rainy,-1.5,6,3.2,71.4
murder,overcast,0.8,0,2.9,59.9
murder,overcast,1.8,0,2.1,55.3
larceny,overcast,5.1,0,5.7,61
robbery,overcast,-3.2,0,1.9,73.4
robbery,snow,-3.9,0,0.8,81
robbery,sunny,0.2,0,2.2,46.1
gewalt,snow,-7.7,0,2.0,52
larceny,sunny,-1.8,0,3.0,38.8
larceny,overcast,-6.2,0,1.9,61.5
robbery,overcast,1.1,0,1.8,72.9
rape,rainy,4.6,1.5,1.3,72.5
rape,sunny,5.4,0,2.0,53.4
gewalt,sunny,8.2,0,1.1,63.9
murder,sunny,8.2,0,1.1,63.9
gewalt,overcast,3.9,0,1.7,54.4
gewalt,rainy,5.4,1,2.4,83.4
gewalt,snow,0.5,2.3,3.3,74.5
gewalt,sunny,0.1,0,2.8,50.6
rape,sunny,12.1,0,2.0,40.6
rape,rainy,12.3,0,2.3,93.1
murder,rainy,11.4,5,2.9,74.3
    
```

Fig. 2 ARFF

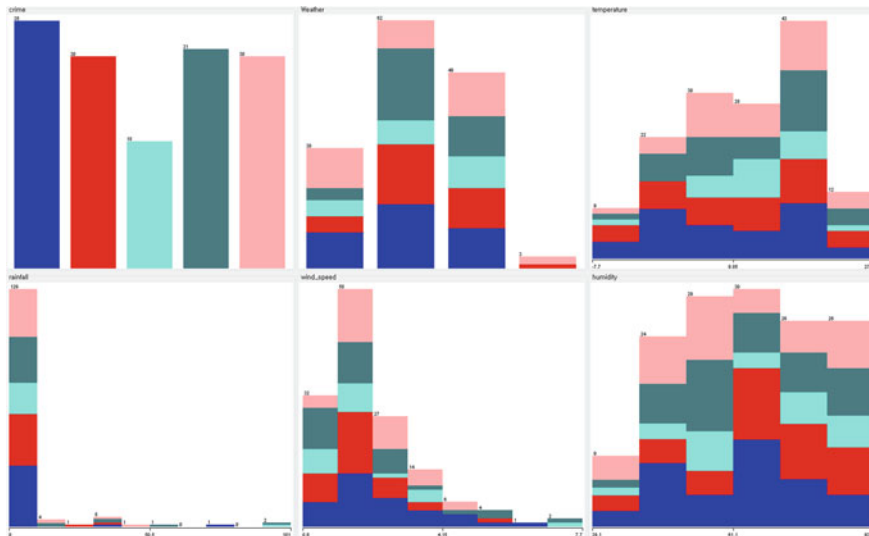


Fig. 3 distribution of the variables

4 Research Results

In this chapter, we made an analysis on what influence meteorological elements have an influence on crime using the C4.5 algorithm as a decision tree.

4.1 Application of Decision Tree, C4.5 Algorithm

ARFF data were applied to a decision tree, C4.5 algorithm. Using “J48-C 0.25 -M 2” through WEKA’s Classify, you can get the results of the C4.5 algorithm as seen in Fig. 4.

Figure 5 shows the visual findings of the C4.5 algorithm .

4.2 Relationship Between Meteorological Changes and Crime by the Decision Tree

The tree used in this study consists of 61 nodes in total, and the ‘Leaves’ node as the last node consists of 32 nodes.

It is interpreted: Sunny weather shows the highest number of assaults in crime type,

Humidity affects crime incidence, compared to other meteorological elements: when the humidity is higher than the basic level 69.9 %, the type of crime changes depending on the properties of wind speed, temperature.

In cloudy weather, murders and assaults occur most frequently. The most influential meteorological variable is precipitation. If the precipitation is higher than the basic level 10 mm, the type of crime changes depending on the properties of wind speed, temperature.

In rainy weather, assaults occur most frequently as seen in the cloudy weather, and the most influential meteorological variable is temperature. If the temperature is higher than the basic level 16.4 °C, the type of crime changes depending on the properties of wind speed, temperature, and humidity.

On snowy days, violence occurs most frequently in terms of crime type.

Fig. 4 J48 cross-validation

```

=== Run information ===
Scheme: weka.classifiers.trees.J48 -C 0.25 -M 2
Relation: Weather_of_prime
Instances: 144
Attributes: 6
    prime
    Weather
    temperature
    rainfall
    wind_speed
    humidity

Test mode: evaluate on training data

=== Classifier model (full training set) ===

J48 pruned tree

Weather = sunny
  humidity <= 86.9
    wind_speed <= 3
      wind_speed <= 1.7: murder (6.0/2.0)
      wind_speed > 1.7
        wind_speed <= 2.1: geweld (7.0/3.0)
        wind_speed > 2.1
          temperature <= -1.5: rape (2.0/1.0)
          temperature > -1.5
            humidity <= 47.5: robbery (2.0)
            humidity > 47.5: geweld (3.0/1.0)
    wind_speed > 3
      temperature <= 3.5: murder (4.0)
      temperature > 3.5: geweld (4.0/1.0)
  humidity > 86.9: larceny (2.0)

Weather = overcast
  rainfall <= 0
    temperature <= 11.3
      humidity <= 87.5
        humidity > 41.6: robbery (2.0/1.0)
        humidity > 41.6: murder (20.0/12.0)
        humidity > 87.5: robbery (5.0/1.0)
      temperature > 11.3
        wind_speed <= 1.8
          temperature <= 16.2: robbery (2.0/1.0)
          temperature > 16.2: larceny (12.0/4.0)
        wind_speed > 1.8
          humidity <= 55.3
            temperature <= 16.7: geweld (4.0/1.0)
            temperature > 16.7: murder (3.0/1.0)
          humidity > 55.3
            temperature <= 21.3
              temperature <= 18.3: rape (3.0)
              temperature > 18.3: murder (3.0/1.0)
            temperature > 21.3: robbery (2.0)
  rainfall > 0: robbery (5.0/2.0)

Weather = rainy
  temperature <= 16.4
    wind_speed <= 1.4: rape (4.0)
    wind_speed > 1.4
      wind_speed <= 2.1
        temperature <= 8.8: larceny (2.0)
        temperature > 8.8
          humidity <= 74.5: geweld (2.0)
          humidity > 74.5: murder (2.0)
      wind_speed > 2.1
        rainfall <= 0.5: rape (3.0)
        rainfall > 0.5
          humidity <= 84.9
            humidity <= 87.5: larceny (2.0/1.0)
            humidity > 87.5: murder (5.0/1.0)
          humidity > 84.9: larceny (3.0/1.0)
        temperature > 16.4
          temperature <= 23.8: geweld (16.0/10.0)
          temperature > 23.8
            temperature <= 24.9: murder (2.0)
            temperature > 24.9: larceny (2.0)

Weather = snow: geweld (3.0/1.0)

Number of Leaves : 32
Size of the tree : 81

Time taken to build model: 0 seconds

=== Evaluation on training set ===
=== Summary ===
Correctly Classified Instances 97          67.3611 %
Incorrectly Classified Instances 47          32.6389 %
Kappa statistic 0.6861
Mean absolute error 0.1623
Root mean squared error 0.2548
Relative absolute error 61.2122 %
Root relative squared error 71.6742 %
Total Number of Instances 144

=== Detailed Accuracy By Class ===


|               | TP Rate | FP Rate | Precision | Recall | F-Measure | ROC Area | Class   |
|---------------|---------|---------|-----------|--------|-----------|----------|---------|
|               | 0.8     | 0.166   | 0.822     | 0.8    | 0.7       | 0.911    | murder  |
|               | 0.833   | 0.093   | 0.78      | 0.833  | 0.891     | 0.927    | robbery |
|               | 0.811   | 0.038   | 0.917     | 0.811  | 0.733     | 0.877    | rape    |
|               | 0.548   | 0.083   | 0.739     | 0.548  | 0.63      | 0.619    | larceny |
|               | 0.733   | 0.148   | 0.684     | 0.733  | 0.639     | 0.81     | geweld  |
| Weighted Avg. | 0.674   | 0.092   | 0.701     | 0.674  | 0.674     | 0.624    |         |



=== Confusion Matrix ===


| a  | b  | c  | d  | e  | ← classified as |
|----|----|----|----|----|-----------------|
| 28 | 2  | 0  | 2  | 3  | a = murder      |
| 4  | 19 | 0  | 0  | 7  | b = robbery     |
| 1  | 0  | 11 | 3  | 3  | c = rape        |
| 8  | 1  | 1  | 17 | 4  | d = larceny     |
| 4  | 3  | 0  | 1  | 22 | e = geweld      |


```

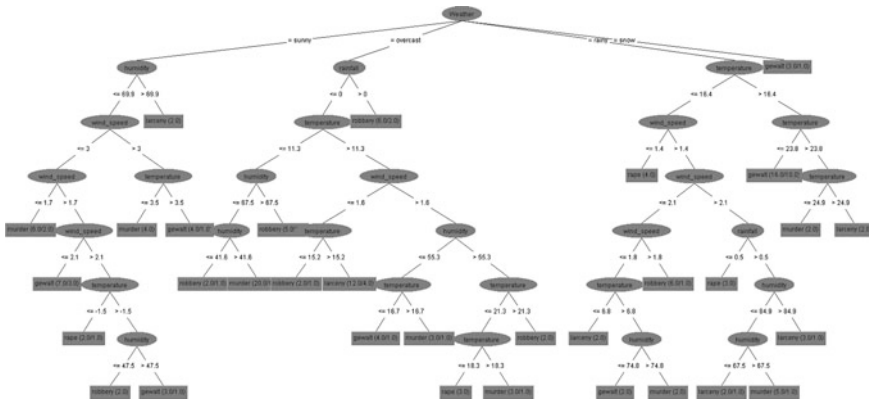



Fig. 5 J48 Visualiz tree

5 Conclusion

In this study, through the J4.5 algorithm of decision tree, we were able to figure out what type of crime (murder, robbery, rape, theft, and violence) occurs depending on meteorological variables (weather, temperature, precipitation, wind speed, and relative humidity). As a result of the analysis, the crime rate was higher in the following weather conditions: less than 10 mm in precipitation, less than 4 m/s in wind speed, and 50 % in humidity. It was observed that in these conditions, the highest incidence of crime was a violence incident and that the highest incidence weather of crime was when weather is cloudy.

In the United States and Europe, a wide variety of researches have been conducted in dealing with the relationship between weather and crime. But in Korea, there has been few studies to analyze the relationship between weather and crime. Therefore, this study is expected to serve an important preceding paper in this field, based on the findings of this study.

Acknowledgments This work was supported by a grant from Kyonggi university advanced Industrial Security Center of Korea Ministry of Knowledge Economy

References

1. Lee Y-H, Kim Y-S (2010) Weather, the day of week, and the number of crime. Korea Assoc Crim Psychol
2. Cheatwood D (2009) Weather and crime. In: Mitchell Miller J (ed.), 21st century criminology: a reference handbook. SAGE Publications, Inc, Thousand Oaks pp 51–58, NcPRAM http://www.samsung.com/global/business/semiconductor/products/fusionmemory/Products_NcPRAM.html

3. McLean Iain (2006) Climatic effects on incidence of sexual assault. *J Forensic Leg Med* 14:16–19
4. McCleary, R Chew KSY (2002) Winter is the infanticide season: seasonal risk for child homicide. *Homicide Stud* 6(3):228–239
5. Rotton J, Cohn EG (2000) Weather, disorderly conduct, and assaults: from Social contact to social avoidance. *Environ Behav* 32(2):651–673
6. Noh C-H, Cho K-C, Ma Y-B, Lee J-S (2009) Grid resource selection system using decision tree method. *Korea Soc Comput Inf* 13(1):1–10
7. Leem Y-M, Kwag J-K, Hwang Y-S (2005) A feature analysis of industrial accidents using C4.5 algorithm. *Korean Soc Saf*

A Case Study for the Application of Storage Tiering Based on ILM through Data Value Analysis

Chun-Kyun Youn

Abstract In the recent, due to explosion of Digital Universe, the performance of computer and storage system is reducing. Therefore, the upgrade and capacity expansion needs is growing. Countermeasure for this problem is required fundamental and long-term solutions rather than piecemeal expansion. In this paper, we establish a data management policy for an enterprise through the operational status of storage system and the analysis of data value of it, and implement ILM-based tiered storage system on the basis of these. The results of this study shows the overall throughput was improved about 21 % compared to the existing system, it is very effective to maintain continuous quality and reduce operating costs in the long term aspect.

Keywords ILM · Data value analysis · Storage · Tiering · Data management method

1 Introduction

According to the 2Q survey which was conducted for storage engineers in domestic companies by IT World Korea, 40 % of them have gone through difficulty in data increase and storage management, which is big problems for storage now, and only 29 % of them has separate data storage systems by utility and importance of data. ILM (Information Life-cycle Management) which is

C.-K. Youn (✉)

Department of Internet Contents, Honam University, #330, Eodeungno, Gwangsan-gu, Gwangju, South Korea

e-mail: chqyoun@honam.ac.kr

tiering storages by data value and Tiering technology with low implementation costs are a solution for problems caused by explosive data increase but most of companies except for 10.8 % of major companies don't plan to implement storage tiering or don't know the technologies [1–3].

Therefore, this study would like to suggest the guide to solutions for storage problems, caused by explosive data increase, which many companies are now facing by showing how much and effective they can increase utility of existing storage systems through application of ILM and Tiering for the existing storage systems in companies and analysis of their effects as long-term and basic solutions for storage problems.

2 Analysis of System Operation Environments

The company which I analyzed is a manufacturer which sells and exports its products and provides statistical service for internal report.

2.1 Status of Key Works

Work importance and usage frequency were determined based on number of users in the company and influence when the works were not performed was surveyed through interviews with employees. The results are as described in Table 1.

2.2 Status of Data Management

2.2.1 Attribute and Importance of Data

Data attribute was reviewed based on the work systems in Table 1. Most of work data were created, maintained, and managed through DB and email application. They could be divided largely into structured and unstructured data as described in Table 2.

In this study, data with relatively high I/O requests and high importance are called as “active data” and data with relatively low I/O requests are called as “inactive data”.

2.2.2 Backup and Deletion of Data

Review on policies of data storage and backup to evaluate backup methods for data and deletion process, which was the last step in the life cycles of data, created Tables 3 and 4.

Table 1 Importance of key work systems

Frequency	Importance	No. of users	Work system	Influence	
Over 1 time weekly	Low	Less than 5 % of whole employee	Production history management system	Low productivity	
			Inventory control system		
			Business management system		
			Production statistics system		
			Basic statistics system		
			Assets Management System		
Over 1 time daily	Medium	Less than 25 % of whole employee	Overseas market information system		
			Extranet homepage system		
			Inventory control system		
	High	Less than 50 % of whole employee	Geographic information system		
			More than 75 % of whole employee		Knowledge portal system
					Business email system
Over 1 time daily	High	More than 75 % of whole employee		Personnel management system	Lawsuit
			Electronic approval system		
			Work management system		

Table 2 Data types of key works

Monthly average number of process	Data attribute	System	Host
Less than 100	Structured	Production history Mgm' system	Unix server
		Inventory control system	Unix server
		Production statistics system	Unix server
		Basic statistics system	Unix server
		Assets management system	Unix server
		Extranet homepage system	Window NT server
Less than 1,000	Structured/unstructured	Geographic info' system	Unix server
	Unstructured	Overseas market info' system	Unix server
		Business management system	Unix server
		Knowledge portal system	Window NT server
		Personnel management system	Window NT server
Less than 5,000	Structured/unstructured	Electronic approval system	Unix server
		Work management system	Unix server
	Unstructured	Business email system	Unix server

Table 3 Storage and backup periods of key work systems

Importance of work	Work system	Backup method	Backup interval
High	Business email system	Storage backup (1st)	Incremental backup (daily)
	Knowledge portal system	LTO drive backup (2nd)	
	Personnel management system	Disaster recovery backup (3rd)	Whole backup (weekly)
	Electronic approval system		
	Work management system		
Medium	Extranet homepage system		
	Geographic information system		
	Overseas market information system		
Low	Inventory control system		
	Production history Management system	LTO drive backup (1st)	Whole backup (weekly)
	Business management system		
	Inventory control system		
	Production statistics system		
	Basic statistics system		
	Assets management system		

2.3 Analysis of DB Operation Environments

Features of application workload, overall DB I/O capacity, and area information of table space were analyzed to assess operation environments of DB application using StatsPack, AWR Report, and performance management package. Analysis of workload type of operation DB applications showed that they were typical DBs for decision-making and mass data storage with features of many-read and few-writes as shown in Fig. 1.

The survey of I/O lists in table space showed that a specific table space(-DATA_0, INDEX_0) had high Read Count, 5 and 3 million each, and average response times of the table space were very low, 9 and 5 ms each, as shown in Fig. 2.

2.4 Analysis of Storage Operation Environments

The survey of allocation lists of systems and storages in the company showed that each operation host was connected to storages through dual SAN switches as shown Fig. 3. Major internal components of the storages were consisted of dual 8 GB memory and 146 GB-disk groups. All disk groups are fiber channel type HDD supporting 15,000 rpm.

Table 4 Data storage interval of work systems/past due capacity

Importance of work	Data storage interval	Work system	Allotted capacity (GB)	Capacity to delete (GB)	Percentage (%)
High	5 Years	Business email system	1,200	320	26.7
		Personnel management system	120	10	8.3
		Knowledge portal system	600	400	66.7
		Electronic approval system	300	50	16.7
		Work management system	600	100	16.7
Medium	3 Years	Extranet homepage system	120	0	0.0
		Overseas market information system	240	20	8.3
		Inventory control system	600	60	10.0
		Geographic information system	2,400	800	33.3
Low	2 Years	Production history management system	1,200	450	37.5
		Business management system	120	20	16.7
		Production statistics system	1,200	400	33.3
		Basic statistics system	960	300	31.3
		Inventory control system	600	250	41.7
		Assets management system	720	320	44.4
		Total	-	-	10,980

Load Profile

	Per Second	Per Transaction	Per Exec	Per Call
DB Time(s):	64.8	0.2	0.02	0.14
DB CPU(s):	3.9	0.0	0.00	0.01
Redo size:	2,105,159.7	5,091.2		
Logical reads:	26,125.1	65.2		
Block changes:	10,572.2	25.6		
Physical reads:	9,627.9	23.3		
Physical writes:	3,782.3	9.2		
User calls:	478.6	1.2		
Parses:	2.3	0.0		
Hard parses:	0.0	0.0		

Fig. 1 Physical read and write of DB application

Tablespace	Reads	Av Reads/s	Av Rd(ms)	Av Blks/Rd	Writes	Av Writes/s	Buffer Waits	Av Buf Wrt(ms)
DATA_0	5,268,726	5,846	9.00	1.00	2,091,970	3,209	100	4.20
INDEX_0	3,397,545	3,770	5.18	1.00	170,842	190	1,865	5.43
UNDOTBS1	103	0	1.26	1.00	25,331	28	35	0.00
SYSAUX	703	1	3.12	1.18	357	0	0	0.00
SYSTEM	28	0	8.57	1.00	35	0	0	0.00
UNDOTBS2	1	0	10.00	1.00	1	0	2	5.00
USERS	1	0	10.00	1.00	1	0	0	0.00

Fig. 2 I/O and response time of table space

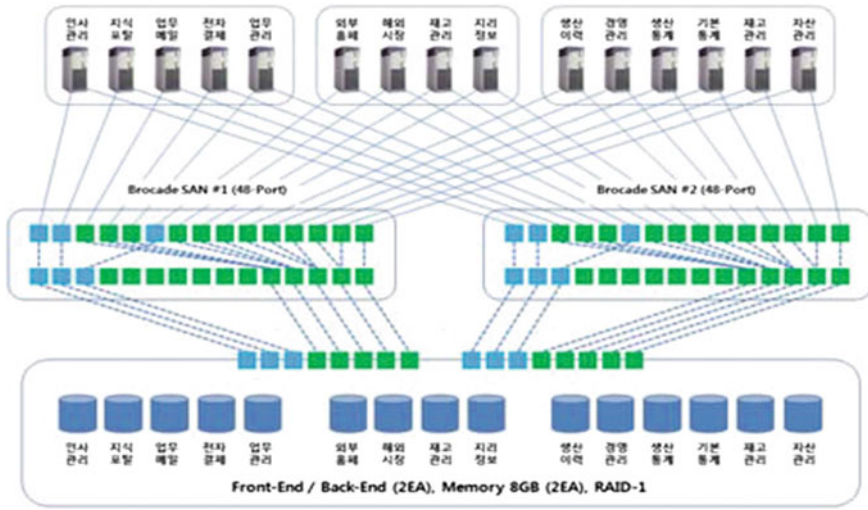


Fig. 3 Volume allocation configuration of operation system

Table 5 Capacity estimation of operation environment of storages

RAID type	Disk group#	RPM (K)	Capacity (GB)	Quantity (EA)	Physical capacity(GB)	LV (GB)	LV (EA)	Actual capacity(GB)
RAID-1 (Mirror)	# 1	15	146	40	5,840	12	243	2,920
	# 2	15	146	60	8,760	12	365	4,380
	# 3	15	146	60	8,760	12	365	4,380
	Total	15	146	160	23,360	-	973	11,680

The analysis of disks capacity showed that physical capacity of the whole disk was 23,360 GB and actual capacity was 11,680 GB (12 GB LV * 973 EA) as described in Table 5. And Table 6 shows the file system and spare volume was estimated about 700 GB.

Figure 4 is a graph showing I/O occurring in all devices in the storages. The left red area is for typical active data which are about 15 % of the whole storage area and whose average occurrence time of I/O is over 100 times and the right green area is for typical inactive data which are about 40 % and less than 10 times.

2.5 Issue of Storage Management and Operation

2.5.1 Systematic Data Management Policy

The issues in the data management policy are as follows:

1. The data management policy as described in Tables 3 and 4 was not applied to actual operation and about 32 % of the whole storage area had data to delete.

Table 6 Storage allocation of major work systems

Work importance	OS version	Work system	LV (GB)	LV (EA)	Allocated capacity(GB)
High	IBM AIX 53	Business email system	12	100	1,200
	Window NT (Window 2008SE)	Personnel management system	12	10	120
		Knowledge portal system	12	50	600
	HP – UX 11.31	Electronic approval system	12	25	300
		Work management system	12	50	600
Medium	Window NT	Extranet homepage system	12	10	120
	IBM AIX 53	Overseas market information system	12	20	240
		Inventory control system	12	50	600
		Geographic information system	12	200	2,400
	Low	IBM AIX 53	Production history Mgm. system	12	100
Business management system			12	10	120
Production statistics system			12	100	1,200
Basic statistics system			12	80	960
SUNOS 5.10		Inventory control system	12	50	600
		Assets Management System	12	60	720
		Total	–	–	915

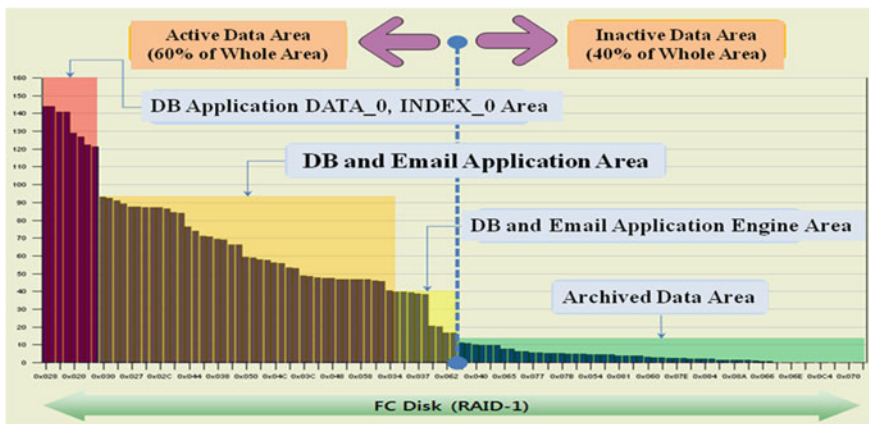


Fig. 4 The details of I/O per business hour

This caused inefficiency of the storage capacity management, increasing complexity of data management, and additional storage capacity

2. There is not backup policy for data after storage period. Therefore data to delete are still included in incremental and whole backup so that backup costs and capacity are continuously increasing

2.5.2 Operation and Protection Mechanism for Data Groups

RAID-1 disk which the company is now using is most popular for DB application volume in spite of high price since its write function is powerful and write penalty is low. However, since both active and inactive data are in RAID-1 as shown in Fig. 6 differential protections cannot be applied to them. In addition, inactive archive data are stored in expensive RAID-1 disk so that they increase customer's TCO.

2.5.3 Storage Media for Data

In general, disk media have various capacities and their operation costs vary. However, active and inactive data with very different features stored in FC (Fiber Channel) disk media with same physical features so that data storage costs are increased.

3 System Improvement

3.1 Establishment of Efficient Data Management Policy

I have improved following items to establish policies for efficient data operation and management with systematic classification of data values through type analysis of data.

1. I classified data values and set data grades as described in Table 7
2. I established standards for disk types and data protection level, as described in Table 8, to guarantee differential data usability and response time based on data grades as described in Table 7
3. I established new management policy for inactive data based on life cycles. The new policy is to delete the inactive data after conducting separate tape backup for them and, when necessary, reuse them from separate backup server file system

3.2 Configuration of ILM-Based Tiering Storage

3.2.1 Measurement to Guarantee Faster Response Time and Data Usability

Since the existing FC disks is not suitable for storage media for the active data area, I suggested new storage tiers with SSD (Solid Status Drive) which can provide higher performance to improve the lowered response time of applications caused by intensive I/O. In addition, efficiency is provided for the inactive data

Table 7 Importance of major works' data

Importance	Work system	Data grade		
		Platinum	Silver	Bronze
High	Business email system	DB data, intensive I/O	Application installation area email and DB data area	Application operation log storage area
	Personnel management system			
	Knowledge portal system			
	Electronic approval system			
	Work management system			
Medium	Extranet homepage system	N/A	DB data area	Application installation area and operation log storage area
	Overseas market info' system			
	Inventory control system			
	Geographic information system			
Low	Production history mgm' system	N/A	N/A	Application installation area and statistics data storage area for long-term storage
	Business management system			
	Production statistics system			
	Basic statistics system			
	Inventory control system			
	Assets management system			

using SATA (Serial Advanced Technology Attachment) disks whose speed is low but disk efficiency is high compared to price.

3.2.2 Establishment of Tiering Through Re-configuration of Storage

If inactive and active data are in the same physical disk groups, response time of DB applications for active data may be negatively influenced by I/O interference between applications when intensive I/O is occurred by mass batch jobs of inactive data. I classified work areas per data grade as described in Table 9 based on Table 7 for re-configuration of disk groups to physically divide the two data areas to minimize such problems.

I reconfigured disk groups as described in Table 10 using tiering concept based on data in platinum, silver, and bronze grades as described in Table 9 and new storage media for scalability.

Table 8 Storage infrastructure environments per data grade

RAID type	Importance of data	Data grade	Capacity requirement (IOPS)	Feature
RAID-1	High	Platinum (New)	Over 5,000	Low disk efficiency (50 %). Higher costs than RAID-5. Lower write penalty than RAID-5
	Medium	Silver	(Now)	
120-167				
RAID-5 (7D + 1P)	Low	Bronze (New)	80	High disk efficiency (75 %). Lower costs than RAID-1 High random and sequential R/W High parity write penalty

Table 9 Data grade of major work systems

Importance	Work system	Existing allocated capacity (GB)	Estimated capacity per data Area (GB)		
			Platinum	Silver	Bronze
High	Business email system	1,200	0	960	240
	Personnel management system	120	24	72	24
	Knowledge portal system	600	120	420	60
	Electronic approval system	300	60	180	60
	Work management system	600	180	360	60
Medium	Extranet homepage system	120	0	96	24
	Overseas market info' system	240		192	48
	Inventory control system	600		480	120
	Geographic information system	2,400		480	1,920
Low	Production history mgm' system	1,200	0	0	1,200
	Business management system	120			120
	Production statistics system	1,200			1,200
	Basic statistics system	960			960
	Inventory control system	600			600
	Assets management system	720			720
	Total	Business email system	10,980	384	3,240

4 Test Environments and Evaluation

Test environments were configured to verify utility of ILM-based tiering storage through comparison of key performance indexes before and after data migration. Baseline data collected before improvement and data collected after improvement

Table 10 Re-configuration of storage

Data Grade	Configuration of disk						
	Disk group	Type	RAID type	Capacity (GB)	Quantity (EA)	Physical capacity (GB)	Available capacity (GB)
Platinum (new)	# 4	SSD	RAID-1	300	4	1,200	600
Silver (existing)	# 1-3	FC	RAID-1	146	160	23,360	N/A
Bronze (new)	# 5	SATA	RAID-5 (7D + 1P)	1,000	1,000	16,000	14,000

were compared and analyzed using VDBench [12] and Workload Analyzer [13] for objective reliability and validity. Various types of I/O workloads were generated using VDBench for the test.

4.1 Comparison of I/O Occurrence

I compared number of I/O occurring from 10 devices each in both active and inactive data areas per second before and after data migration. Number of I/O for active data was significantly decreased to about 21 % of the former number but inactive data didn't show meaningful change as shown in Figs. 5 and 6.

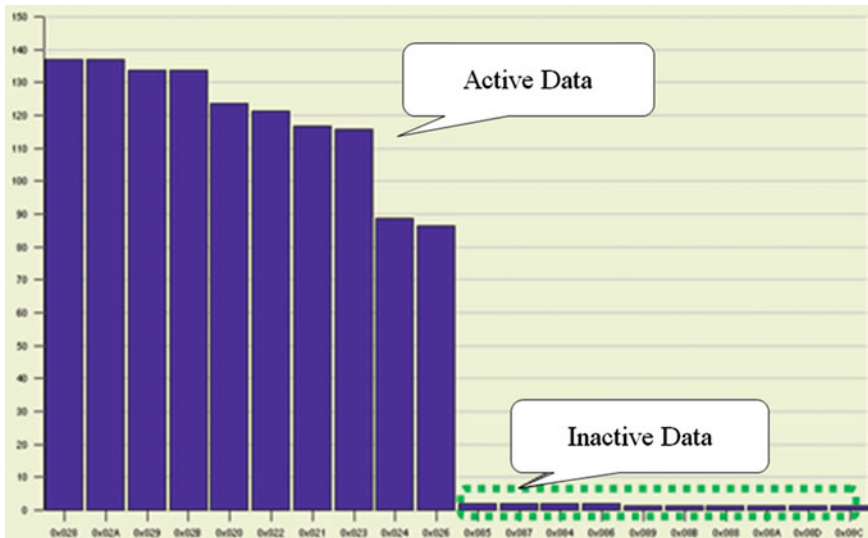


Fig. 5 Total I/Os per second (before)

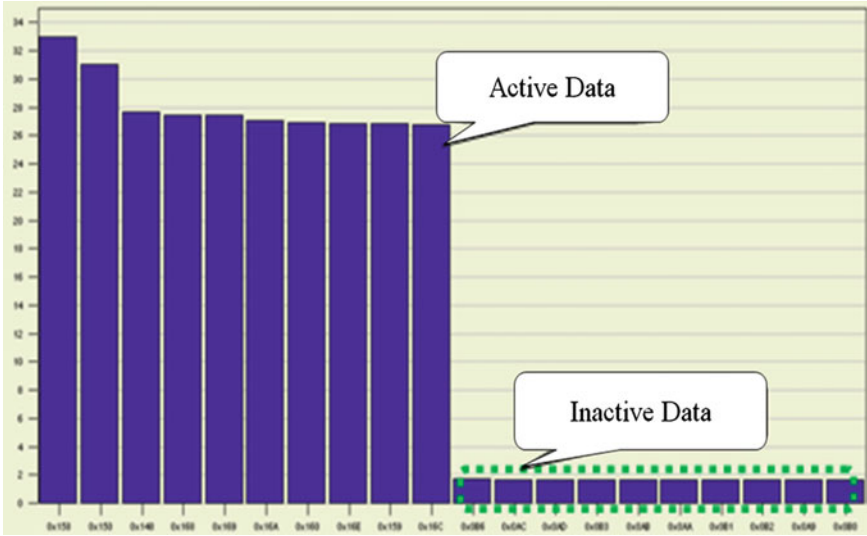


Fig. 6 Total I/Os per second (after migration)

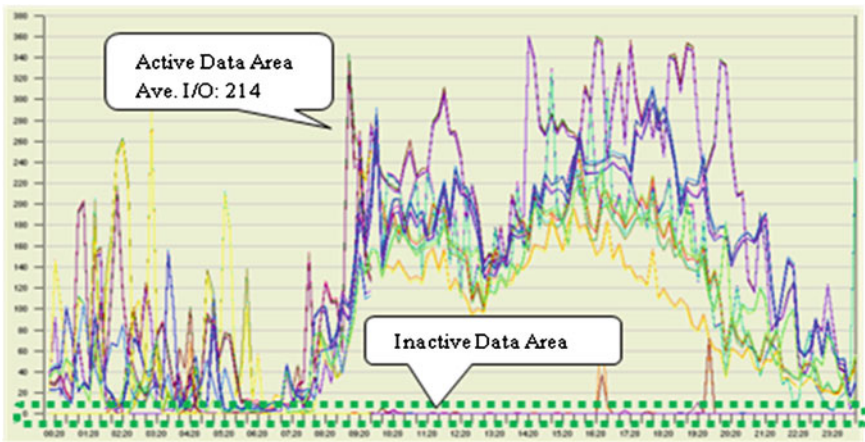


Fig. 7 Device I/Os per second. in business hour (before)

I measured number of I/O occurring in business hour. Number of I/O for the active data which significantly influence the system performance was decreased by about 6 times after the storage tiering as shown in Figs. 7 and 8 showing 8 devices for active data and 2 devices for inactive data with highest I/O occurrence. That was because the active data areas were evenly distributed in the disks and general performance of the storages and DB applications was significantly improved. However, data I/O occurring of the inactive per second was less than five which is too small to measure differences.

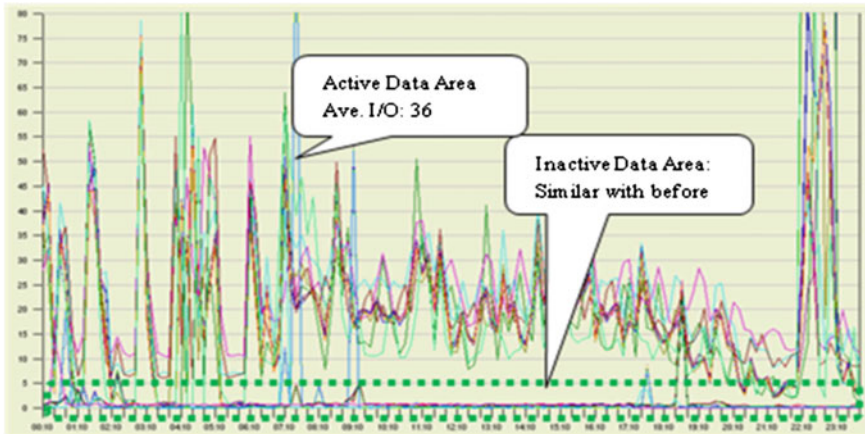


Fig. 8 Device I/Os per second. in business hour (After migration)

4.2 Comparison of Device Transaction Volume

I measured transaction volume per second for active data device areas which applications mainly use. Response time of DB applications was negatively influenced by concentrated transactions (2,000 KB in average) on specific devices (0 × 020, 0 × 021, 0 × 022, 0 × 023) before the storage tiering as shown in Fig. 11. But the transactions were distributed (480 KB in average) after data migration and bottleneck was improved by about 2.8 times as shown in Figs. 9 and 10.

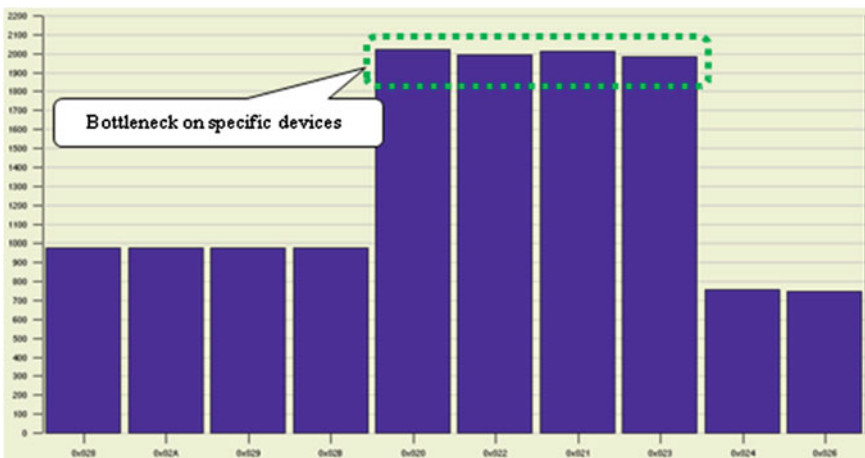


Fig. 9 Transaction volume per second (KB) (before)

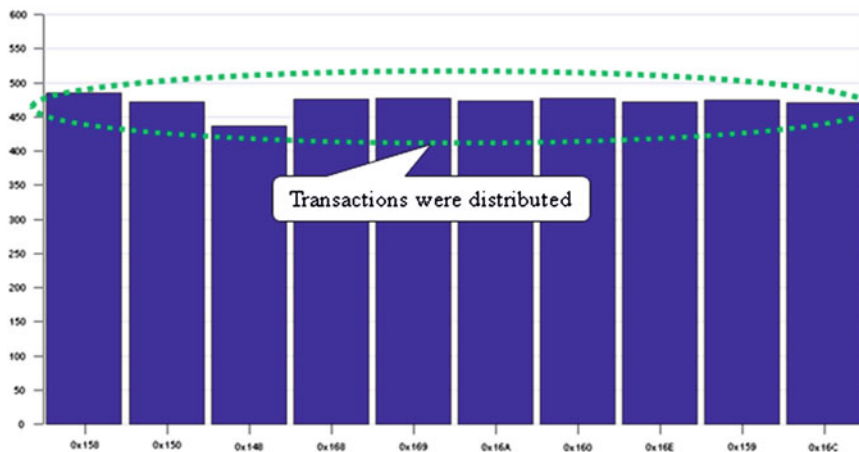


Fig. 10 Transaction volume per second (KB) (after migration)

4.3 Comparison of Read/Write per Disk Media

The average value of read sizes for SSD disk media is 120 KB which is double of that of FC disk media as shown in Figs. 11 and 12. An average read times for active data areas in existing FC media and SSD media, after migration, are 42 ms

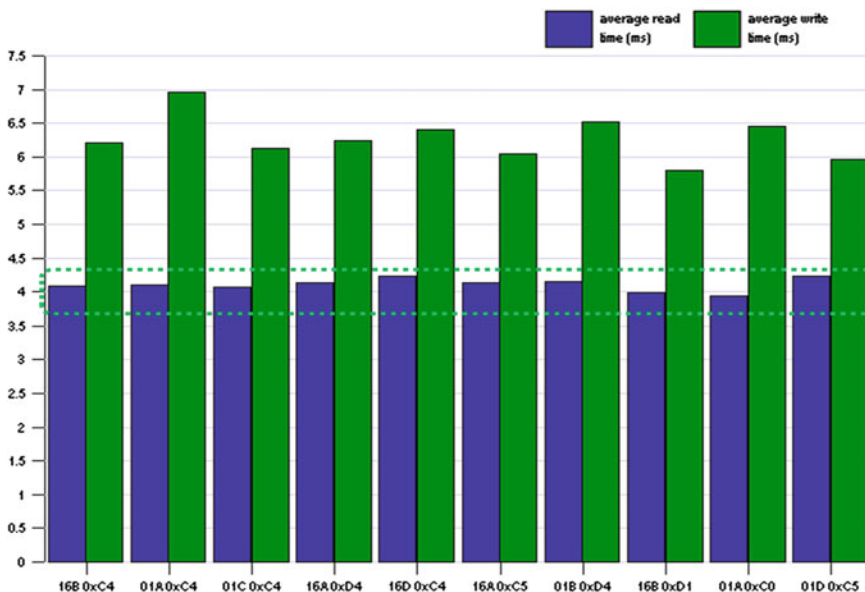


Fig. 11 Average read/write time (ms) per disk (before)

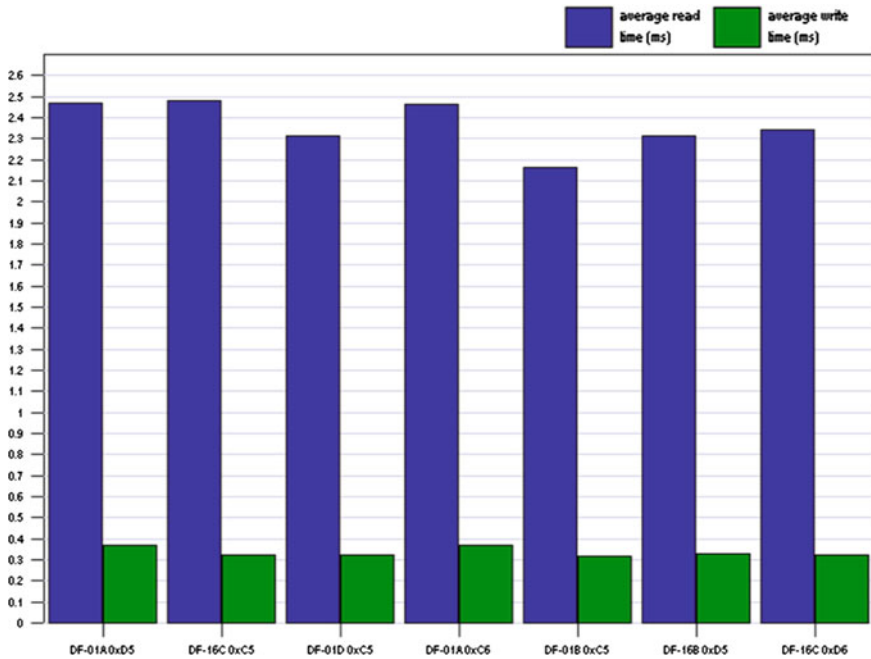


Fig. 12 Average read/write time (ms) per disk (after migration)

and 24 ms each. It was improved by 2 times. For write time, the average values before and after migration were 62 ms and 0.36 ms each. It was improved by 17 times and shall improve the response time of DB applications significantly.

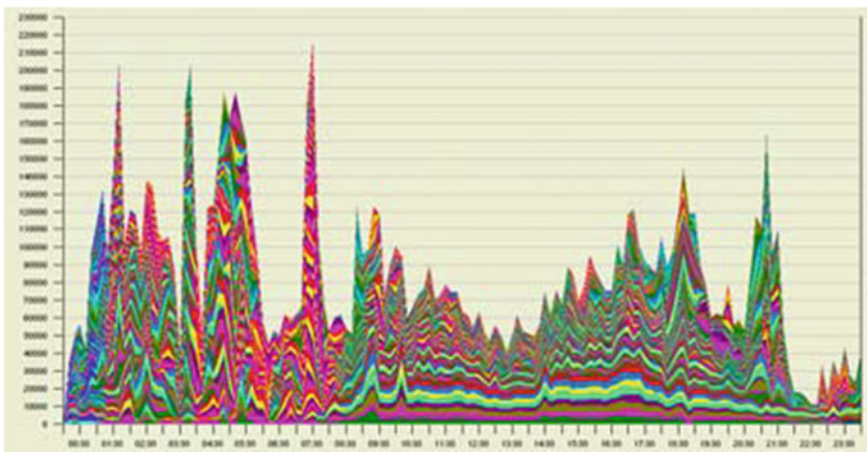


Fig. 13 Total throughput to-from hosts (before)

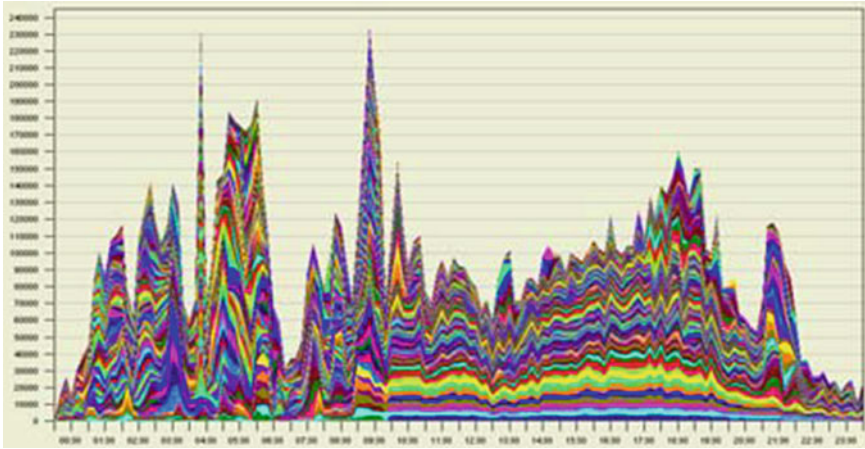


Fig. 14 Total throughput to-from hosts (after migration)

4.4 Comparison of Throughput for Storage System

I measured and compared overall throughput of the storage system for 1 week to confirm how much the partial improvements contributed to whole storage system. The throughput for the existing system and the tiering system were about 70,000 KB/s and 85,000 KB/s each as shown in Figs. 13 and 14. It was improved by 21 %.

5 Conclusions

Improvements through this study are as follows;

1. Overall throughput of the storage system was improved by 21 % and new SSD media improved read time by 2 times and write time by 17 times. Bottleneck was improved by 2.8 times by the tiering. Number of I/O occurring in active data which significantly influence the system performance was decreased by 6 times
2. This storage tiering test confirmed that data management policy based on ILM through data value analysis was a very effective cost-saving solution considering consistent quality of systems under long-term view
3. Efficiency for storage management was improved and efficient data management policy was built through application of information life-cycle concept to existing data management policy
4. This study can be a good reference for most companies which don't have systematic storage systems to solve explosive data increase caused by various service requests from customers

5. The tiering based on ILM through data value analysis enables companies not only to utilize existing disk media but to save system maintenance costs through enhanced system performance and efficient investment for future expansion under long-term view

References

1. <http://www.ciokorea.com/ciolibrary/9138> Sharp increase of data, we need a storage-oriented strategy, IDG Research White paper, 2011
2. Kweon H (2004) Acceptance of ILM strategy. *Netw times* 135:137–139
3. Twentyman J (2006) The ILM imperative: smart use of the data storage infrastructure is helping many companies tackle data management and compliance burdens-but an information life-cycle management strategy requires a lot of upfront effort. *EI Ark Group* 3(5):12–15
4. Baffa C (2005) ILM for life sciences: the next big storage play?. *Comput Technol Rev* 24
5. EMC (2011) <http://korea.emc.com/collateral/article/on25-summer.pdf>, Information Life-cycle Management, EMC
6. Amelang M (2005) ILM, tiered storage and active archive form a powerful trio Companies embrace information lifecycle management to meet explosive data growth challenges. *World Oil* 226(11):63–66
7. Jeong SR, Shin DW (2009) Improving the utilization and efficiency of enterprise architecture (EA) through the implementation of information lifecycle management (ILM)-based EA system. *Ocean Sci J* 6(2):107–121
8. Lee HS (2000) A plan to build enterprise storage system using storage area network concept. Yonsei University Major in Computer Engineering Graduate School of Engineering
9. Kim D (2009) Archiving system construction based on database for ILM implementation. *Korean Inf Process Soc* 16(02):0737–0738
10. Im J-H, Lee C-G, Lee Y-J (2005) The information value-based document management technique using the information lifecycle management theory. *Korea Soc Simul* 14(4):19–30
11. Kim SE (2010) Storage technology issue: cloud data center should reduce the Information management complexity. *Netw Times* 202:131–144
12. <http://vdbench.sourceforge.net/>
13. <http://www.datastorageconnection.com/doc.mvc/EMCs-Workload-Analyzer-0001>

A Neural Network Mixture Model for Green Warranty Diffusion

Sang-Hyun Lee, Sang-Joon Lee and Kyung-II Moon

Abstract The purpose of this paper is to assist in measuring all costs associated with product warranties including the environmental problems and in estimating the potential warranty cost savings. The concept of the green warranty is emphasized in this paper because of its effect on increasing the scope of warranty cost savings. This paper suggests a new concept for the design of warranty system that combines some of neural network approaches in green IT's point of view. In particular, Gompertz function is used as the transfer functions in the model. The academic importance of this study is that Gompertz can be a type of mathematical model for green warranty claims, where warranty growth is slowest at the start and end of warranty lifetime period. To apply the model to warranty data, the practitioners need not identify parametric distributions for the failure attributes. To demonstrate the model, this paper develops a neural network mixture model for the automotive warranty data.

Keywords Green warranty · Neural network · Warranty diffusion · Gompertz function

S.-H. Lee
Department of Cultural Industry Management, Honam University,
Gwangju, Korea
e-mail: leesang64@honam.ac.kr

S.-J. Lee (✉)
School of Business Administration, Chonnam National University,
300 Youngbong-dong Buk-gu, Gwangju, Korea
e-mail: s-lee@chonnam.ac.kr

K.-I. Moon
Department of Computer Engineering, Honam University, Gwangju, Korea
e-mail: kimoon@honam.ac.kr

1 Introduction

The demand for green warranty diffusion is essential in this society. Until now the warranty model have been extended or altered to fit in with the heuristic situation. The fuzzy logic [1, 2], the neural network [3] and the knowledge management [4] are introduced to this model. In regard to this warranty, a method was discussed to estimate hazard rate from incomplete and unclear warranty data [5, 6]. Lawless [7] and Majeske [8] studied warranty data with focus on automobile industry. Recently, Lee and et al. presented many approach for warranty claims with artificial method [9–13].

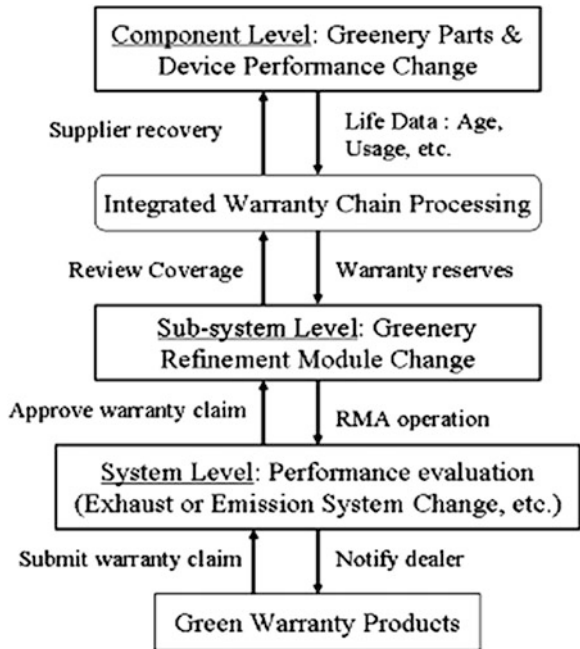
In this paper, we propose the mixture model combined neural network with Gompertz function, traditional diffusion model to predict and model the green warranty. This paper is organized as follows: Sect. 2 introduce a green warranty process. Section 3 propose a neural network mixture model. Section 4 demonstrates the application of our model.

2 Green Warranty Process

When considering warranty issues, an important concept to keep in mind is the warranty chain. Like the supply chain for purchasing and manufacturing, the warranty chain extends the scope of warranty activities beyond the walls of a single company to encompass suppliers, manufacturers, OEMs, distributors, dealers, repair centers, policy carriers, and customers.

Figure 1 represents an environmental parts warranty system around the environmental parts in the general warranty system. Once some warranty claims are submitted to the dealer and repair center, some classification works and performance evaluation are processed in overall system level. If there are some valid warranty claims, invalid parts are tested in the sub-system levels, for instance, the exhaust system of automotive parts. These steps are labor intensive, time consuming warranty processes. Next step is composed of the repair work or replacement in the component level, and reserved the warranty data such as age and usage in the warranty database. Some replacement or repaired parts are again tested in the sub-system levels. These warranty works are corresponding to the Parts Return or Return Merchandise Authorizations (RMA). Since the results of RMA work give effect to the overall system, many tests are needed as a whole. In the sub-system and component levels, the warranty data analysis is corresponding to a knowledge base system. The reason is that many uncertainties are existed and some qualitative evaluations are required by the expertise. The remaining levels above the sub-system are connected to some warranty degree determination. An approximate reasoning method can be used for such warranty degree determination since the exact warranty degree determination is difficult. The reason is that there are many qualitative factors such as seasonality and assembly skills. Further, it is required the multidimensional analysis considered both the age and usage variables [10, 11].

Fig. 1 Warranty process for environmental parts



3 Neural Network Mixture Model

The Gompertz function [4] can be used as a sigmoid function. The objective is to approximate a smooth scalar function of q inputs h : $R_q \rightarrow R$ using a feed-forward Gompertz function network. The set of nonlinear weight equations that relates the neural network’s adjustable parameters to the warranty data is obtained by imposing the training requirements on the network’s output and gradient equations. First, consider a multilayer feed-forward network with one layer of hidden neurons and one output neuron. When an input vector, representing time and a time-dependent quantity such as amount of use, is propagated through the network, for the current set of weights there is an output warranty, denoting the actual warranty amount. The objective of supervised training is to adjust the weights so that the difference between the network output z and the actual output w is reduced. This requires an algorithm that reduces the absolute error, which is the same as reducing the squared error, where Network Error = $z - w = E$. The algorithm should adjust the weights such that E^2 is minimized (Fig. 2).

Fig. 2 Neural network mixture model

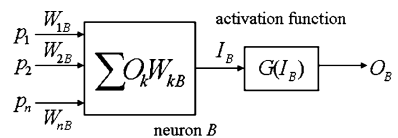


Table 1 Warranty summary by contingency table

Age (months)	Usage (mileage: 1000 km)															Total
	100-105	105-110	110-115	115-120	120-125	125-130	130-135	135-140	140-145	145-150	150-	Total				
3	2	2	0	1	0	0	0	0	0	0	0	0	0	0	5	
6	3	0	1	1	1	0	0	0	0	0	0	0	0	0	7	
9	1	1	0	1	0	0	0	0	0	0	0	0	0	0	3	
12	4	0	1	0	1	0	0	0	0	0	1	1	1	1	8	
15	0	3	0	3	1	0	0	0	2	1	1	1	0	0	11	
18	2	1	0	1	0	0	0	0	1	0	0	0	1	1	6	
21	1	1	2	0	1	0	0	0	1	0	0	0	0	0	6	
24	1	0	1	0	0	2	0	0	0	0	0	0	0	0	4	
27	1	0	2	0	2	0	0	0	0	0	0	0	0	0	5	
30	5	2	2	0	0	1	0	0	1	0	0	0	0	0	11	
33	3	3	0	1	1	1	2	1	1	0	0	0	0	0	12	
36	16	11	12	5	7	4	5	1	1	0	2	1	1	64	64	
39	46	26	29	21	20	11	5	14	6	6	9	6	6	193	193	
42	45	40	31	23	23	14	21	7	10	7	9	7	7	230	230	
45	77	54	47	39	28	33	17	13	16	12	12	7	7	343	343	
48	84	72	58	73	55	37	42	31	19	19	19	15	15	505	505	
51	134	99	83	73	65	41	39	30	27	20	20	21	21	632	632	
54	110	78	76	65	40	35	34	35	20	20	20	22	22	535	535	
57	55	36	38	22	31	24	19	27	14	23	10	10	10	299	299	
60	33	26	25	17	20	23	17	15	11	10	10	10	10	207	207	
63	29	21	21	16	10	10	7	14	5	6	8	8	8	147	147	
66	13	11	13	6	6	4	4	2	6	6	6	2	2	73	73	
69	7	5	11	5	4	3	4	5	4	1	1	3	3	52	52	
72	3	2	5	6	3	3	1	3	1	4	4	0	0	31	31	
75	1	3	7	2	3	3	3	3	2	2	2	1	1	30	30	
Total	676	497	465	381	322	249	220	207	142	145	115	115	115	3419	3419	

4 Application

We used the number of warranty claims related to O₂ sensors in the automobile warranty domain as a measure of the environmental parts warranty growth. Oxygen sensors are a product that have been around for more than 20 years, yet most motorists do not even know they have one or more of these devices on their vehicle.

The only time most people even become aware of an oxygen sensor’s existence is if they get a check engine light and there is a code that indicates an O₂ sensor problem their vehicle fails an emissions test because of a sluggish or dead O₂ sensors. But in most cases, they will not have a clue as to know to diagnose or test this mysterious little device that is often blamed for all kinds of drive ability and emissions ills. We collected the pure number of environmental parts warranty from January 2002–March 2008 from the warranty reports published by automobile company in Korea.

Table 1 denotes a summary data by using the two-way contingency table. It displays the warranty counts for each combination of two-attribute warranty variables (age and mileage). Since there can be multiple models on the same car type or multiple car types used for a single O₂ sensor, the count represents entire ones rather than the number of special cars.

The box plot shows that both age and usage have a significant effect on warranty claims (see Fig. 3). Usage has the higher warranty response (about 310); age has the lower warranty response with about 136. It shows statistically significant differences in warranty accuracy ($p = 0.0137 < 0.05$) between the age and usage variables.

Table 2 denotes the usage weighted data to adjust for the differences between the age variable and usage variable. Weighting by each proportion of the usage

Fig. 3 Box plots of age and usage

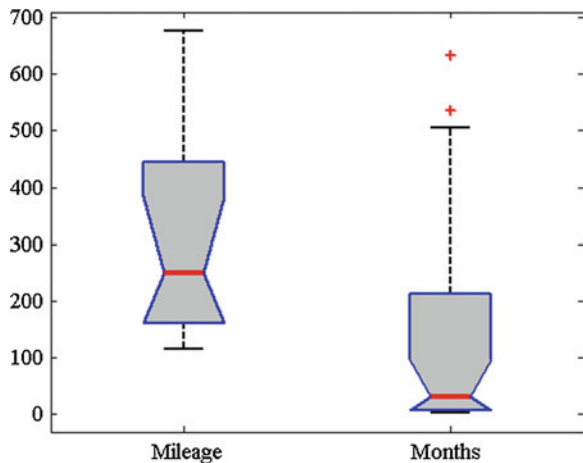


Table 2 Weighted warranty claims

Age (months)	Weighted warranty	Cumulative	Age (months)	Weighted warranty	Cumulative	Age (months)	Weighted warranty	Cumulative
3	4.20	4.20	30	9.32	57.25	57	266.54	3153.80
6	6.01	10.21	33	10.51	67.76	60	185.10	3338.90
9	2.55	12.76	36	55.61	292.56	63	129.90	3468.80
12	6.9	19.66	39	169.19	461.75	66	64.60	3533.40
15	9.93	29.59	42	202.57	664.32	69	46.38	3579.78
18	5.25	34.84	45	300.93	1412.70	72	27.81	3607.59
21	5.23	40.07	48	447.45	1860.15	75	27.11	3634.70
24	3.52	43.59	51	555.72	2415.87			
27	4.34	47.93	54	471.39	2887.26			

Fig. 4 Gompertz function fitting

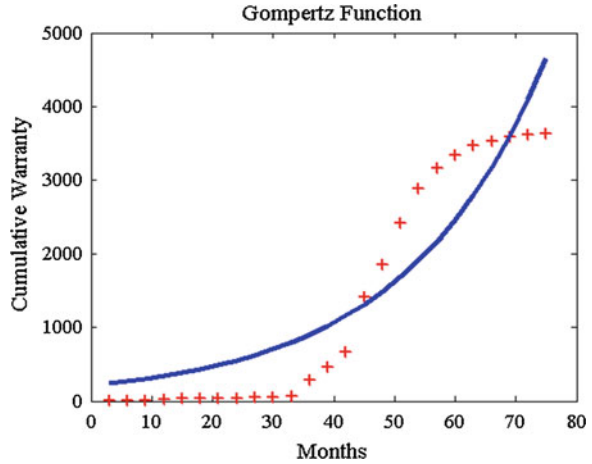
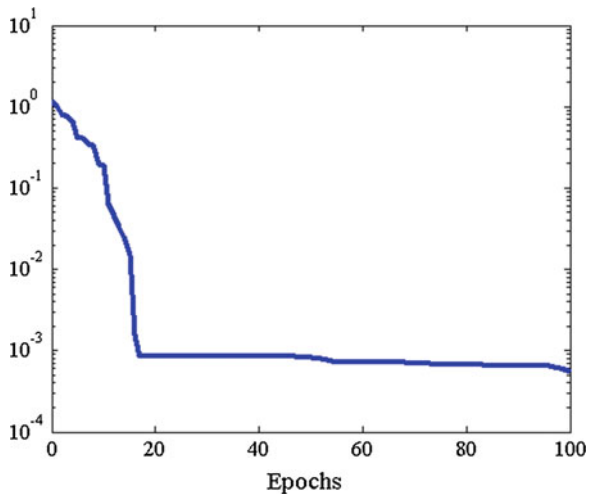


Fig. 5 Network error

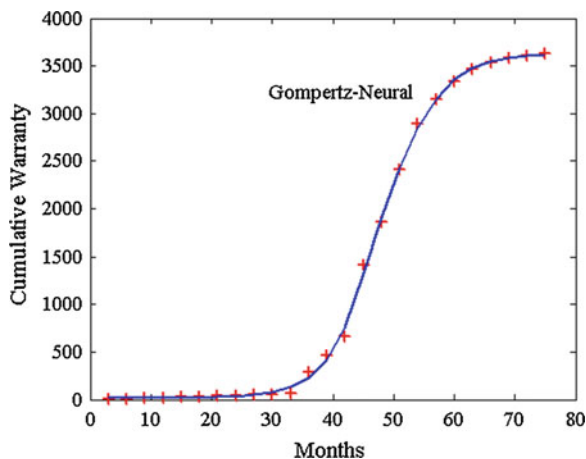


variable can be a compromise to minimize the actual differences between the age and usage variables.

Figure 4 denotes a Gompertz fitting result to the cumulative data of Table 2. The initial estimates of the parameters a , b and c are 0.0002, 41.247 and 0.001, respectively. Using the Gauss–Newton method, the final estimates of Gompertz parameters are $1.2081e-15$, 41.999 and $9.489e-3$, respectively. The root mean squared error (RMSE) as a performance measure is 319.4940. RMSE is the preferred performance measure when different methods are compared for decision making.

For the simulation of suggested method, we use a 1-3-1 architecture (one input layer, three hidden units and one output unit), and Gompertz function as the transfer function, and it was trained with a learning rate of 0.1, the momentum

Fig. 6 Gompertz-neural fitting



term was set at 0.07, and 100 iterations. As the parameter estimates for the NN model, 0.00965, 0.1039 and 0.31142 are the estimated final weights from input neuron to hidden units 1, 2, and 3 respectively.

Also, 0.99991, 1.0279, and 0.14512 are the estimate weights from hidden nodes 1, 2, and 3 respectively, to the output neuron. The NN model converges after 18 epochs with the objective function (sum of the absolute deviations among sample points) value of approximately 0.00017 (see Fig. 5).

The NN model has smaller RMSE (182.1164) than that of the pure Gompertz model (319.494) in the calibration data. Figure 6 represents that the neural net is approximating the warranty diffusion smoothly. In summary, the results indicate that the Gompertz NN model is superior to the traditional diffusion models in forecasting the warranty growth. In particular, it is significant to note that the environmental parts warranty was down for discussion from April 2004. The S-shaped curve is well-established for modeling diffusion due to the environmental parts warranty.

5 Conclusion

This paper makes a contribution to environmental parts warranty research by suggesting a novel approach to model its diffusion. The warranty diffusion models are variants of S-curves. External influences such as environmental parts warranty are largely ignored when it comes to modeling. This is perhaps because external effects can be domain-sensitive or occur at any stage of the warranty life cycle. Random occurrence of external effects causes problems in modeling warranty diffusion process with any pre-determined mathematical relation such as an S-curve. The mathematical models that fit S-curves to diffusion data thus essentially treat external perturbations as random error. These random errors reduce the

accuracy of the forecasts of these models. The results seem to support the argument that the more flexible approach of a neural network model is better than the conventional models for forecasting the warranty diffusion, especially when there is cause to believe that external factors perturb the diffusion phenomenon. However, this represents the results from one actual dataset. The validation of the current study is further strengthened by conducting a simulated experiment that compares the models across various datasets that are created to systematically examine the effects of different types of disturbances in the warranty data. Specifically, one must consider four characteristics of the external effects: magnitude, direction, stage of the warranty life cycle of diffusion, and type of S-curves.

References

1. Yadav NS, Chinnam RB, Goel PS (2003) A fuzzy logic based approach to reliability improvement estimation during product development. *Reliab Eng Syst Saf* 80:63–74
2. Lolas S, Olatunbosun OA, Steward D, Buckingham J (2007) Fuzzy logic knowledge base construction for a reliability improvement expert system. In: *Proceedings of the world congress on engineering, 2007*
3. Ferrari S, Stengel RF (2005) Smooth function approximation using neural networks. *IEEE Trans Neural Netw* 16(1)
4. Liao SH (2003) Knowledge management technologies and applications—literature review from 1995 to 2002. *Expert Syst Appl* 25:155–164
5. Rai B, Singh N (2003) Hazard rate estimation from incomplete and unclean warranty data. *Reliab Eng Syst Saf* 81:79–92
6. Kim HG, Rao BM (2000) Expected warranty cost of two-attribute free-replacement warranties based on a bivariate exponential distribution. *Comput Ind Eng* 38:425–434
7. Lawless JF (1998) Methods for the estimation of failure distributions and rates from automobile warranty data. *Lifetime Data Anal* 1:227–40
8. Majeske KD (2003) A mixture model for automobile warranty data. *Reliab Eng Syst Safety* 81:71–77
9. Lee S-H, Lee S-J, Moon K-I, Kim B-K (2008) Decision of abnormal quality unit lists from claim database. *J Inf Process Syst* 4(3)
10. Lee S-H, Cho S-E, Moon K-I (2010) Fast fuzzy control of warranty claims system. *J Inf Process Syst* 6(2)
11. Lee S-H, Cho S-E, Moon K-I (2010) Neural network approach for greenery warranty systems. In: *The 6th international conference on intelligent computing, ICIC 2010 Changsha, China, 18–21 Aug 2010*
12. Lee S-H, Lee S-J, Moon K-I (2010) A fuzzy logic-based approach to two-dimensional automobile warranty system. *J Circuits Syst Comput* 19(1):139–154
13. Lee S-H, Lee S-J, Moon K-I (2011) A two-attribute green warranty model for automobile warranty data. In: *International conference on computer convergence technology, Seoul Korea, 20–22 Oct 2001, pp 444–449*

Part IX
Web Technology
and Software Engineering

Generation of User Interest Ontology Using ID3 Algorithm in the Social Web

Jong-Soo Sohn, Qing Wang and In-Jeong Chung

Abstract It is feasible to collect individual user interests from social networking services. However, there have been few studies of the interests of domain users. In this paper, we propose an approach for ontology generating the interests of SNS domain users by employing semantic web technology and ID3 algorithm. In our approach, domain ontology is generated by a decision tree, which classifies the domain web pages and the domain users. Experimental test shows ontology of the interests of domains users regarding USA presidential candidates. We expect that our results will be beneficial in the field of computer science, such as recommendations, as well as other fields including education, politics, and commerce. Proposed approach overcomes the problem of domain user classification and lack of semantics by composing decision tree and semantic web technology.

Keywords Semantic web · Ontology · ID3 algorithm · SNS · FOAF · Interest extraction · Social web · Classification · Election · OWL

J.-S. Sohn · Q. Wang · I.-J. Chung (✉)
Department of Computer and Information Science, Korea University,
208 Seochangri, Sejong City, South Korea
e-mail: chung@korea.ac.kr

J.-S. Sohn
e-mail: mis026@korea.ac.kr

Q. Wang
e-mail: wangqing@korea.ac.kr

1 Introduction

The interests of user have an important role in various areas such as recommendations, commerce, and political activities, although less information is provided by users than is available on networks. Thus, the working efficiency is very low if we have to identify the interests of users one by one. The interests of individual users are valuable for personalized services, but they cannot satisfy the requirements of public services because public users lack the universality. By contrast, the interests of public users lack personality, although they possess universality. The interests of domain users are different from individual interests and public interests. They are the common interests of users in the same domain and they have different personalities because users come from a different domain, although they have the same universality because users are in the same domain. Therefore, the value of the interests of domain users is greater than the interests of individual user' and public users in other fields in addition to computer science. However, only a few studies have addressed the interests of domain users. The process of extracting the interests of domain users has the following two problems [1].

- Defining the user domain. It is simple for a human to define a user domain, but this is a very difficult task for a computer, because machines do not understand semantics.
- Classifying different domain users. Defining a user domain is very difficult for computers. Computers are unable to classify different domain users based on a traditional classification if this requires an understanding of semantics.

Therefore, we propose an approach for extracting the interests of domain users, which exploits semantic web technology for collecting them from SNS. We create a domain ontology that helps computers understand semantics. Semantic web technology is a maturing technological field, which continues to be the focus of much research

Our methodology addressed the difficulties of defining user domains and the classification of different domains to identify the interests of domain users. With our approach, domain ontology is generated by a decision tree that classifies the domain webpages and domain users. Experimental test shows ontology of the interests of domains users regarding USA presidential candidates. The result of our research has great value for computer science and other fields including education, commerce, and political activities.

2 Related Work

Folksonomy is a very important part of web 2.0, which is used to classify SNS such as the Delicious website. This is a new method that allows users to classify information [2], which is also associated with a semantic problem [3].

Research into this area has included recommendations, user interests, and classification. Illig et al. proposed a tag recommendation method for folksonomy [4]. In [5], Shan et al. described item recommendation based on folksonomy. Szomszor et al. proposed the modeling of user interest based on folksonomy while Kawase et al. proposed a classification for user interest patterns based on folksonomy [6, 7]. Neither classification nor the extraction of user interests of based on folksonomy can achieve classification on their own. They still require other tools or cooperation with traditional classifiers, because folksonomy is associated with the semantic problem.

Few studies have addressed classification methods based on semantics using ontology. Some researchers have focused on the interests of domain users, while others have addressed the interests of individual users. Lipczak proposed a tag recommendation method based on the interests of individual users [8]. Yin et al. proposed the semantic modeling of individual user interests and collaborative tagging systems for individual users [9]. Sasaki et al. proposed a method for extracting preference terms, which excluded unrelated pages from user interests where the user also referred to individual users [10]. White et al. proposed a method for predicting the interests of individual users based on contextual information [11]. However, the interests of domain users have greater value than individual interests and public interests for computer science and other fields.

3 Generation of User Interest Ontology Using ID3 Algorithm

3.1 Generating Decision Trees

Decision tree is a popular classifier, which is characterized by the property that samples are subjected to a sequence of decision rules before they are assigned to a unique class [12]. We acquired training data from the Delicious website to generate a domain decision tree. The domain decision tree is used to create domain ontology. And every pathway in a decision tree can be written using a tag rule such as the Web Ontology Language OWL-DL, so the sub-tree can be translated into domain ontology without loss during classification based on semantics. Decision trees are part of the classification when generating ontology. We use the ID3 algorithm to generate a decision tree for a domain [13]. OWL was developed as a more formal and more powerful ontology language than RDFS. OWL-DL supports users who require maximum expressiveness while retaining computational completeness and decidability [14].

In generate a decision tree for domain within specific limits, we collected a random sample set from the Delicious website as training data, which consisted of the top 5 tags and webpage links, which we stored as a table. The webpage “The Choice: Comment: The New Yorker” is marked by some users on the Delicious website. We extracted the webpage link “The Choice: Comment: The New Yorker”

Fig. 1 An item of training data in the ID3 table

obama	romney	bachmann	politics	election	2012	calss
YES	NO	NO	YES	YES	NO	Obama

and the top 5 tags for Obama, politics, election, New Yorker and an endorsement of this paper, and we recorded them in a table as training data for the decision tree.

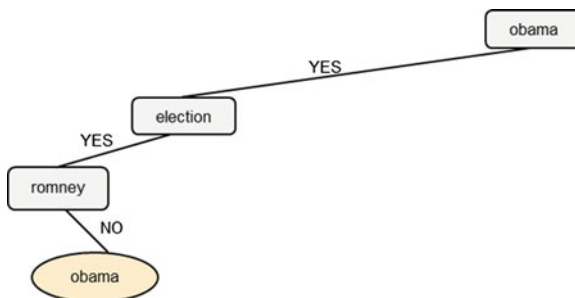
We prepared a table to describe a situation for ID3, which consisted of all the keywords and classes. Next, we determined the relationships between each piece of training data and keywords. In each piece of training data, if any tag matched any keyword, we recorded the cell of the keyword as “YES” in the ID3 table. Cells of keywords were recorded as “NO” when there no tag matched with it. We recorded the cells for Obama, politics, and election as “YES” whereas the others were “NO” in the ID3 table, as shown in Fig. 1. Next, we check the content of the webpage and record its class. Finally, all of the training data is recorded in the table.

Information gain is a measure of the information theory, which characterizes the impurity of an arbitrary collection of examples. The method used for information gain is covered in [15]. After the information gains are calculated for each attribute, we can draw the domain decision tree. Every attribute will have a position based on the results by information gain.

Generated decision tree contains several pathways in different sub-trees that lead to different classes, so the same class may have several pathways. Each pathway consists of several keywords that can be expressed as a tag rule using OWL-DL.

For example, Fig. 2 shows a pathway in the domain decision tree. Based on the calculation of the information gain, the keywords ‘obama’, election, and ‘romney’ are attributes in a pathway in the domain decision tree. We can use the OWL-DL to write a tag rule as follows: $obama \sqcap election \sqcap \neg romney$. Because of the top 5 tags, every webpage on the Delicious website also has a tag rule, which is stored in our webpage database. Based on all of the pathways in the decision tree, we can write tag rules using OWL-DL for every class in the domain ontology.

Fig. 2 Apathway in the domain tree



3.2 *Ontology Learning*

To infer and classify webpages into different domain classes, the domain ontology has to be trained using the webpage data. The webpage data is used to train the ontology, which is also classified. This is an important part of classification, which we discuss further in this chapter, and it consists of data storage, learning, and inference.

We store webpage data in the database, which consists of the URL of a webpage and tags. We write a tag rule for each webpage to infer the class of the webpage based on the top 5 web page tags using OWL-DL. The learning rules are as followings.

- If the tag rule of the webpage exists in the class “*DeliciousWebpage*”, we create a new instance to store the URL and the instance belongs to the subclass “*Condition_n*”, which contains the same tag rule. If the tag rule of the webpage does not exist in the class “*DeliciousWebpage*”, we create a new subclass “*Condition_{n + 1}*” and store the tag rule as a property in “*Condition_{n + 1}*”. We then create a new instance to store the URL and this instance belongs to the subclass “*Condition_{n + 1}*”.
- If we let the tag moves to the class “*Webpagetag*” in the ontology. As with the class “*DeliciousWebpage*”, if the tag exists in the class, we create a new instance to store the tag and the instance belongs to the subclass “*Tag_n*”, which contains the same classes such thatUS, USA, and America belong to the same subclass “USA”.
- If the tag does not exist in the class “*Webpagetag*”, we create a new subclass “*Tag_{n + 1}*” and create a new instance to store the tag and the instance belongs to the subclass “*Tag_{n + 1}*”.

4 Experiments and Evaluations

We conducted an experiment on a domain containing three US presidential candidates. In this experiment, we selected three USA presidential candidates in 2012 as the test domain. They were the Democrat Party’s Barack Obama, and Mitt Romney and Michele Bachmann from the Republican Party. There were 497,849 webpages on Obama, 12,271 webpages on Romney, and 4,570 web pages on Bachmann on theDelicious website. We selected 100 random webpages for each candidate for classification.

We selected sixty web pages of Obama, Romney, and Bachmann from the Delicious website as a training set to make the table for the decision tree.

And we calculated the information gain using the ID3 algorithm and determined all the attributes of the decision tree. Obama, Romney, and Bachmann were the candidates and these three domains were contained in the same tree, as shown in Fig. 3.



Fig. 3 The decision tree for three domains

In the decision tree, every pathway described by OWL-DL corresponded to a rule. Based on all of the rules in the tree, we produced an ontology for classifying the experimental data and separated the data into three classes, as shown in Fig. 4.

For the justification, we had calculated precision, recall and accuracy [16]. We compared the three types of domain user interests in the experiment with several other types of user interests in our previous studies. We checked 300 web pages and recorded their classes. Table 1 shows the calculation results for the precision, recall, and accuracy of the three domains for their classes. The first column lists the number of quality inbound links for 100 domain web pages. The other three

TBox/RBox	
(1)	ObamaWebpage \equiv DeliciousWebpage
(2)	ObamaWebpage \equiv hashtag \exists ((Obama \cap Election) \sqcup (Obama \cap _2012 \cap Politics))
(3)	ObamaWebpage \equiv \neg Romney
(4)	BachmannWebpage \equiv DeliciousWebpage
(5)	BachmannWebpage \equiv hashtag \exists (Bachmann \cap ((Politics \cap ($_2012$ \sqcup Election)) \sqcup ($_2012$ \cap Election)))
(6)	BachmannWebpage \equiv \neg Obama
(7)	RomneyWebpage \equiv DeliciousWebpage
(8)	RomneyWebpage \equiv hashtag \exists (Romney \cap Politics \cap (Election \sqcup _2012))
(9)	RomneyWebpage \equiv \neg Obama
(10)	RomneyWebpage \equiv \neg Bachmann
(11)	{Romney, Politics, Election, Obama, Bachmann, _2012...} \in \exists WebpageTag TagOf $_$ \equiv hashtag

ABox	
	\langle Webpage $_1, \dots, \text{Webpage}_n$ \rangle : hashtag
	\langle romney, politics, election, obama, bachmann, _2012, ..., \rangle : TagOf

Fig. 4 Ontology of the domains

Table 1 Results for precision, recall, and accuracy for the candidate Obama class

No.	Precision	Recall	Accuracy
60	0.875	0.824	0.917
65	0.882	0.833	0.923
70	0.895	0.811	0.914
75	0.895	0.811	0.920
80	0.895	0.811	0.925
85	0.900	0.818	0.929
91	0.905	0.826	0.923

columns show the results for precision, recall, and accuracy. The evaluation results showed our approach was reliable for determining the interests of domain users, because it had 91.5 % precision and 93.1 % accuracy for classification.

5 Conclusions

SNS is stimulating the development of the Internet. It accelerates the speed of information sharing between the people, and a vast amount of useful information is available on the interests of users. Users' interests have an important role for computer science and the social sciences. The interests of domain users have greater value, but domain interests have rarely been studied. The process to identify the interests of domain users presents two difficulties: defining user domains and classifying them.

In this study, we addressed these problems of defining user domains and their classification. We proposed an approach based on semantic web technology for identifying the interests of domain users from SNS. Domain ontology was generated by a domain decision tree that classified domain webpages and domain users using ID3 algorithm. Our experiment to determine the interests of domain users on USA presidential candidates showed that our approach had high precision and accuracy. We hope that our research will have value for computer science, such as recommendations, and for other fields such as education, political activities and commerce.

References

1. Zhuge H (2010) Socio-natural thought semantic link network: a method of semantic networking in the cyber physical society perth. In: 24th IEEE international conference on advanced information networking and applications, pp 19–26
2. Zhang T, Lee B, Kang S, Kim H, Kim J (2009) Collective intelligence-based web page search: combining folksonomy and link-based ranking strategy. *Computer and Information Technology*, 2009, pp 116–171

3. Pi S, Liao H, Liu S, Lin C (2011) Framework for classifying website content based on folksonomy in social bookmarking. In: Intelligent computing and information science, communications in computer and information science, vol. 135. pp 250–255
4. Illig J, Hotho A, Jäschke R, Stumme G (2011) A comparison of content-based tag recommendations in folksonomy systems. In: Knowledge processing and data analysis, Lecture Notes in Computer Science, vol. 6581/2011, pp 136–149
5. Shan S, Zhang F, Wu X, Liu B, He Y (2011) Ranking tags and users for content-based item recommendation using folksonomy. Computing and Intelligent Systems, Communications in Computer and Information Science, pp 32–41
6. Szomszor M, Alani H, Cantador I, O'Hara K, Shadbolt N (2008) Semantic modelling of user interests based on cross-folksonomy analysis. In: The semantic web—ISWC, Lecture Notes in Computer Science, 2008, vol. 5318/2008. pp 632–648
7. Kawase R, Herder E (2011) Classification of user interest patterns using a virtual folksonomy JCDL'11, Ottawa, Canada, ACM 978-1-4503-0744-4/11/06, 13–17 June 2011
8. Lipczak M (2008) Tag recommendation for folksonomies oriented towards individual users. In: ECML PKDD Discovery Challenge, pp 84–95
9. Yin D, Hong L, Xue Z, Davison, BD (2011) Temporal dynamics of user interests in tagging systems. In: Twenty-Fifth AAAI conference on artificial intelligence
10. Sasaki K, Okamoto M, Watanabe N, Kikuchi M, Iida T, Hattori M (2011) Extracting preference terms from web browsing histories excluding pages unrelated to users' interests. In: SAC'11, TaiChung, Taiwan, pp 21–25 March 2011
11. White RW, Bailey P, Chen L (2009) Predicting user interests from contextual information. In: 32nd international ACM SIGIR conference on research and development in information retrieval, ACM New York, USA, pp 19–23
12. Argentiéro P (1982) An automated approach to the design of decision tree classifiers. In: IEEE transactions on pattern analysis and machine intelligence, vol. Pami-4, no. 1
13. LópezMántaras R (1991) A distance-based attribute selection measure for decision tree induction. Mach Learn 6(1):81–92
14. Panigrahi S, Biswas S (2011) Next generation semantic web and its application. IJCSI Int J Comput Sci Issues 8(2):385–392
15. Gruber T (2008) What is an ontology. Encyclopedia of database systems, vol. 1. Springer-Verlag
16. vanRijsbergen CJ (1979) Information retrieval, Butterworth-Heinemann Newton, MA

Collective Intelligence Based Algorithm for Ranking Book Reviews

Heungmo Ryang and Unil Yun

Abstract IIR (Internet Information Retrieval) system searches important documents on the internet by measuring the importance of these documents. For this purpose, various ranking techniques are proposed and adopted. In this paper, we propose ReviewRank, a ranking technique for finding book reviews. With an increasing number of people buying books online, reviews of books written by other people have become more important. General ranking techniques measure the importance of documents based on references or quotations between documents through hyperlinks. However, they are not suitable for book reviews. In this paper, we analyze characteristics of the importance of book reviews based on voluntary participation or evaluation of people called as collective intelligence, and proposes measures for considering the importance. We also suggest a ranking algorithm which adopts ReviewRank for finding book reviews. Experimental results show that ReviewRank outperforms previous ranking techniques for both general IIR system and searching book reviews.

Keywords Book review · Information retrieval · Ranking technique · Collective intelligence

H. Ryang · U. Yun (✉)

Department of Computer Science, Chungbuk National University, 410, Gaesin-dong, Heungdeok-gu, Cheongju, Republic of Korea
e-mail: yunei@chungbuk.ac.kr

H. Ryang

e-mail: riangs@chungbuk.ac.kr

1 Introduction

There are a huge number of documents on the internet, and the number has been increasing. Therefore, the importance of searching meaningful documents on the internet has increased. IIR (Internet Information Retrieval) system looks for these important documents by measuring its importance through various ranking techniques. For measuring the importance of the documents, various ranking techniques [4, 10] have been proposed. Some of them measures the importance based on references or quotations between documents through hyperlinks [2, 3]. One of these ranking techniques is PageRank [9] which is adopted by Google search engine (<http://www.google.com>) which is the most famous and becomes the fundamental algorithm of IIR. However, it is difficult to measure the importance of book reviews using these ranking techniques since there are little references or quotations unlike general documents on the internet.

Recently, with an emerging of SNS (Social Network Service) such as Twitter and Facebook and explosive increasing numbers of data on it, researches for SNS information retrieval have been studied [6, 11]. Ranking techniques proposed in these researches measure the importance of an article written by user such as tweet in Twitter or a document based on network relationship constructed by users. However, network relationship between users of online bookstore is not almost constructed unlike between SNS users. Thus, ranking techniques for SNS information retrieval are not also suitable for measuring the importance of book review.

In this paper, we analyze characteristics of important book review based on voluntary participation or evaluation of people called collective intelligence and propose measures for considering the importance of book reviews. Moreover, we propose Review Rank, a ranking technique based on correlation analysis between the importance and the measures. By adopting ReviewRank proposed in this paper to IIR system, not only book reviews explicitly evaluated by users are important but also potentially important book reviews can be found. Potentially important book review means that it is not explicitly evaluated as important by voluntary participation or evaluation of people but it has similar characteristics to them. In this paper, we propose a ranking algorithm which adopts ReviewRank for searching important book reviews.

The remainder of the paper is organized as follows. We begin with describing influential related work in Sect. 2. In Sect. 3, we analyze characteristics of important book reviews based on evaluation by collective intelligence and propose ranking technique, ReviewRank, for measuring the importance of book review. Additionally, we propose ReviewRank algorithm for searching important book reviews. In Sect. 4, we show performance of our ranking technique through various evaluation experiments. Finally, Sect. 5 summarizes our contributions.

2 Related Work

TF-IDF (Term Frequency-Inverse Document Frequency) [1] which is one of traditional ranking techniques is a numerical statistic reflected how important a word is to a document in a collection. TF-IDF is calculated by multiplying tf which means frequency of a word in single document by idf which is a measure as general importance of word. In this paper, we employ not only ReviewRank for measuring the importance of book review but also TF-IDF for finding book reviews that contain words of query.

With the advent of Web 2.0, collective intelligence [8, 12] which means shared or group intelligence that emerging from the collaboration and competition of many individuals has appeared. A ranking technique measure the importance of document on the internet based on book marks by people has been proposed in [7]. In this paper, we propose ranking technique to measure the importance of book reviews in the online bookstores based on evaluations of users. The first research of ranking techniques for measuring the importance of book reviews in the online bookstores has proposed in [10], the approach considers characteristics of book reviews about the length of contents and the number of reply of book review for measuring the importance of book review. In this paper, we propose ReviewRank based on the number of reply, the average length of reply, and the rate of participation of the user who wrote the book review for focusing on user voluntary.

3 Ranking Based on Collective Intelligence

In this Section, we analyze characteristics of important book reviews in the online bookstores based on evaluations by collective intelligence and propose the measures for checking importance of book reviews based on analysis results. Moreover, we propose ReviewRank which is a ranking technique based on correlation [5] analysis between the importance and the measures and suggest a ranking algorithm which adopted ReviewRank for searching important book reviews.

3.1 Ranking Measures

To analyze characteristics of important book review, we use collected dataset which is described details in Sect. 4. The result is that the average number of evaluation tends to increase when the average length of reply or the number of reply written by reviewer's themselves, or the number of reply increases. It means that the probability becomes higher when the value of each measure becomes

higher. From the analysis, we propose the measures which are the average length of reply, the number of reply written by reviewer's self, and the number of reply for checking the importance of book reviews based on the analysis.

3.2 Analyzing Dataset Based on Collective Intelligence

ReviewRank proposed in this paper is a ranking technique based on results from evaluation of book reviews, and these reviews are targets in searching by collective intelligence. Correlation between each measure proposed in Sect. 3.1 can be different depending on evaluation to dataset by collective intelligence. Therefore, it is required to perform analyzing correlation between each measure of book reviews in dataset by retrieval system based on evaluation by collective intelligence. To analyze correlation, we calculate the Pearson correlation coefficient [5] between each measure with the number of evaluation by collective intelligence through as following Eq. (1). The Pearson correlation is defined only if both of the standard deviations are finite and both of them are nonzero, and it cannot exceed 1 in absolute value. If correlation between two variables x and y is strong, the Pearson correlation coefficient is close to either -1 or 1 .

$$r(x,y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \quad (1)$$

In Eq. (1), x is the set of the number of evaluations and y is the set of values of the each measure. The values of the Pearson correlation coefficient calculated through correlation analysis are used for determining the reflection rate of each measure in calculating ReviewRank of book review. If the same dataset and the different user groups who have the different tendency each other are given then the results of searching book reviews are different with the same query by retrieval system which adopts ReviewRank.

3.3 Calculating Review Rank

After analyzing correlation between each measure with the number of evaluations by collective intelligence, ReviewRank referring the importance of book review is calculated based on the analysis results. To calculate ReviewRank through Eq. (2), correlation coefficient of each measure is used as the adoption rate.

$$\text{ReviewRank} = \sum r_j(x,y) \times y_i \quad (2)$$

In Eq. (2), $r_j(x, y)$ is a correlation coefficient of each measure and y_i is the measure value of book review. That is, ReviewRank of book review can be

obtained through the sum of each multiplication of $r_j(x, y)$ by y_i . To search important book reviews contained words of query, the importance of each word in book review has to be calculated. Morphemic analysis is performed for calculating the importance of words contained book review by eliminating stop words such as prepositions. After that, the importance of each word appeared in book reviews of dataset is calculated. In this paper, we use TF-IDF for calculating the importance of each word through Eq. (3).

$$\text{TF-IDF} = tf \times (\log_2^N - \log_2^{d_k} + 1) \quad (3)$$

In Eq. (3), tf is a frequency of word in book review, N is the total number of book reviews in dataset, and d_k is the number of book reviews which contain the word. Finally, word importance is calculated by adopting ReviewRank to calculate TF-IDF value through Eq. (4).

$$\text{Ranking} = \text{TF-IDF} + (\text{TF-IDF} \times \text{ReviewRank}) \quad (4)$$

Values of calculated importance of each word based on ReviewRank referring the importance of book review contained those words are saved to inverted files. Inverted file which includes information for searching such as ranking score, url of book review, frequency of word is created by each word. Moreover, information in inverted file is sorted in descending ranking score order. If a measure is not correlated with the importance of book review then the measure cannot have an effect on the results of searching book reviews. We can apply various measures to ReviewRank even if some of them are only effective in a single target since its ranking score becomes zero when ReviewRank is applied in the other targets. Therefore, we can use only one ReviewRank algorithm for various targets by considering all measures for each target. In addition, ranking score of a word of Eq. (4) in a book review is always greater than ranking score of the book review of Eq. (2).

3.4 Algorithm Based on ReviewRank

In this Section, we propose and illustrate a ranking algorithm adopting ReviewRank for implementing IIR system of book review based on collective intelligence. Fig. 1 shows the algorithm proposed in this paper. To analyze criteria for determining book reviews as important collected dataset from a certain target, Analyzing function uses the Pearson correlation coefficient and calculates the adoption rate of each measure with tendency of collective intelligence. For this, Analyzing function first extracts values from book reviews (line 3), where values indicates each measure in the book review such as the number of replies. Then, extracted values are added to MVL list which is the values list of the measures (line 4).

```

Function Analyzing( dataset )
1. Set all of values in AL to 0
2. For each book review in dataset
3.   Extract values from book review
4.   Add extracted values of the measures to MVL list
5.   Add extracted value of evaluation to EVL list
6.   Calculate sum of extracted values and values in AL
7. end
8. Calculate averages of each measure in AL
9. For each values of the measures of book review in VL
10.  Calculate the Pearson Correlation Coefficient
11.  Add calculated value to CL
12. end

Function CalcWeight( dataset, MVL, CL )
1. For each book review in dataset
2.   weight  $\leftarrow$  0
3.   For each value of the measure in MVL
4.     Multiply measure's value by the Pearson Correlation Coefficient
5.     weight  $\leftarrow$  weight + calculated_value
6.   end
7.   Add weight to WL
8. end

Function CalcRank( dataset, WL )
1. For each book review in dataset
2.   weight  $\leftarrow$  weight value of current book review in WL
3.   Extract keywords from book review by analyzing morphemes
4.   Add extracted morphemic keywords to KL
5.   For each morphemic keyword in KL
6.     Calculate TF-IDF value of morphemic keyword
7.      $RR \leftarrow TF-IDF + ( TF-IDF \times weight )$ 
8.   end
9. end

```

Fig. 1 Algorithm based on collective intelligence for searching book reviews

Especially, evaluation values in respect of the book review by collective intelligence are required for determining and analyzing the importance of the book review. Thus, the evaluation value is also extracted (line 3) and added to EVL list (line 5), where EVL list refers to evaluation value list. After that, sum of each extracted value is calculated and accumulated in AL which means each accumulated measure value (line 6). Averages of each accumulated value in AL are calculated (line 8) for computing the Pearson correlation coefficient (line 10).

After analyzing the dataset, CalcWeight function is called to calculate the importance of each book review, and MVL and CL, which refers to the calculated Pearson correlation coefficient values, are delivered for this. CL is used as the adoption rate of each measure. CalcWeight function computes weights of each

book review in dataset using Eq. (2) (lines 1 to 6), and then the weight is added to WL (line 7), where WL refers to book review weight list. Finally, ranking score of each word in book review is calculated by CalcRank function. In this stage, weight of each book review is calculated by CalcWeight function is used to calculate final ranking scores. A ranking score of a word is computed based on weight of a book review which contains that word. For this, CalcWeight function first extracts keywords from the book review by morpheme analysis (line 3), and then each extracted keyword is added to KL which is keyword list (line 4). For each morphemic keyword in the KL, first TF-IDF value is calculated using Eq. (3) and then ranking score is computed by Eq. (4) (lines 6 to 7).

4 Experiments

In this Section, we present performance evaluation of ReviewRank with TF-IDF and LengthRank which is proposed in [10] for evaluating importance of book review. Moreover, we also perform experimental evaluation with TF-IDF and Google search engine (<http://www.google.com>) by using search operator to limit target dataset as book reviews in online bookstores.

We have collected book reviews (about 0.11 million reviews) from Amazon (<http://www.amazon.com>) and GoodReads (<http://www.goodreads.com>) which are online bookstore and collecting book reviews site respectively. That is, the dataset used in every experiment reflects tendencies of both user groups in Amazon and GoodReads. In this Section, we present performance evaluations of precision and recall. All algorithms were written in Microsoft Visual C ++ 2010, and run with the Windows 7 operating system on an Intel Pentium quad-core 3.2 GHz CPU with 8GBytes main memory.

4.1 Evaluation of the Precision

We compare performance of ReviewRank with LengthRank, TF-IDF, and Google search engine by evaluating precision. Precision is the rate of retrieved relevant reviews by the searched K reviews. We used 5 random keywords and set K as 50 for evaluation of precision.

Figure 2a presents the results of precision evaluation of ReviewRank with TF-IDF and LengthRank which use the same dataset. In this paper, we refer to important review as that a review has greater than and or equal to the average value of evaluation of collected dataset. We can observe that ReviewRank outperforms TF-IDF and LengthRank in every case of Fig. 2a. Since ReviewRank is based on collective intelligence and book review is determined as important by user groups while the others do not apply that.

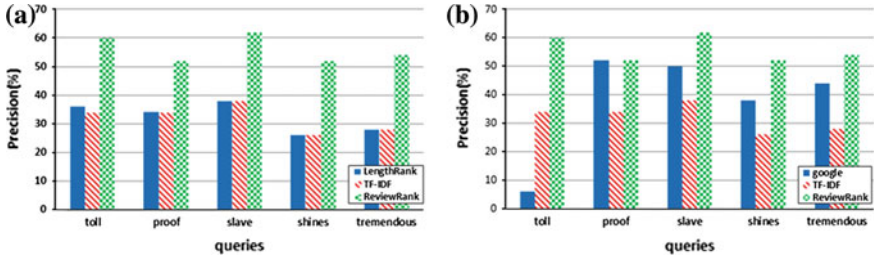


Fig. 2 Precision evaluation of reviewrank with TF-IDF, lengthrank, and google

Fig. 3 Recall evaluation of reviewrank with TF-IDF and lengthrank

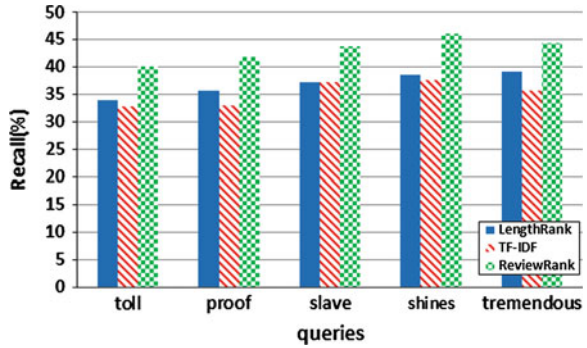


Figure 2b is a graph of the results of precision evaluation of ReviewRank with TF-IDF and Google search engine, and shows that ReviewRank outperforms Google on searching book reviews. Since ranking technique adopted by Google measures importance based on references or quotations while there are little references or quotations between book reviews. Therefore, we can see that the ReviewRank shows the better performance in almost every case. In addition, Google also outperforms TF-IDF in almost case except when a keyword “toll” is used.

4.2 Evaluation of the Recall

Second, we evaluate performance through recall which is the fraction of reviews relevant to the retrieved reviews using 5 queries which are the same as queries used in Sect. 4.1. Note that answer dataset is required for performing evaluation of recall. Therefore, we perform recall experiment of only ReviewRank and LengthRank so that they can use the same with answer dataset.

Figure 3 is a graph of the results of recall evaluation, and we can observe that ReviewRank shows performance better than LengthRank and TF-IDF in every case. In addition, LengthRank also outperforms TF-IDF in every case.

5 Conclusions

In this paper, we proposed ranking technique, ReviewRank, based on collective intelligence for evaluating importance of book reviews and presented a ranking algorithm adopting ReviewRank for searching book reviews. From the performance test, the experimental results showed that ReviewRank outperformed previous ranking techniques for both the internet information and book review retrieval on Precision and Recall. We expect that our research will take effects to not only searching book reviews but also retrieval area based on collective intelligence.

Acknowledgments This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 2012-0003740 and 2012-0000478).

References

1. Aizawa AN (2003) An information-theoretic perspective of Tf-idf measures. *J Inf Process Manag* 39(1):45–65
2. Alyguliev RM (2007) Analysis of hyperlinks and the ant algorithm for calculating the ranks of web pages. *ACCS* 41(1):44–53
3. Dou Z, Song R, Nie JY, Wen JR (2009) Using anchor texts with their hyperlink structure for web search. *SIGIR*, New York, pp 227–234
4. Duan Y, Jiang L, Qin T, Zhou M, Shum HY (2010) An empirical study on learning to rank of tweets. In: *COLING* pp 295–303
5. Egghe L, Leydesdorff L (2009) The relation between Pearson’s correlation coefficient r and Salton’s cosine measure. *CoRR*
6. Gayo-Avello D (2010) Nepotistic relationships in Twitter and their impact on rank prestige algorithms. *CoRR*
7. Huang JJS, Yang SJH, Chen JYL, Li I, Hsiao IYT (2010) A social bookmarking-based people search service building communities of practice with collective intelligence. *IJOI* 1(2):83–95
8. Krol D, Lopes HS (2012) Nature-inspired collective intelligence in theory and practice. *Inf Sci* 182(1):243–263
9. Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: bringing order to the web. Technical Report, Stanford InfoLab
10. Ryang H, Yun U (2011) Effective ranking techniques for book review retrieval based on the structural feature. *Lecture Note in Computer Science, ICHIT, CheJu Island*, pp 360–367
11. Teevan J, Ramage D, Morris MR (2011) #TwitterSearch: a comparison of microblog search and web search. *WSDM*, pp 35–44
12. Tumer D, Wolpert D (2011) Collective intelligence, data routing and Braess’ paradox. *CoRR*

Ranking Techniques for Finding Correlated Webpages

Gwangbum Pyun and Unil Yun

Abstract In general, when users try to search information, they can have difficulties to express the information as exact queries. Therefore, users consume many times to find useful webpages. Previous techniques could not solve the problem effectively. In this paper, we propose an algorithm, RCW (Ranking technique for finding Correlated Webpages) for improving previous ranking techniques. Our method makes it possible to retrieve not only basic webpages but also correlated webpages. Therefore, RCW algorithm in this paper can help users easily look for meaningful information without using exact queries. To find correlated webpages, the algorithm applies a novel technique for computing correlations among webpages. In performance evaluation, we test precision, recall, and NDCG of our RCW compared with the other popular system. In this result, RCW guarantees that it finds the number of correlated webpages greater than the other method, and shows high ratios in terms of precision, recall, and NDCG.

Keywords Webpage analysis · Correlation searching · Ranking technique · Information retrieval

G. Pyun · U. Yun (✉)

Department of Computer Science, Chungbuk National University, 410, Gaesin-dong, Heungdeok-gu, Cheongju, Republic of Korea
e-mail: yunei@chungbuk.ac.kr

G. Pyun

e-mail: pyunb@chungbuk.ac.kr

1 Introduction

Recent internet information retrievals have deviated general keyword-based retrieval manner, and various approaches have been actively studied such as applying various effects, using user's information [1], understanding contents of webpages [2, 3], using associated queries [4, 5], and utilizing webpage's links [2]. However, although there are previous algorithms that classify topics identifying webpages or look for webpages closer to queries, they still have constraints in terms of methods analyzing and matching queries [6]. Since most of the users input a part of keywords or relevant keywords to find useful information, they may try to find inaccurate webpages many times. Our algorithm proposed in this paper, RCW (Ranking technique for finding Correlated Webpages) can seek valid webpages although users even input a part of keyword or inaccurate keywords. The more correlated webpages are found, the higher probability of getting meaningful webpages becomes. Therefore, this correlated retrieval is an expanded method more than methods by directory selection or query. Procedure of our algorithm is as follows. If users require queries in the algorithm for internet retrieval, the algorithm searches correlated webpages comparing contents with queries. Here, using ranking technique proposed in this paper, the algorithm seeks valuable webpages according to their correlated ranking order. Organization of this paper is as follows. In Sect. 2, we describe related work for the proposed algorithm. Section 3 presents information of our algorithm, RCW, and Sect. 4 conducts performance evaluation of our proposal compared with a famous ranking technique. Finally, we describe conclusions in the Sect. 5.

2 Related Work

In information retrieval field, algorithms with various retrieval effects have been studied. PageRank [2] used in Google (www.google.com) computes relevance of webpages identifying links in webpages and retrieves valid webpages according to queries. In [7], authors proposed an algorithm for retrieving information by using not only queries but also other data such as GPS coordinate data, where the algorithm searches for a variety of local information in Singapore referring to user-location data. PIC [8] is a method based on topics, which conducts information retrieval utilizing popularity of webpages. The algorithm extracts popular keywords from webpages of any website, analyzes the topic of the website through the extracted results, and then servers information based on the topic to users. WAS [9] provides techniques for analyzing and indexing webpage structures. This performs clustering webpages depending on their structure types to retrieve webpages easily. MLNs [10] uses Markov Logic Network so as to obtain webpage information. The algorithm converts word-connection data composing webpages to graph forms and constructs Markov model using the graph data, thereby extracting important topics and keywords effectively.

3 Technique for Finding Correlated Webpages

3.1 Calculating Correlation List of Webpages

Our algorithm uses words of webpages to consider correlation among webpages. Webpages having a correlated topic commonly use the same or correlated words. For example, if a webpage selected by user is related to medical science, the webpage frequently uses words related to medical science. Then, any webpage correlated to the webpage will also use the same words. Therefore, we have to compare words of the selected webpage to that of the other webpages. Given two webpages A and B, a method for calculating the number of words between them is as follows. (1) Bring a list that organizes all the words in the webpages. (2) Match words of the webpages A and B with indexing numbers in the list. (3) Sort A and B in ascending order of index numbers. (4) Calculate indexing number of which A and B are matched. (5) In the result, the number of (4) becomes correlated score. That is, given correlated score, S and certain webpages A and B, keywords consisting of A are $a_1, a_2, a_3, \dots, a_n$, i.e. $A = \{a_1, a_2, \dots, a_n\}$ and keywords consisting of B are b_1, b_2, \dots, b_m , i.e. $B = \{b_1, b_2, \dots, b_m\}$, O is a set of the words that A and B are matched. This is denoted as follows.

$$O = \{o_1, o_2, o_3 \dots o_r | r \leq n, n \leq m\} \tag{1}$$

$$S = \sum_{i=1}^r o_i \tag{2}$$

Correlated score between two webpages can be calculated by (2). If correlated score is high between webpages, it can be determined that the two webpages are highly correlated webpages since they use many same words. Figure 1 shows a method for calculating the number of the same words between two webpages, A and B. Consider the webpage A has n words and B has m words. If $a_1 = b_1$, $a_3 = b_4$, and $a_4 = b_6$, then the number of words in the set, O is 3. Therefore, the correlated score between the webpages A and B becomes 3.

Lemma 1 Given certain webpages A, B, and C, a correlated score between A and B denotes S_{ab} , and one between A and C denotes S_{ac} . If $S_{ab} < S_{ac}$, A is correlated with B more than C.

Proof A correlated score of two webpages is decided as words which consist of webpages. Given webpages P, Q, and R, if $P = Q$, S_{pq} = the word number of P = that of Q. On the other hand, if $P \neq R$, $S_{pr} \leq S_{pq}$ since each page has a different set of words. Therefore, certain two pages with different contents have a value, S lower than those with the same contents. According to this property, correlation of A with C becomes more relevant webpages than that of A with B if $S_{ab} < S_{ac}$. □

However, it is not possible to calculate correlated scores of all webpages whenever users search. Thus, we calculate correlated scores in advance.

Fig. 1 Calculation of correlated score

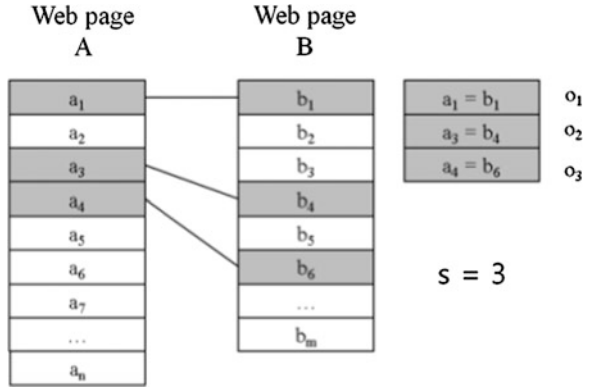
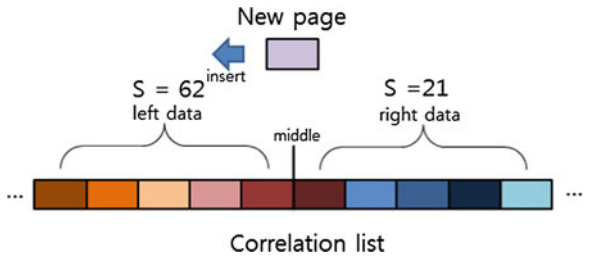


Fig. 2 Inserting new page in the correlation list



Naïve solution is that if there is N webpages, then $N-1$ correlated scores are saved to each webpage. However, it is an inefficient solution since this method has to store N^2 correlated scores in total. Therefore, we propose an efficient method to save the correlation rate. The method assigns a unique number which indicates characteristics of webpage to each webpage. Webpages with the high correlation have unique numbers close to each other, and vice versa. If a user selects a certain webpage then it shows webpages which have adjacent unique numbers. This method is very efficient since it requires only N spaces, and it is based on binary search to assign the unique numbers as shown in Fig. 2.

The method searches left data if a target is smaller or right data if it is larger compared with a central data of sorted data. However, general binary search cannot find a unique number according to webpage’s contents. The reason is that the algorithm cannot know whether a topic of a collected webpage is higher than that of a compared webpage. For this reason, we use a variant of original binary search which can compare left or right representative data based on the central data. Using this manner, we know whether a collected webpage has a disposition of left data or right data. Comparing the dispositions, correlated score, S comparing two webpages is used. Here, the method computes S between a webpage of left representative data and a collected webpage, and does S between the right and that. In the two representative data, a side with a higher score is more likely to be a position of data correlated with a disposition of a collected webpage. After comparing data by modified binary search, a central data becomes an edge data

while the webpage with a higher score becomes a central data. Then, in common with original binary search, this compares correlated score using a new left or right representative data, and reduce a scope gradually. Iterating this process continuously, adjacent two data are remained eventually. The algorithm inserts the collected webpage into the right position of larger data compared with the adjacent two data, and the collected webpage has a unique number. When a certain webpage is selected after assigning unique numbers in all collected webpages by this variant manner, webpages adjacent to the unique number have a word structure correlated with the selected webpage. That is, they become webpages getting correlated contents.

Lemma 2 If webpages, A and B are adjacent in correlation list, these two pages are closely correlated in terms of contents.

Proof Our algorithm inserts a webpage into a corresponding position of correlation list based on binary search method to consider left or right correlated score. The algorithm recursively conducts a search moving toward S with a higher value. Supposing that I is a inserted page, M is a moved page depending on above property, and R is a page with a reverse orientation of M, $S_{im} \geq S_{ir}$ is always satisfied. Webpages near to the inserted page therefore become correlated webpages with the highest values of correlation list. Consequently, the lemma 2 is clear due to the above properties. □

3.2 Ranking Technique

The algorithm computes ranking of webpages to search webpages correlated to what users want. Ranking manner is based on tf-idf [11]. When users request queries in the retrieval algorithm to look for internet information, the algorithm identifies contents from the words of a query and retrieves correlated webpages using tf-idf manner. Then, if users select a based webpage, this shows webpages correlated with the selected page. Given a based page selected from a user, selected page and correlated page adjacent to the selected page, target page, we can calculate distance of correlation and then obtain new ranking scores according to the computed distance. Given term frequency, tf and inverse document frequency, idf, ranking score depending on correlated distance proposed is as follows.

$$Correlation\ range = \frac{1}{\sqrt{|selected\ page - target\ page|^2}} \tag{3}$$

$$Ranking\ score = tf \times idf \times (1 + Correaltion\ range) \tag{4}$$

Selected page means a unique number in a selected webpage and target page reflects a unique number in a webpage of which ranking score is calculated. Correlation range is to compute distance of unique numbers between selected page

and target page. The more adjacent correlation range is, the closer a value is to 0. In contrast, the more distant it is, the closer one is to 1. To compute new ranking scores, our algorithm uses unique numbers of correlation list stored in inverted index. In order to show correlated webpages close to the based page to users, the algorithm includes word frequency of queries in operating ranking score. Since users input queries to obtain meaningful information, the algorithm has to retrieve corresponding webpages near to the based page or topic.

3.3 RCW Algorithm

In this section, we present RCW algorithm which implements correlated webpage retrieval techniques. Figure 3 shows overall RCW algorithm and sub functions. RCW algorithm divides words of webpages in the first line, and then finds keywords. Then, the algorithm assigns indexing numbers to the words comprising webpages. In the line 4, RCW searches a position of the currently inputted webpage in correlation list collected through `find_location` function. RCW algorithm is terminated after inputting the new webpage information into the searched position. Iterating this algorithm, all of the webpages are sorted depending on their correlation and saved in the inverted index. Then, these are utilized to calculate ranking scores.

4 Performance Evaluation

In this Section, we run performance evaluation of proposed ranking technique. Our ranking technique is run with Intel Pentium 3.2 Ghz CPU, 8 Gbyte memory, Window7 OS, and C++ language. For the morpheme analysis for the RCW, we use clucene [12]. As experimental data, the collection of internet news in <http://washingtonpost.com> is used for the performance evaluation, where target data range from 01/01/2011 to 12/31/2011. All of the collected webpages have each category and the number of that webpages are 32,227, where the categories of the collected webpages consist of politics, business, life, style, entertainment, sports, region, and world. Since these categories are actually classified by The Washington Post Newspaper, that is reliable data. Our algorithm uses the above categories as an important element when evaluating correlated webpages. Different webpages with the same category can be regarded as the correlated webpages. In other words, since each category is classified by topic, the same category can have the same topic, and thus we can decide that these webpages with the same category are the correlated webpages.

Fig. 3 Algorithm for RCW

```

Algorithm RCW
input: correlation list L
      new Webpage
      word list W
output: extended correlation list L'
begin
1. isolation words in a new Webpage
3. page word list ← find_word_ID(p, W)
4. insert location p ← find_location(p, L)
5. insert W in the p
6. return L'
end
function find_location(left location, right location, p, L)
output : insert location
begin
13. mid ← [(left + right)/2] // mid
14. left data ← [(left + mid)/2] // left data
15. right data ← [(mid + right)/2] // right data
16. Sleft ← calculate_correlation(left data)
17. Sright ← calculate_correlation(right data)
18. if Sleft < Sright then finding left area
19. if Sleft > Sright then finding right area
end
function calculate_correlation(page A, page B)
c : unique number list of A
r : unique number list of B
begin
20. foreach c ← A
21.   for each r ← B
22.     if c = r then increase count
23.   end foreach
24. return count
end
    
```

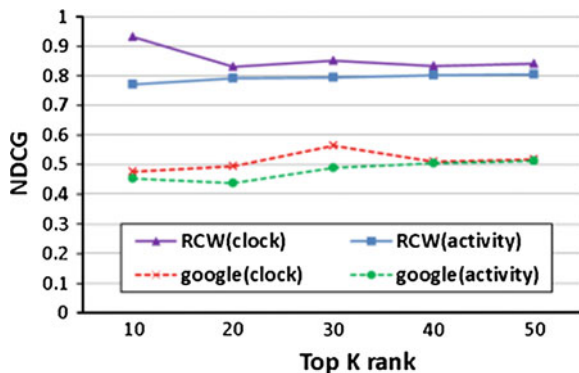
4.1 NDCG Test

To evaluate performance of our proposal, RCW and Google, we applied Normalized Discounted Cumulative Gain (NDCG) method. NDCG has a high value if relevant document is located in the high rank and a low value if that is located in low rank. NDCG is calculated by the following formula.

$$NDCG_k = \frac{DCG_k}{IDCG_k} DCG = rel_1 + \sum_{i=2}^k \frac{rel_i}{\log_2 i}$$

NDCG is defined as ratio between DCG and IDCG. If K is a top-Kth webpage, rel is a relevant score of a top-i-th webpage in the searched webpages. Relevant score is assigned as follows: 3 scores when a certain webpage is the same category as the selected webpage and their contents are closely correlated, 2 scores when they have the same category and are relevant, 1 scores when they have the same category and are not relevant or when they do not have the same category and are relevant, or 0 scores when they do not have the same and are not relevant. IDCG is to calculate DCG after sorting top-K webpages that is searched as the ideal ranking order in relevant score descending order. For the evaluation, we first conduct

Fig. 4 NDCG test



keyword retrieval and select a base webpage, and then measure the searched result as NDCG using RCW ranking method. Figure 4 presents NDCG result of the two keywords, 'clock' and 'activity', where K's range of these is from 10 to 50. In the result, our RCW received at least 0.7 or more in all cases while Google received lower values than that of RCW in all cases.

4.2 Precision (Mean Average Precision) Test

Second, we test precision and Mean Average Precision (MAP) of RCW algorithm. Data used in this experiment are also the collection of the news webpage in common with the above case. We compare ours with the Google's algorithm, PageRank. For the fair evaluation, searching data of both two techniques is limited to the news webpages of the Washington Post Newspaper in 2011. Here, criteria of evaluation, precision means the ratio that user-desired webpages exist in the top-K webpages. Keywords used in the experiment are randomly created. Assume that K is 50 and it is used for precision evaluation. Webpages that users want target the internet news pages with a category equal to the user-selected webpage. Figure 5 shows the result of precision evaluation based on ten random keywords. As a result, RCW technique showed higher precision than Google. Google generally showed 10–20 % precision ratio while our RCW showed at least 45 % or more. In the next experiment, we evaluated MAP of the two algorithm increasing top-K. K is increased from 20 to 100, and average precision is computed with ten keywords. Figure 6 illustrates MAP of the keywords. In common with Fig. 6, RCW presented more precision result than that of Google in all cases. Since the webpages with top ranking become the most correlated webpages, the precision of RCW algorithm becomes higher.

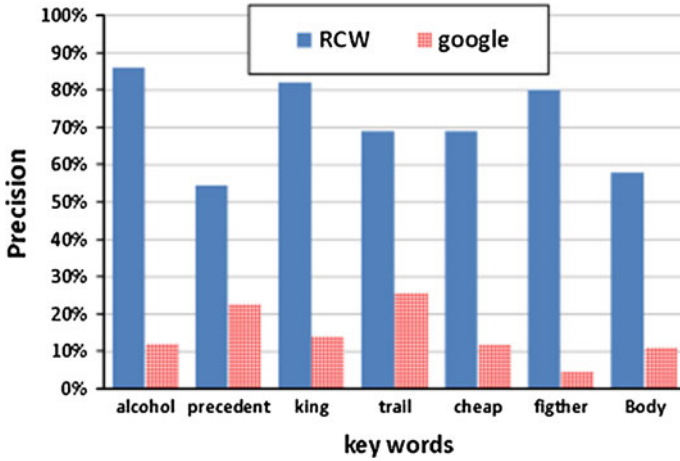
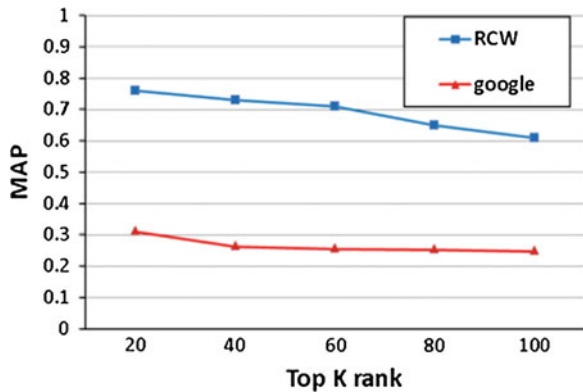


Fig. 5 Precision test of random keywords

Fig. 6 MAP test of random keywords



4.3 Recall Test

Third, we experiment recall comparing our RCW with Google, where recall means a ratio of which the ranking technique finds corresponding answer pages in all of the answer pages. This answer pages for the recall measurement are in advance made using categories of newspapers. Answer pages becoming criteria of evaluation are the webpages with the category and keyword equal to the selected page. Recall is a ratio of which corresponding answer pages occurs in all searching result. In common with above, RCW is compared to Google. Then, ten random keywords in this evaluation are used and random webpages related to these keywords are used in the experiment. Figure 7 shows the result of recall evaluation in ten keywords. From Fig. 6, RCW shows better results than that of Google in all

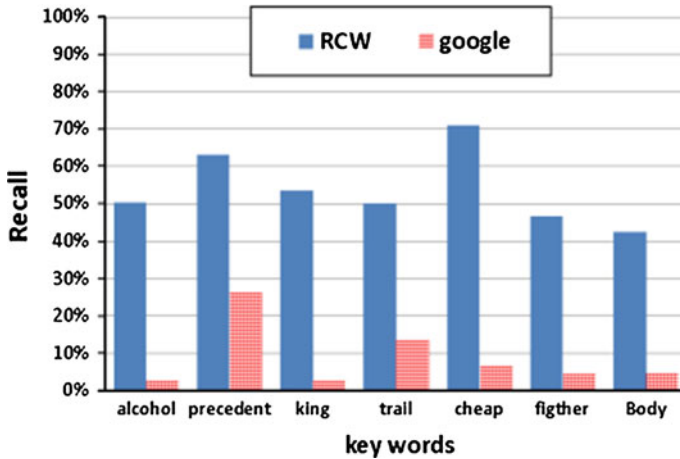


Fig. 7 Recall of random keyword

cases. Google shows about 10 % recall result on average while RCW shows higher recall result.

5 Conclusion

In this paper, we proposed a ranking algorithm that finds correlated webpages efficiently. Our algorithm finding correlated webpages is a more expanded and improved method than previous keyword or directory search method. In the experimental result, our algorithm, RCW showed higher precision and recall than those of previous retrieval technique. We expect that this proposed algorithm has a various effect in internet information retrieval.

Acknowledgments This research was supported by the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 2012-0003740 and 2012-0000478).

References

1. Hulth A, Karlren J, Jonsson A, Bostrom H, Asker L (2010) Automatic keyword extraction using domain knowledge. *Lect Notes Comput Sci* 472–482
2. Ishii H, Tempo R (2010) Distributed randomized algorithms for the page rank computation. *IEEE Control Syst Soc* 55(9):1987–2002
3. Ermelinda O, Massimo R (2011) Towards a spatial instance learning method for deep web pages. In: *Industrial conference on data mining (ICDM)*, pp 270–285

4. Fu L, Meng Y, Xia Y, Yu H (2010) Web content extraction based on webpage layout analysis. In: Information technology and computer science (ITCS), pp 40–43
5. Baillie M, Carman M, Crestani F (2011) A multi-collection latent topic model for federated search. *Inf Retrieval* 14(4):390–412
6. Ricardo Y, Carlos C, Flavio J, Vassilis P, Fabrizio S (2007) Challenges on distributed web retrieval. In: International conference on data engineering, pp 15–20
7. Flora T (2011) Web-based geographic search engine for location-aware search in Singapore. *Expert Syst Appl (ESWA)* 38(1):1011–1016
8. Song G, Yajie M, Liu Y, Chunping L (2009) Topic-based computing model for web page popularity and website influence. In: Australasian conference on artificial intelligence, pp 210–219
9. Costantinos D, Christos M, Yannis P, Evangelos T, Athanasios T (2010) A web page usage prediction scheme using sequence indexing and clustering techniques. *Data Knowl Eng (DKE)* 69(4):371–382
10. Sandeepkumar S, Sahely B, Sundararajan S, Rajeev R, Prithviraj S (2011) Web information extraction using markov logic networks. In: Knowledge discovery and data mining (KDD), pp 1406–1414
11. Metzler D (2008) Generalized inverse document frequency. In: Conference on information and knowledge management, pp 399–408
12. CLucene Project web page <http://clucene.sourceforge.net/>

Square-Wave Like Performance Change Detection Using SPC Charts and ANFIS

Dong-Hun Lee and Jong-Jin Park

Abstract While developing software products, performance regressions are always big issues in enterprise software projects. To detect possible performance regressions earlier, many performance tests are executed during development phase for thousands or ten thousands of performance metrics. In the previous researches, we introduced an automated performance anomaly detection and management framework, and showed Statistical Process Control (SPC) charts can be successfully applied to anomaly detection. In this paper, we address the special performance trends in which the existing performance anomaly detection system hardly detects the performance change especially when a performance regression is introduced and recovered again. Generally the issue comes from that the fluctuation gets aggravated and the lower and upper control limits get relaxed with the fixed sampling window size while applying SPC charts. To resolve the issue, we propose to apply automatically tuned sampling size, and to build the optimized Fuzzy detection system. ANFIS is adopted as a Fuzzy inference system to determine the appropriate sampling window size. Using the randomly generated data sets, we tune fuzzy rules and fuzzy input/output membership functions of ANFIS by learning. Finally we show simulation results of the proposed anomaly detection system.

Keywords Performance anomaly · Statistical process control (SPC) chart · Fuzzy theory · Adaptive neuro-fuzzy inference system (ANFIS)

D.-H. Lee

SAP Labs Korea TIP, Banpo-dong, Seocho-gu, Seoul, South Korea
e-mail: dong.hun.lee@sap.com

J.-J. Park (✉)

Department of Internet, Chungwoon University,
Hongseong-eup, Hongseong-gun, Chungcheongnam-do, South Korea
e-mail: jjpark@chungwoon.ac.kr

1 Introduction

During recent decades, performance issues have become more critical to release a software product supporting diverse functionalities in enterprise software projects. To prevent performance regression during software development and reduce the overall cost to fix the issues, it is widely used to run performance tests not just before releasing the product but during development. More than thousands of performance metrics are monitored and more than hundred thousands of performance measurements are estimated even one day to find possible performance regressions. Such performance results having big data characteristics cannot be monitored manually, but require a kind of automated performance monitoring, analysis and reporting system for swift feedback and lower cost to keep the desirable performance.

For automated performance monitoring and estimation, we need a periodic build and test system for the source code under development, and need an efficient result management repository. In the previous researches [1, 2], we proposed an effective software development process to enable the fast feedback of performance regressions, and introduced a performance anomaly detection and analysis framework to reduce the overall cost for performance issue management. We introduced statistical process control (SPC) charts as a tool to detect the performance anomalies and addressed how to build an automated anomaly detection system considering the characteristics of software performance trends.

In this paper, we address the special performance trend in which the existing anomaly detection system can hardly detect the noticeable performance change, and propose the revised detection algorithm which dynamically tunes the sampling window size considering the characteristics of each performance trend, and applying Fuzzy theory to find an appropriate size of the moving window. This problematic situation happens when a performance regression is introduced and recovered again a while later, and the fluctuation gets aggravated and the lower and upper control limit get relaxed so that the existing system hardly detect the noticeable performance change. In another paper to be published soon, we already showed the feasibility of our approach using the real performance data. Here we introduce the optimized Fuzzy detection system built by hybrid learning mechanism of Adaptive Neuro-Fuzzy Inference System (ANFIS) using randomly generated sample data.

In Sect. 2, we present our general approach to detect performance anomalies among massive performance results. In Sect. 3, we explain the fuzzy system by ANFIS to automatically tune sampling window size while applying SPC charts. The simulation results are presented in Sect. 4, and Sect. 5 concludes the paper.

2 Software Performance Anomaly Detection

2.1 Big Data Analysis Using SPC Charts for Performance Anomaly Detection

We showed in [1] that SPC charts are very effective to detect performance anomalies during software development. SPC charts are mainly used in manufacturing process to keep the stable production process, and many charts have been already researched [3]. Recently they began to be applied to software engineering area as seen in [4, 5]. In [1], we also showed that too many past samples can make the detectability worse while applying SPC charts to performance anomaly detection, because the performance trend of a software product keeps changing slightly as the source code is being modified, and too much old data can increase the standard deviation and fluctuation. In [1], we proposed that about 40–50 samples are appropriate for SPC charts in software performance monitoring.

As seen in Fig. 1, we maintain a fixed size (W) of sampling window including the last performance measurement, and calculate the required statistics to apply SPC charts. The exceeded data from upper/lower limits in the charts are detected as anomalies.

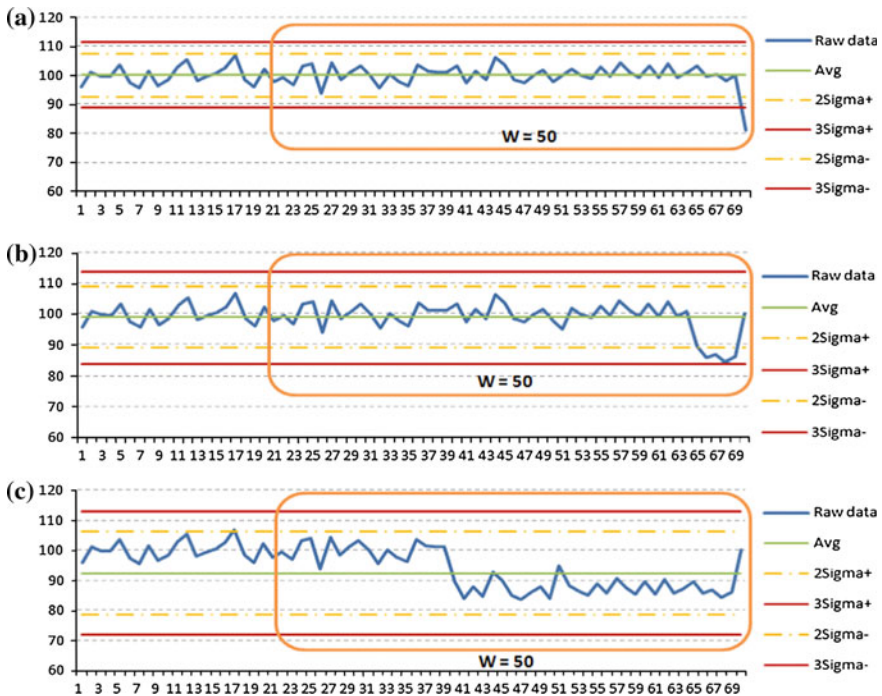


Fig. 1 Performance trend having different recovery time after a performance regression

We don't report all anomalies found in the charts, but only the ones within the recent N samples to reduce the duplicated anomaly reports and prevent missing ones by developers. The N value can be configured for each test according to the test duration or their characteristics. One more benefit from this report window is that we can notify a performance recovery indirectly, when a performance has been recovered and there is at least one anomaly within the recent N samples. In the case that a performance is recovered after a performance drop, it is not easy to detect it by SPC charts because the recovered performance falls in the normal range and the existing anomalies also fall in the normal range due to the increased standard deviation by big fluctuation and relaxed control limits in SPC charts.

Figure 1a shows that a performance regression occurred in the latest test results, which would be easily detected by SPC charts. The chart (b), and (c) in Fig. 1 depict that the performance has been recovered after a while later, but the dropped performance increases the standard deviation within the sampling period and relaxes the upper/lower limits in SPC charts, which causes that almost all data get to fall in the normal range. In this paper, we address how to report this kind performance change.

2.2 SPC Chart with Moving Window Using Automatically Tuned Sampling Period

In the previous section, we showed that the existing fixed sampling window has some limitation to detect the performance change once a performance anomaly has occurred. To handle this kind of performance change, we propose the flexible sampling window size according to the performance trend as seen in Fig. 2. In case that the performance is recovered in a short time period like in Fig. 1b, the recent performance drop can be detected by SPC charts by increasing the sampling window ($W1$ to $W2$) as seen in Fig. 2a, because the old stable values help to keep the standard deviation small and prevent the control limit relaxed too much. Rather, in case that it takes a bit long time to make the performance recovered like in Fig. 1c, the decreased window size would help detect the performance change. Within the shorten sampling period $W2$ in Fig. 2b, the dropped performance become a new base line and the recovered performance can be detected as a big performance change by SPC charts.

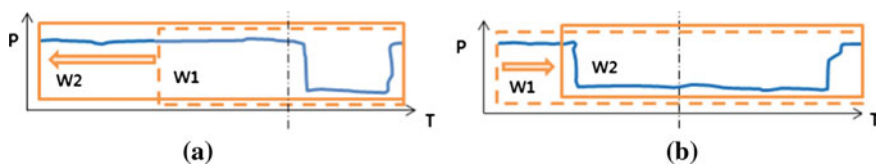


Fig. 2 Flexible sampling window on applying SPC charts

The main issue to apply this idea is how to determine the appropriate sampling size according to the diverse performance trend. In Sect. 3, we explain how Fuzzy theory can be applied to handle this issue.

3 Design of a Fuzzy System to Determine an Appropriate Sampling Window Size

3.1 Fuzzy Input and Output Variables

It is widely known that fuzzy inference system can express effectively a complex and uncertain system, which can be hardly defined by the existing mathematic models, using a set of fuzzy IF–THEN rules. First, we should define the input/output variables to determine the appropriate sampling window size using Fuzzy inference system. While monitoring the various performance trends having performance regressions and recoveries, we noticed that several statistical values show big differences between the first and second half of a sampling window, and there are some relationships among them. Based on this observation, we propose the following three statistical values as the input variables of our Fuzzy model. The first one is the overall fluctuation over the whole sampling period as expressed in Eq. (1), which indicates if there is a big performance change or not within the sampling window. The second one is the difference of the mean values of the first and second half of the sampling window as expressed in Eq. (2). The last one is the fluctuation of the first half of the sampling window as defined in Eq. (3). The sampling window size becomes the output variable.

$$Fluctuation_All = 100 * Stdev / Mean_All \quad (1)$$

$$DM = |Mean_H1 - Mean_H2| / Mean_All \quad (2)$$

$$Fluctuation_H1 = 100 * Stdev_H1 / Mean_All \quad (3)$$

3.2 Construction of a Fuzzy Inference System Using ANFIS

To build a fuzzy inference system, we adopt ANFIS, which is a kind of neuro-fuzzy system proposed by Jang [7] and considered to be universal approximator. Its inference system corresponds to a set of fuzzy IF–THEN rules that have learning capability to approximate nonlinear functions. In this paper, ANFIS identifies Takagi–Sugeno type fuzzy inference system. ANFIS identifies the parameters of antecedent variables on-line as well as the ones of consequent ones utilizing the error back-propagation method and the least-square method.

Especially ANFIS is very effective when there are many parameters to be identified as the number of fuzzy rules increase or many data to be used in modeling. In this paper, we build and simulate ANFIS using Fuzzy logic toolbox of Matlab.

To have fuzzy system learn by ANFIS and be effective for diverse performance trends and enhance the detectability of the system, we generate various randomized data set having different distribution as seen in Fig. 3. Each data set follows the normal distribution having 100 as the mean value. We generate seven kinds of data set so that its standard deviation would be 3,4,5,6,7,8, and 9 respectively. In Fig. 3, you can see 4 data sets having 3,5,7, and 9 standard deviation respectively.

Based on these seven kinds of data sets, we generate ten variants for each data set, assuming that a big performance regression occurred and the performance would be recovered a while later. The ten variants have different recovery time, that is, 5 to 45 samples as a recovery period. For example, Fig. 1a shows that a performance regression is found at the last sample for the data set having three standard deviation. Figure 1b shows that the performance regression has been resolved after five sample period for the same data set. Figure 1c means that the performance is recovered after 20 sample period.

Additionally we assume three different kinds of anomalies for each data set, that is, the magnitude of the anomaly is 4,5, and 6 standard errors (=4,5,6 * standard deviation). In total, we generate 210 kinds of data sets according to the distribution, recovery period and anomaly size. To train ANFIS, we generated 630 data sets and calculated expected window sizes for each data set.

We defined three fuzzy membership functions for each input variable and Fig. 4 shows the results of learning by ANFIS. In total 27 fuzzy rules have been generated through ANFIS.

The final leaning error is 4.95. The fuzzy rules are the type of Takagi–Sugeno as follows;

Rule i:

If F_all is A_j and DM is B_k and F_H1 is C_l , Then $y_i = p_i * F_all + q_i * DM + r_i * F_H1 + s_i$

Where, $i = 1, \dots, 27$, j, k and $l = 1, \dots, 3$, j, k, l are numbers of membership functions, A_j, B_k, C_l fuzzy membership functions and p_i, q_i, r_i, s_i consequent parameters.

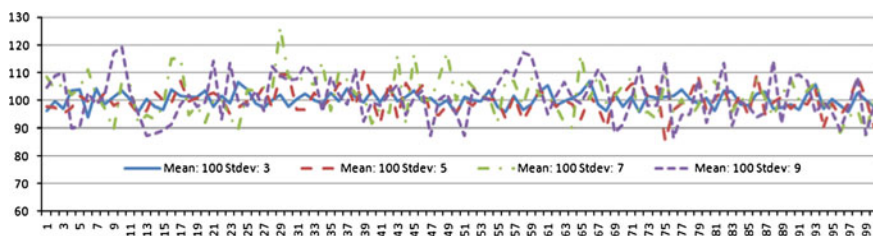


Fig. 3 Sample base data for learning of ANFIS

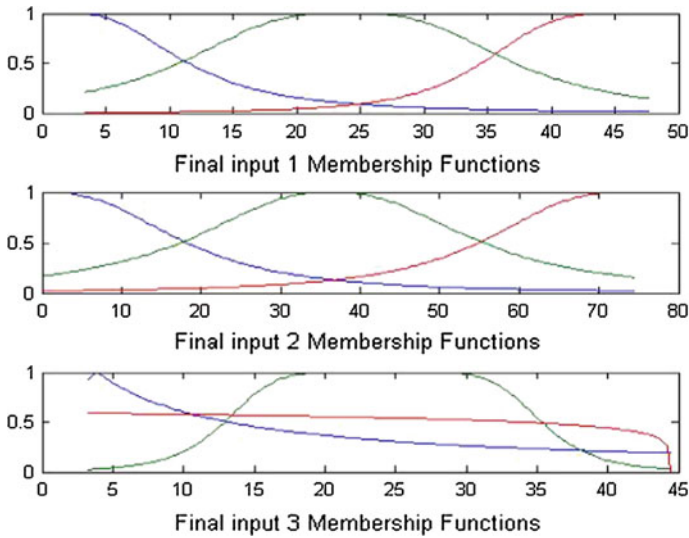


Fig. 4 Fuzzy membership functions of input variables obtained by ANFIS

4 Simulation and Results

Table 1 shows the simulation results by the proposed system using fuzzy logic Toolbox of Matlab. Randomly generated three data sets having 4, 6, and 8 standard deviation are used as verification data. As seen in the table, ANFIS could propose the appropriate window size successfully in 28 cases among 30 showing more than 93 % detectability.

Table 1 Simulation results by ANFIS

Recovery period	Expected window size	Proposed window size by ANFIS		
		Stdev: 4 Anomaly: 5 sigma	Stdev: 6 Anomaly: 5 sigma	Stdev: 8 Anomaly: 5 sigma
0	50 \leq W	54	54	55
5	50 \leq W	48	55	55
10	50 \leq W	58	53	59
15	70 \leq W	84	77	79
20	90 \leq W	92	90	77
25	28 \geq W	19	21	18
30	33 \geq W	21	27	27
35	38 \geq W	32	34	34
40	45 \geq W	40	40	42
45	50 \geq W	44	44	46

5 Conclusions

Recently performance management approaches and analysis methods are discussed as software performance engineering (SPE). This paper is closely related to one of the future goals of SPE mentioned in [8]—“better methods and tools for interpreting the performance results and diagnosing performance problems”.

In this paper, we addressed the special performance trend in which the existing performance anomaly detection system hardly detects the performance change. To handle the issue, we proposed the automatically tuned sampling window size while applying SPC charts. We implemented the optimized Fuzzy detection system built by fuzzy learning mechanism of ANFIS to determine the appropriate sampling window size. We adopt ANFIS as a fuzzy inference system and train input/output membership functions using the randomly generated data sets having diverse fluctuations, anomalies and recovery periods. ANFIS can simulate the system which is represented by the data with small error. Finally we showed that the proposed system can detect the performance changes with more than 90 % of detectability.

References

1. Lee DH, Cha SK, Lee AH (2012) A performance anomaly detection and analysis framework for DBMS development. *IEEE Trans Knowl Data Eng* 24(8):1345–1360. doi: [10.1109/TKDE.2011.88](https://doi.org/10.1109/TKDE.2011.88)
2. Lee DH (2012) Performance anomaly detection and management using statistical process control during software development *J KIISE Softw Appl* 39(8):639–645
3. Montgomery DC (2005) *Introduction to statistical quality control*, 5th Edn. Wiley, New York
4. Komuro M (2006) Experiences of applying SPC techniques to software development processes. In: *ICSE '06: Proceedings of the 28th international conference on Software engineering*, pp 577–584
5. Cangussu JW, DeCarlo RA, Mathur AP (2003) Monitoring the software test process using statistical process control: a logarithmic approach. *ACM SIGSOFT Softw Eng Notes* 28(5):158–167
6. Park J-J, Choi G-S (2001) *Fuzzy control systems*. KyowooSa, Seoul
7. Jang JSR (1993) ANFIS: adaptive network based fuzzy inference system. *IEEE Trans Syst Man Cybern* 23(3):665–685
8. Woodside M, Franks G, Petriu DC (2007) The future of software performance engineering. In: *International conference on software engineering, 2007 Future of software engineering*, pp 171–187. doi: [10.1109/FOSE.2007.32](https://doi.org/10.1109/FOSE.2007.32)

Hybrid Standard Platform for E-Journal Usage Statistics Management

Youngim Jung and Jayhoon Kim

Abstract The increasing availability of usage data for e-journals and the global standard project for usage statistics report has lead increasing interest on collecting and utilizing usage statistic information. However, collecting and integrating large-scale usage reports generated and transferred in a various way by publishers hinders librarians to utilize the usage statistics. Thus we have implemented an automatic collection and visualization system of e-journal usage statistics in this work. By using the suggested system, librarians can focus on more specific and concrete usage data of their institution. In addition, visualized statistics information can be utilized in the composition of various reports on budget or operation of the libraries.

Keywords Usage statistics · Hybrid collection · COUNTER · SUSHI · Scraping · Web · OpenAPI

1 Introduction

Currently, usage statistics of e-journal is necessary for various parties: usage statistics are increasingly performing a role in driving collection development and service decisions for librarians. College administrators need accountability for expenditure and objective basis for flat budget or budget cut. Scholars want to

Y. Jung (✉) · J. Kim

Department of Overseas Information, Korea Institute of Science and Technology Information, 245 Daehangno, Yuseong-gu, Daejeon 305-806, Republic of Korea
e-mail: acorn@kisti.re.kr

J. Kim

e-mail: jay.kim@kisti.re.kr

know how many times the publications have been read and top most read journals among their research interest. Publishers apply usage data to meet the needs of librarians and researchers needs quickly and well, and to improve their publication and marketing process [1]. These growing needs for quantitative analysis in use of electronic journals have led the availability of usage statistics online. Usually, usage statistics of e-journals are generated and provided by each publisher. The usage statistics reports are provided in a various way. Librarians can (1) visit each web site¹ as to view or download the usage statistics report of their own institution, (2) receive the reports regularly via e-mail from publishers, (3) view the reports collected by library automation system or (4) achieve the usage report by file transferring from publisher on request. Web sites where librarians view or download the usage report differ from publisher to publisher, it takes much time and cost for librarians to explore each publisher's website and gather the reports [2].

The aims of this study are (1) to implement a hybrid platform for collecting and managing usage statistics of e-journals, and (2) to provide integrated usage statistics service to over 350 institute including academic, research, public, corporate and medical libraries at nationwide level.

The rest of this paper is as follows: in Sect. 2, current status and problems in collecting and providing usage statistics are described and related work will be discussed. Section 3 illustrates the implementation of the usage statistics of e-journals management system based on web scraping and SUSHI protocol. We summarize analyzed statistical information on usage provided by our visualization and analysis module as well. Section 4 concludes this paper and suggest next step of the study.

2 Related Work

2.1 SOAP and WSDL

Web services have emerged as the core architecture of choice for current distributed web environment. Two important Web services specifications are Web Services Description Language (WSDL) and Simple Object A Protocol (SOAP). WSDL provides a standard language to precisely specify all the information necessary for communication with a Web service, including the interface of the service, its location, and the list of communication protocols it supports. SOAP is the most commonly used Web services communication protocol for information exchange. SOAP specifies how a message-and the data within it- may be

¹ Publisher's web site refers to the web site where the usage reports are provided for librarians. Some publishers may serve the reports in their official web sites; others provide them in separated web sites for providing usage statistics data only.

represented and wrapped in XML. SOAP is a popular choice as the common underlying protocol for interoperability between servers and clients in heterogeneous environment [3].

2.2 Fetching Data by Scraping

Web scraping (also called screen scraping) is a technique of extracting information from websites. Web scraping is closely related to web indexing, which indexes information on the web using a bot and is a universal technique adopted by most search engines. In contrast, web scraping focuses more on the transformation of unstructured data on the web. More specifically, it transforms unstructured web data in HTML format, into structured data that can be stored and analyzed in a central local database or spreadsheet. Uses of web scraping include online price comparison, weather data monitoring, website change detection, research, web mashup and web data integration. Web scraping is generally considered an ad hoc, inelegant technique, often used only as a “last resort” when no other mechanism is available. Aside from the higher programming and processing overhead, output displays intended for human consumption often change structure frequently. Humans can cope with this easily, but computer programs will often crash or produce incorrect results.² Indeed, [4] reported that scraping engines often failed in the trial to collect usage statistics from publisher websites when there is any change in account information for log-in or any modification in URL, directory or structure of the websites. Thus more robust and standardized method for fetching usage data from publishers is required.

2.3 Transferring Usage Statistics Reports via SUSHI

The Standardized Usage Statistics Harvesting Initiative (SUSHI) Protocol standard (ANSI/NISO Z39.93-2007) defines an automated request and response model for the harvesting of electronic resource usage data utilizing the web services framework. It is intended to replace the time-consuming user-mediated collection of usage data reports. The protocol was designed to be both generalized and extensible, meaning it could be used to retrieve a variety of usage reports. An extension designed specifically to work with COUNTER reports [5] is provided with the standard, as these are expected to be the most frequently retrieved usage reports. The standard is built on SOAP for transferring request and response messages. The GetReport method is used for transferring ReportRequest as the input message and returning ReportResponse as the output message as shown in

² http://en.wikipedia.org/wiki/Web_scraping

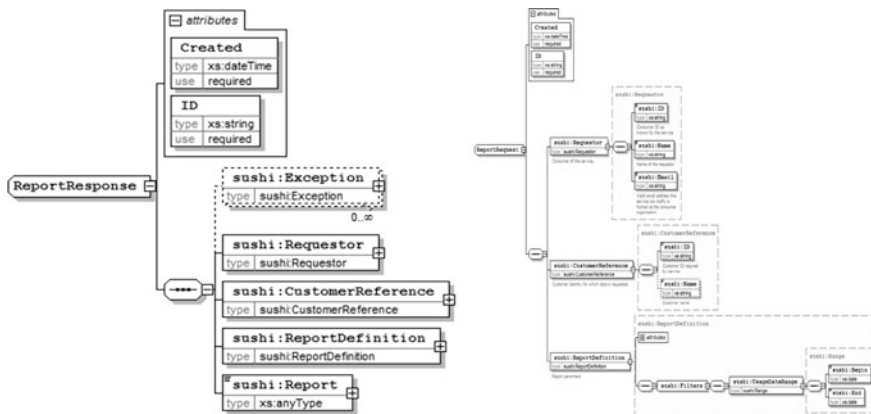


Fig. 1 Request-reply report diagram of SUSHI

Fig. 1. The standard includes a versioned WSDL to describe the Web service namespace and operations, and a generalized XML schema with the syntax of the SUSHI protocol [6].

The diagram of ‘GetReport’ and ‘ReportResponse’ are shown in Fig. 1.

In order to reduce the burden of librarians to visit publisher websites and gather the usage reports, function for collecting usage reports automatically has been implemented in several commercialized library automatic systems. Serials Solution’s 360 COUNTER and ExLibris bX’s UStats use SUSHI protocol. However, only thirty-eight publishers throughout the world adopt SUSHI currently, and major publishers such as Elsevier, Wiley & Blackwell or Springer do not participate in SUSHI project yet.³ In addition, many minor publishers having no sufficient budget for implementing SUSHI server provide usage statistics data through their websites or e-mail only. In this paper, therefore, hybrid method of scraping and standard transferring protocol is studied for usage statistics from publishers including major and minor ones.

3 Hybrid Platform for Managing Usage Statistics

The overall architecture of the suggested system is illustrated in Fig. 2. This system includes a data collection part, a data integration part and a data service part. In Sects. 3.1, 3.2 and 3.3 the three parts will be described in detail, respectively.

³ <http://sites.google.com/site/sushiserverregistry/>

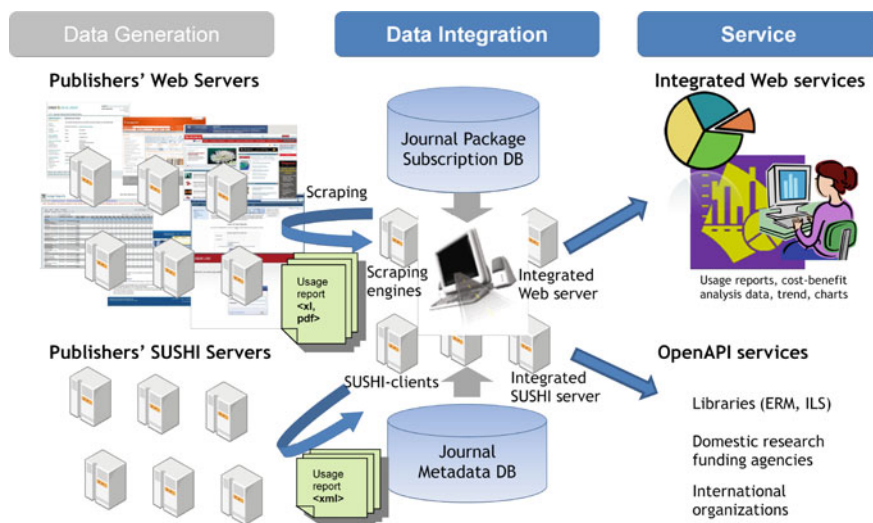


Fig. 2 Overall system architecture

3.1 Hybrid Module for Fetching Usage Statistics Report

Collection of usage statistics reports proceeds in a machine to machine way in the suggested system. As shown in the overall system architecture, two types of engines for fetching usage statistics reports are used, which are scraping engines and SUSHI client modules.

Usage reports are generated on a monthly basis by publishers; however the reports are generated variously and irregularly. A scheduler activates the scraping engines and SUSHI engines to collect usage reports on a scheduled date. If the collection of reports fails on the scheduled date, then the scheduler activates each engine on the alternative date.

Scraping engines log in to publisher's web sites using each institution's account information. Then the engines scrape the usage reports in the specific URL and store the reports in usage statistics DB. Twenty-four scraping engines and eight SUSHI client modules are implemented in this system as shown in Table 1. Both scraping engine and SUSHI-client module have been developed for ACS usage statistics collection [7].

Table 1 Scraping engine and SUSHI-client module for fetching statistics

Collection method	Number of publisher	Publisher name
Scraping	24	ACS, AIP, APS, ASCE, ASME, Berkely, BMJ, Brill, Elsevier, IEEE, IOP, JSTOR, Karger, NPG, OSA, OUP, Pion, NAS, Sage, Science, SPIE, Spinger, Thieme, Wiley
SUSHI	8	ACS, AR, BioOne, CUP, MAL, ProjectMUSE, RSC, WG

3.2 *Integration of Statistics with Journal Subscription Data and Metadata*

COUNTER JR1 compliant reports collected from publishers only provide ‘number of successful full-text article requests by month and journal’ year by year. These simple data cannot give sufficient insights or implications for analyzing cost versus use, e-journal use trend or various summary data for users. For the multi-layered and comprehensive analysis of use, consortium subscription databases are linked in this system. Table 2 presents various data provided in the suggested system.

In addition, integrated journal usage reports with e-journal metadata enables for researchers to make further analysis on journal use distribution by subject, comparative analysis of usage versus citation of e-journal, and others.

3.3 *Hybrid Platform for Web Service and OpenAPI Service*

The suggested system provides a two-way service for human beings and machines: which are web service and openAPI service, respectively.

Users log in the web system, and then can view the multi-layered statistical data as shown in Table 2 as well as the original journal usage reports collected from publishers. In addition, all kinds of report can be downloaded in excel files and various graphs, charts and tables are downloaded in jpg or pdf format from the web service as shown in Fig. 3. By utilizing the downloaded reports, graphs, charts and tables, users can compose various reports on the use, cost and purchase of e-journals conveniently.

Table 2 Multi-layered Statistical Data

Category	Information title
Journal level	Original JR1 report
	Bibliographic data of journal (metadata, subject, Impact Factor)
	Breakdown of publisher usage (title and year)
	Breakdown of institutional usage (title and year)
	Which titles have highest use (institution, year and package)
Package level	Breakdown of package usage (title and year)
Summary report	Summary of publisher usage
	Summary of consortial package usage
	Summary of institutional usage
	Summary of overall usage for the last 10 years
	Summary of overall number of reports collected for the last 10 years
	Use by institution types
Cost report	Journal package cost
	Cost per articles accessed (CPA)

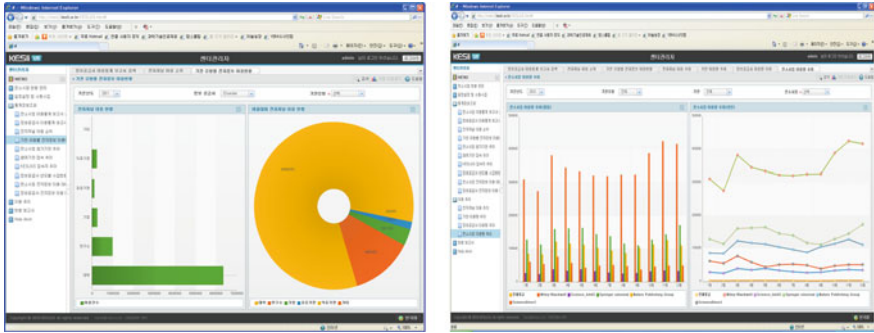


Fig. 3 Statistical data visualized in graph, chart and pie in web service

OpenAPI service is designed for the machine-to-machine transferring of bulk statistical data in way. An individual institution who has purchased journal (package)s may want to provide its journal usage reports in its own electronic resource management (ERM) system or integrated library system. The suggested system can function as a SUSHI server of an integrated usage statistics registry in South Korea for providing integrated journal usage reports according to SUSHI standard protocol. Recently, journal usage data are regarded as complementary measures to the traditional citation-based evaluation of journal impact [8]. Thus a research funding agency needs to consider not only citation count but also various metrics of journal impact including usage. An international organization related journal usage data such as ProjectCOUNTER may require a regional usage statistics of journals, as well. For sharing the impact of a specific journal throughout South Korea, OpenAPI service is provided for those bodies to search the entire usage of a journal with its title, publisher name and ISSNs.

3.4 Implementation Result

Table 3 presents the usage statistics reports collected by the suggested system. Statistics data since 2000 are constructed for 359 libraries. Figure 4 illustrates the performance of scraping (blue bar) and SUSHI (red bar). Scraping engines run much slower than SUSHI due to the large scale data of major publishers and their technical limitation.

Table 3 Entire Usage Statistics Reports Collected from 31 Publishers

Publisher	Years											
	2012	2010	2009	2008	2007	2006	2005	2004	2003	2002	2001	2000
AAAS	9	95	95	95	95	95	95	95	95	95	95	95
ACS	5,200	5,200	5,040	5,040	5,040	0	0	0	0	0	0	0
AIP	1,863	2,423	2,101	1,910	1,970	2,398	387	0	0	0	0	0
APS	684	692	700	677	676	0	0	0	0	0	0	0
AR	41	1,148	1,148	1,148	1,148	1,148	1,148	1,148	1,148	0	0	0
ASCE	402	2,686	2,867	2,509	1,246	1,145	0	0	0	0	0	0
ASME	529	597	650	432	350	240	220	0	0	0	0	0
Berkely	456	570	570	570	570	570	570	570	570	570	570	570
BioOne	1,512	1,512	1,509	1,508	0	0	0	0	0	0	0	0
BMJ	855	1,035	1,059	1,059	1,037	1,039	1,059	1,059	1,059	0	0	0
Brill	0	1,514	1,416	0	0	0	0	0	0	0	0	0
CUP	17,618	17,980	19,234	13,426	12,732	9,163	13,061	13,410	17,752	0	0	0
Elsevier	0	562,889	589,297	583,500	578,138	0	0	0	0	0	0	0
IEEE	0	32,018	31,125	26,936	18,358	0	0	0	0	0	0	0
IOP	6,649	7,957	7,957	7,661	0	0	0	0	0	0	0	0
Wiley	395,556	395,370	383,420	365,190	312,582	83,520	71,544	57,715	46,864	0	0	0
JSTOR	0	246	25,728	26,349	23,478	22,858	24,858	25,144	0	0	0	0
Karger	950	2,172	2,162	1,034	0	0	0	0	0	0	0	0
MAL,	840	1,086	1,072	1,072	1,072	0	0	0	0	0	0	0
NAS	34	34	34	34	34	34	34	34	34	34	2	2
NPG	0	1,452	1,401	1,795	1,621	718	0	0	0	0	0	0
OSA	483	483	437	0	0	0	0	0	0	0	0	0
OUP	0	17,141	16,607	16,637	4,878	4,711	3,365	2,680	0	0	0	0
Pion	0	34	34	0	0	0	0	0	0	0	0	0
PM	502	496	490	490	0	0	0	0	0	0	0	0

(continued)

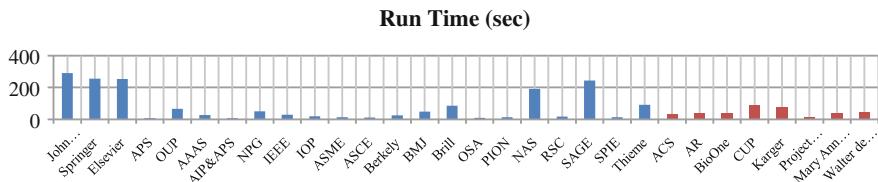


Fig. 4 System performance of scraping and SUSHI

4 Conclusion and Further Work

In order to collect and construct large-scale usage statistics data from 31 publishers and to manage visualized and analyzed statistical information, a hybrid standard platform has been implemented in this work. Whereas SUSHI-client module operate fast for a limited number of publishers data, scraping engines cover most and large scale usage statistics data in a slower tempo. At the moment both methods have their strength and weakness. Multi-layered and comprehensive statistical information presented in graphs, charts and tables provided in this platform will be utilized for librarians to write reports and make an important decision on purchasing and canceling e-resources. For generating in-depth analysis statistical data, automatic method of integrating usage data and metadata of e-journals will be studied as our next step. Recognition and identification of institution, publisher and journal needs to be studied since names were/are/will be changing over time. For the consistent usage statistics service over time, we need to keep tracking down the list of names and the changes and investigate systematic recognition and identification method. In addition, we plan to design and implement a subscription decision support system by utilizing data mining technique for selecting best journals (packages) to purchase in considering the use, cost and impact of e-journals.

References

1. Borghuis M (ed) (2005) What counts and what doesn't: an insider's guide to usage reports. Library connect, vol 7. Elsevier, San Diego
2. Jung Y, Kim J, You B (2011) Development of automatic collecting system of e-journal usage statistics based on screen scraping. In: 2011 Korea computer congress, Kyeongju
3. Head MR, Govindaraju M, Slominski A, Liu P, Abu-Ghazaleh N, Engelen R, Chiu K, Lewis MJ (2005) A benchmark suite for SOAP-based communication in grid web services. In: 2005 Super computing: international conference for high performance computing, networking, and storage
4. Jung Y, Kim J (2012) Improving efficiency of usage statistics collection and analysis in e-journal consortia. J Korean Soc Inf Manag 29(2):7-25
5. Project COUNTER, Release 3 of the COUNTER: code of practice for journal and databases http://www.projectcounter.org/code_practice.html

6. NISO, NISO Standardized Usage Statistics Harvesting Initiative (SUSHI): Z39.93. http://www.niso.org/apps/group_public/download.php
7. Jung Y, Kim J (2012) Analysis of STM e-journal use in south korea by utilizing automatic collection system of e-journal usage statistics. In: 8th international conference on WIS and 13th COLLNET meeting, Seoul
8. Bollen J, Van de Sompel H, Rodriguez MA (2008) Towards usage-based impact metrics: first results from the MESUR project. In: JCDL 2008, Pittsburgh

Study for Predict of the Future Software Failure Time Using Nonlinear Regression

Yoon-Soo Ra and Hee-Cheul Kim

Abstract Software failure time have been proposed in the literature exhibit either constant, monotonic increasing or monotonic decreasing. For data analysis of software reliability model trend analysis was developed. The methods of trend analysis are arithmetic mean test and Laplace trend test. Trend analysis only offers information of outline content. In this paper, we discuss failure time case of failure time censoring, and predict the future failure time using nonlinear regression models (growth, Logistic and weighted type) which error terms for each other are different. The proposed prediction method used the failure time for the prediction using nonlinear regression model. Model selection, using the coefficient of determination and the mean square error, were presented for effective comparison.

Keywords Software reliability · Time censoring · Nonlinear regression

1 Introduction

Computer system failure due to software failures can cause tremendous loss to our society. Therefore, software reliability is an important issue in the software development process. This problem should satisfy the user's requirements and

Y.-S. Ra (✉)

Department of Sport and Leisure Studies, Kwandong University,
Gangneung-si, Gangwon-do, South Korea
e-mail: ysra0820@hanmail.net

H.-C. Kim (✉)

Department of Industrial and Management Engineering, Namseoul University, 21 Maeju-ri,
Seonghwan-eup, Cheonan-si, Chungnam 330-707, South Korea
e-mail: kim1458@nsu.ac.kr

testing costs. Efficiently in order to reduce the costs of variability and reliability of the software must know in advance the cost of testing in software testing (debugging). Therefore, the reliability, the software development process with considerations of cost and emission time, is essential. There is a need to develop a model to estimate the defect content of software products for the end. Until now, many software reliability models have been proposed.

Some of these models for describing the software failure phenomenon are based on the non-homogeneous Poisson process (NHPP). In fact, these models are fairly effective in describing the error-detection process with a time-dependent error-detection rate. However, the assumption is that each time an error occurs the fault that caused it can be immediately removed, leading to no new problems, which is usually called perfect debugging [1]. More recently, Huang [2] incorporated both a generalized logistic testing-effort function and the change-point parameter into software reliability modeling. In attempting to predict software reliability, computer algorithm has been an effective technique in assisting the prediction.

Chiu, Huang and Lee [3] can explain the learning process that software managers to become familiar with the software and test tools for S-type model. In this paper, we measured time to failure, then when the time cutting time prediction of future failure was studied. For predict the future failure time [4], the weighted model, growth model and logistic-type curve regression analysis was used.

2 Related Research

If the arithmetic mean increase pattern, results of the arithmetic mean test represents reliability growth. The Laplace trend test results from the Laplace factor value has varying between -2 and 2 indicate stable reliability [5]. In addition, time series model of the software using the AR (1) model to predict future failure time was studied [6]. In the relatively recent, time-series analysis using a simple moving average, weighted moving average and exponential smoothing to predict the future failure time have been proposed [7]. Also, the S-curve regression model, growth model were compared, predict result of the future failure, logistic model shows better model [8].

2.1 Nonlinear Regression Model

If possible, a linear relationship between the explanatory variable x and dependent variable y , the linear model represents the relationship as follows:

$$y_i = b_0 + b_1x_i + \varepsilon_i, \quad (i = 1, 2, \dots, n) \quad (1)$$

where, b_0 and b_1 are regression coefficients and $\varepsilon_i \sim N(0, \sigma^2)$.

However, the non-linear case (curve linear model) by taking the log or the weight converted the linear model.

2.1.1 Curve Linear Regression Model

The growth and Logistic model regression analysis used, in this paper, to predict the future failure time. The growth model and logistic model are known, respectively, as follows [4]:

$$\hat{h}(y_i) = \beta_0 + (\beta_1 + x_i), \quad (i = 1, 2, \dots, n) \tag{2}$$

$$\hat{h}(1/y_i - 1/\mu) = \ln(\beta_0) + (\ln(\beta_1) \times x_i), \quad (i = 1, 2, \dots, n) \tag{3}$$

where β_0 and β_1 are estimates of b_0 and b_1 , μ is upper boundary value.

2.1.2 Weighted Regression Model

Case of unequal dispersion or variance for the error term, the weighted regression can be used. Namely, in Eq. (1)

$$Var(\varepsilon_i) = \sigma_i^2 = w_i \sigma^2 \tag{4}$$

where, w_i is weight.

In this case, the regression line using the weighted least squares method to predict which are called weighted regression analysis of these series. Of these cases, the regression coefficients are known as follows [9]:

$$\beta_1 = \frac{\sum_{i=1}^n w_i(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n w_i(x_i - \bar{x})^2}, \quad \beta_0 = \bar{y} - \beta_1\bar{x} \tag{5}$$

where, β_0 and β_1 are estimates of b_0 and b_1 ,

$$\bar{x} = \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}, \quad \bar{y} = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} \tag{6}$$

In this paper, we want to use weights $\frac{1}{x_i} \left(Var(\varepsilon_i) = \sigma_i^2 = \frac{\sigma^2}{x_i} \right)$

2.2 Criteria to Determine the Optimal Model

How the predicted value and the actual value of the material to determine whether the error measure is the criterion for better prediction techniques [3, 8]. In general, the coefficient of determination, modify the coefficient of determination, the sum mean squared criterion is possible.

2.2.1 Coefficient of Determination (R^2)

The coefficient of determination is defined as the sum of the square by the regression line (SSR), the sum of squared errors (SSE), and called the sum total squared (SST). R^2 can measure how successful the fit is in explaining the variation of the data. It is defined as follows:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \tag{7}$$

where $SST = \sum_{i=1}^n (y_i - \bar{y})^2$, $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, $= \sum_{i=1}^n (y_i - \hat{y}_i)^2$, and \hat{y} is estimates of y .

2.2.2 Mean Square Error

As the basis for determining the efficient model, the Mean Square Error (MSE) is defined as follows:

$$MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} \tag{8}$$

Because difference the predicted and the actual value, the minimum of MSE model is efficient model.

3 The Proposed Prediction Method

The failures time data [10] summarized in Table 1 for the prediction. For his data, the first, trend data should be preceded [11]. Laplace trend test analysis typically performed. Showing in Fig. 1, the Laplace factor between -2 and 2 appears confidence growing by reliability growth properties. Therefore, using this data, it is possible to estimate the time to failure time. In Table 2, coefficient determination and mean square error is summarized. This table, because weighted regression model than any models value of coefficient of determination is high and mean square error is small, weighted regression efficient model. Summarized in Fig. 2 from the point 1(failure number) to the point 30 (failure number), prediction using weighted model, the growth model and the logistic model are drawing. All models is similar to the true value in this figure, the weighting model shows relatively better prediction and growth model shows lower (underestimate), the logistic model appears higher (overestimate).

In Fig. 3 and Table 3 were summarized, failure time of future were predicted from the point 31(failure number) to the point 40 (failure number) and statistic information using weighted model, the growth model and the logistic model. In this table, because estimate of the growth model appears from 9.742 to 12.478,

Table 1 Failure interval time and failure time

Failure number	Failure interval (second)	Failure time (second)	Failure number	Failure interval (second)	Failure time (second)
1	0.479	0.479	16	1.908	10.771
2	0.266	0.745	17	0.135	10.906
3	0.277	1.022	18	0.277	11.183
4	0.554	1.576	19	0.596	11.779
5	1.034	2.610	20	0.757	12.536
6	0.949	3.559	21	0.437	12.973
7	0.693	4.252	22	2.230	15.203
8	0.597	4.849	23	0.437	15.640
9	0.117	4.966	24	0.340	15.980
10	0.170	5.136	25	0.405	16.385
11	0.117	5.253	26	0.575	16.96
12	1.274	6.527	27	0.277	17.237
13	0.469	6.996	28	0.363	17.600
14	1.174	8.170	29	0.522	18.122
15	0.693	8.863	30	0.613	18.735

Fig. 1 Laplace trend test

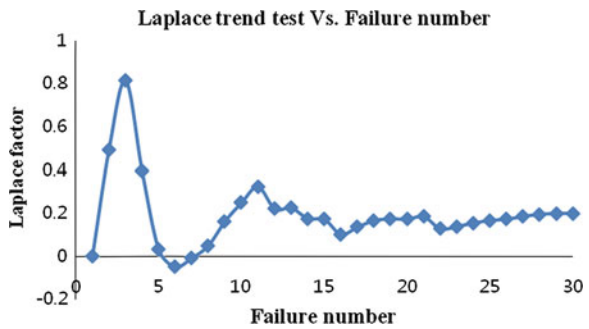


Table 2 Model summary and parameter estimation of each model

Model	Model summary		Parameter estimation	
	R^2	MSE	β_0	β_1
Weight $\left(\frac{1}{x_i}\right)$	0.879	1.598	0.554	0.653
Growth	0.840	31.205	0.318	0.304
Logistic	0.842	30.464	0.729	0.901

Explanatory notes

Explanatory variable: failure number, Dependent variable: failure time, R^2 : Coefficient of determination, MSE : Mean square error
 β_0 and β_1 : estimates of b_0 and b_1

Fig. 2 Predictive failure time for each model from one point time to 30 point time

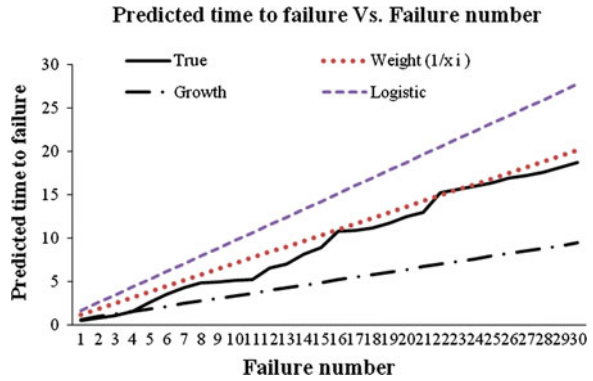


Fig. 3 Predictive failure time for each model from 31 point time to 40 point time

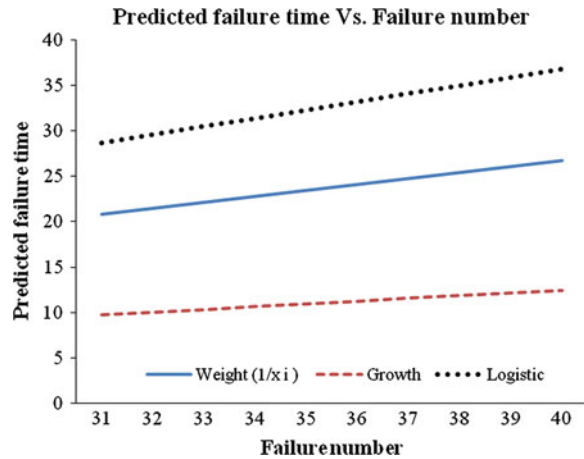


Table 3 Statistic information about predictive failure time data for each model

Statistics	Weight ($(\frac{1}{x_i})$)	Growth	Logistic
Mean	23.7355	11.12	32.7145
Standard error	0.625199901	0.291057841	0.862641824
Median	23.7355	11.11	32.7145
Standard deviation	1.977055681	0.920405708	2.727912969
Variance	3.908749167	0.847146667	7.441509167
Range	5.877	2.736	8.109
Minimum value	20.797	9.742	28.66
Maximum value	26.674	12.478	36.769
Observation number	10	10	10

the growth model than any models shows better range. In addition, for the average, weighted model is the highest, logistic model is lowest. Because logistic model is the highest for the dispersion (standard deviation and variance) and growth model is relatively small. If want to use logistic model, the security about dispersion measures is necessary.

4 Conclusions

As software systems play an increasingly important role in computer systems, intensive studies have been carried out to ensure the software reliability.

Evaluation can be modeled as software reliability growth test time or run-time failure the time, more realistic and relatively efficient model in this area is unknown, but the choice of weights can be a major issue.

Therefore, fluctuations in the quality of the products that are produced in the manufacturing process which, despite more than continue to run without any action or improvement process, so the quality of the product will fall dramatically. This paper, censoring software systems design and related industries, the workers anticipate and consider the weights of the information to pre-censoring data can be analyzed, using weighted prediction method utilizing the software quality could be helpful to the improvement of software reliability.

Acknowledgments Funding for this paper was provided by Namseoul University.

References

1. Gokhale SS, Trivedi KS (1999) A time/structure based software reliability model. *Annal Softw Eng* 8:85–121
2. Huang C-Y (2005) Performance analysis of software reliability growth models with testing-effort and change-point. *J Syst Softw* 76:181–194
3. Kuei-Chen C, Yeu-Shiang H, Tzai-Zang L (2008) A study of software reliability growth from the perspective of learning effects. *Reliab Eng Syst Saf* 93:1410–1421
4. Kim H-C, Shin H-C (2012) The study for software future forecasting failure time using curve regression analysis. *Korea Convergence Secur Assoc* 12(3):115–121
5. Kim H-C, Park H-K (2009) Exponentiated exponential software reliability growth model. *Int J Adv Comput Technol* 1(2):57–64
6. Kim H-C, Shin H-C (2008) The study for software future forecasting failure time using ARIMA AR(1). *Korea Inf Assur Assoc* 8(2):36–40
7. Kim H-C, Shin H-C (2008) The study for comparative analysis software future time using EWMA control chart. *Korea Inf Assur Assoc* 8(3):33–39
8. Kim H-C, Shin H-C (2011) The study for software future forecasting failure time series analysis. *Korea Inf Assur Assoc* 11(3):19–24

9. Kim H-C (2010) Introduction to regression analysis. Biz-Press, pp 131–137
10. Hayakawa Y, Telfar G (2000) Mixed poisson-type processes with application in software reliability. *Math Comput Model* 31:151–156
11. Kanoun K, Laprie JC (1996) Handbook of software reliability engineering. In: Lyu MR (ed) Chapter trend analysis. McGraw-Hill, New York, pp 401–437

Analysis of Threat-Factors for Biometric-Information Processing Systems According to Goal-Oriented Threat- Modeling

Su-Jin Baek, Jong-Won Ko and Jung-Soo Han

Abstract As there is an increasing reliance on information systems in most organizations, there is also an increased risk of security accidents of information systems. Therefore, in order to verify the potential security risks and their results, it is necessary to have a security threat assessment process called threat modeling. However, research in security threat modeling has yet to mature as there is paucity of established techniques and tools to aid the threat modeling and formal analysis process. This study provides a method to design and analyze threats that occur in the bio-information processing system using the visual Goal-oriented threat modeling. In addition, it determines each threat based on the Goal-Threat model and defends itself through measuring repetitive assessment, vulnerability the degree of risk. Then, by designing each organization to perform security checks on its own, it supports to make it possible to easily find vulnerabilities in terms of administration and presents a recommendation to be performed in order to ensure stability.

Keywords Threat modeling · Biometric information processing system · Security threat

S.-J. Baek · J.-W. Ko (✉)

Department of Computer Engineering, KyungHee University, Giheung-gu Gyeonggi-do, Yongin-si 446-701, Korea
e-mail: jwko@khu.ac.kr

S.-J. Baek

e-mail: croso@khu.ac.kr

J.-S. Han

Division of Information and Communication, Baekseok University, Cheonan Chungcheongnam-do, Seoul 330-704, South Korea
e-mail: jshan@bu.ac.kr

1 Introduction

The fast changing information society has a dysfunction of diverse and serious security. In recent days, security threats are gradually becoming specialized, advanced and complicated. Thus, security accidents are also becoming accidents related to specific information (personal information, confidential business information, etc.) from accidents related to the IT infrastructure. In response to changes in the threat of the information society, the measures to respond the threats are also changing. With the advent of information risk society and increased demand for institutional/legal measures in recent years, there has been an increasing need for more systematic and comprehensive measures to address the risks. In order to manage these risks, the measures to protect personal information using the bio-information that have unique features have to be considered instead of infrastructure and technology based approaches [1].

Bio-information has the characteristics of universality, uniqueness and permanence. When viewed from the perspective of authentication, it can act on the positive features. However, it can cause the problems of individual secret information and groups' confidential information to be leaked, which are stored using bio information when bio-information gets leaked by an attacker [2, 3]. Therefore, it is important to identify potential security risks and security risks using bio-information. And it is necessary to have security threat assessment process called threat modeling to determine the level of measures.

Security threat modeling (or simply threat modeling) is a formal process of identifying, documenting and mitigating security threats to a software system. It enables development teams to understand a system's threat profile by examining the application through the eyes of a potential adversary, and helps to determine the highest-level security risks posed to the system [4]. The threat modeling process usually involves identifying information resources to be protected, identifying the entry or access points to these assets, analyzing the threats, evaluating associated risks, and developing mitigating strategies. Ideally, a threat model should be developed during the earliest stages of system development, and then as the application evolves and requirements are better defined, the threat model can be updated as needed [5]. However, research in security threat modeling has yet to mature as there is paucity of established techniques and tools to aid the threat modeling and formal analysis process. While the importance of starting threat modeling during the requirements analysis phase has been well discussed in the literatures, existing modeling notations such as data-flow diagram and attack trees [6], are largely focused on the design and development phases. Moreover, existing work do not integrate threat modeling notations with a formal threat analysis procedure to aid decision making during security requirements analysis.

This paper designs and analyzes risks that occur in the bio-information processing system with the Goal-oriented modeling. In addition, it determines each threat based on the Goal-Threat model and defends itself through measuring repetitive assessment, vulnerability the degree of risk. Then, by designing each

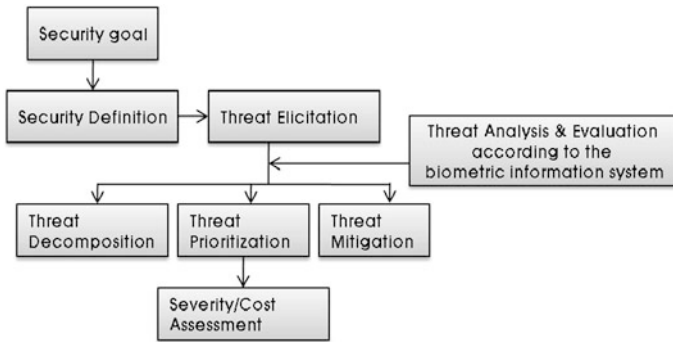


Fig. 1 Goal-threat modeling and analysis about requirement

organization to perform security checks on its own, it supports to make it possible to easily find vulnerabilities in terms of administration and presents a recommendation to be performed in order to ensure stability.

2 Goal-Oriented Threat Modeling for Bio-Information Processing System

The Threat-factors of the bio-information processing system are broadly divided into security threats in each system of the bio system and security threats in the process. In order to understand and analyze these security threats, it measures repetitive assessments, vulnerability assessment and the level of the risk using the Goal-oriented threat modeling as in Fig. 1 to defend adequately.

Also, proposed modeling and analysis process for security threats is depicted in Fig. 1. The process includes five-level steps which encompass setting goal about security requirement, defining what security means for the system, eliciting threats, analyzing threats and their associated risks, and evaluating how countermeasures lead to the achievements of security objectives. The entire process is documented in a threat analysis that forms the threat model to be used throughout the development life cycle.

3 Effect Relationship of Goal-Oriented Threat Model and Bio-Information Processing System

As far as the software is concerned, the requirement plays the roles of the necessary inputs for the software to achieve the intended goals. Therefore, the input requirements are to be extracted by the assets and then analyzed based on the

scenarios so that the goals to be changed can be identified. The requirement goals can be achieved by putting together at least several of the processes. That is, the scenarios to fulfill the required goals should be defined, followed by the extraction of the abstract goal requirements. Here, the information regarding the threat requirements from the security is displayed as the contextual information, which is assumed to be input in the forms of requirements in natural language. Then, the groups that are composed of the assets required for the goals by the security requirement are defined, with which sub-goals are determined by taking various steps as shown in Fig. 2 below.

The goal-threat models for the security requirement are then modeled into goal graphs, and the requirements are analyzed based on the goal and the scenarios about bio-Information Processing system. In this method, the detailed flows for the achievement of the higher-level goals are disclosed by the scenarios and then the low-level goals are identified based on the threat-factor and the flow of the scenarios. In order for the fulfillment of the system's goals, a multiple number of the threat-factor are carried out. Also, each of the threat-factors can be related to more than one

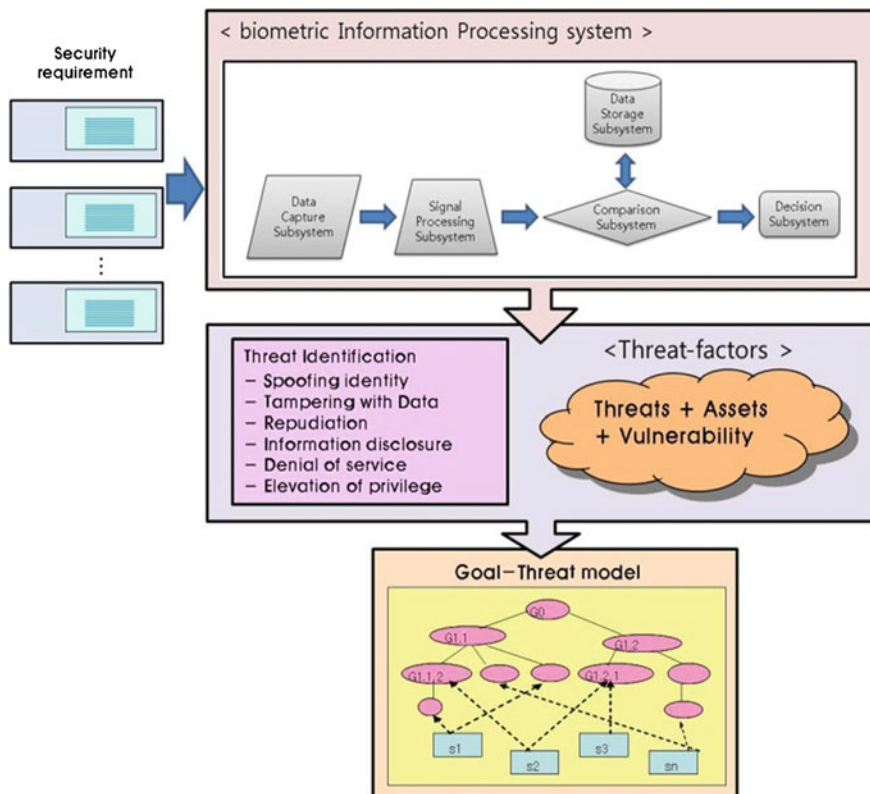


Fig. 2 The influential relationships between the biometric information system and the security requirement to extract the goal-threat model

of the goals. If any of the threat-factors are considered not important for the fulfillment of the goals, it will be set as non-relevant to the goal. With this, it becomes possible not only to identify the threat-factors that are running currently, but also to anticipate one or more actions that could possibly be executed next time.

A Bio-Information processing system is composed of the followings as seen in Fig. 2. It has bio-capture system data to recognize the bio information, signal processing system to process bio information, data storage system to store bio reference and recognition reference and the comparison system to calculate the values when the score, which is based on the measurement of the bio information stored by the data storage system and the bio information processed by the signal processing system, is more or less than the specified value. It also has a decision system to determine the personal identification through determining the processed result [7]. The threats that occur in these systems are extracted through the threat, assets and vulnerabilities of the security objectives. The threats did not occur yet; refer to information about the behavior and intentions with the potential losses. The assets refer to the value that the information system possesses. And the vulnerability refers to the information about the flaw or weakness that can damage the information system. So, it is to generate the security Goal-threat model through by the information about these threats and STRIDE. Therefore, the security goal is defined in six different abstraction Categories, which are respectively Spoofing identity(S), Tampering with Data(T), Repudiation(R), Information Disclosure(I) Denial of Service(D), and Elevation of Privilege(E).Threat classification schemes such as STRIDE can be used to elicit security threats for a biometrics Information Processing system. The system context and environment can be cross-examined against these classes of threats to determine if they are relevant to the system being developed.

4 Analysis of the Threat-Factors By Goal-Oriented Threat Modeling

The easiest way to apply the Goal-oriented threat modeling to the bio-information processing system is to consider what impacts it will give to each component and the relationship and connection with other system components. Therefore, it is imperative to observe each part of the bio-information processing system and to verify whether there are any of the relevant threats in accordance with security objectives within the threat range of STRIDE. And it is imperative to perform the following process repetitively for the sake of appropriate defense in accordance with Goal threatening model.

1. Confirm the impacts on each system as displaying detailed Goal threats found through Goal-Threat model.
2. Grade each Goal-Threat model Goal threat based on criticality, impact and likelihood. Assign a number on a scale of 1 to 10 as 10 to be the most serious

level of importance. As for likelihood, set 1 for the item most likely to occur and 10 for the item least likely to occur. Dividing criticality by likelihood will give overall risk. In other words, the overall risk is calculated with $\text{Risk} = \text{Criticality}/\text{Likelihood}$.

3. Mitigate each threat by selecting the appropriate techniques or technologies for defense against security threats associated with the construction and processing of bio-information processing system.
4. Since it is hard for a system to determine all threats at an initial phase, it may cause a vulnerability of application even as processing through one item. Thus, it is important to start from the first step once an analysis of the Goal-Threat model is under way. And it should be done from the upper level to the lower level.

According to the process described above, each threat Goal-Threat model is analyzed and evaluated. And depending on the severity of the threat, it will be made to defend against security threats.

5 Conclusions

Therefore, it is important to identify potential security risks and security risks using bio information. And it is necessary to have security threat assessment process called threat modeling to determine the level of measures. Threat modeling plays a significant role in the design of the overall security model for a system because it can help to ensure that security is built into applications, rather than addressed as an afterthought. This paper provides a method to design and analyze threats that occur in the biometric information processing system using the visual Goal-oriented threat modeling.

Further, this study identified each threat based on the Goal-threat model of the Goal and performed the repetitive assessments through STRIDE threat to find out what impacts the threats would give to each component and the relationship and connection with other system components. Upon analyzing and evaluating each threat of the Goal-Threat model in accordance with the proposed process, it will be made to defend against threats based on criticality. Then, by designing each organization of the bio information processing system to perform security checks on its own, it will be made to support it to easily find vulnerabilities in terms of management and to present a method to ensure such protection in order to ensure stability. According to the process described above, each Goal-Threat model is analyzed and evaluated. And depending on the severity of the threat, it will be made to defend against security threats.

Acknowledgments This work was supported by the Security Engineering Research Center, granted by the Korea Ministry of Knowledge Economy.

References

1. Arun AR, Nandakumar K, Anil KJ (2006) Handbook of multibiometrics. Springer, New York
2. Shin YN, Kwon MJ, Lee YJ, Park JI, Chun MG (2009) Biometric and Identity reference protection. *J Korean Inst Intell Syst* 19(2):160–167
3. Shin YN, Kim YJ, Chun MG (2011) Operational management for biometrics hardware security module and PKI. *J Korean Inst Inf Technol* 9(5):207–216
4. Swiderski F, Snyder W (2004) Threat modeling. Microsoft Press, Redmond
5. Park KY, Yoo SG, Kim J (2011) Security requirements prioritization based on threat modeling and valuation graph. *Commun Comput Inf Sci* 206:142–152
6. Baek SJ, Han JS, Song YJ (2012) Security threat modeling and requirement analysis method based on goal-scenario, IT convergence and security. In: Proceedings of the international conference on IT convergence and security 2011/2012, pp 419–424
7. ISO/IEC JTC1 SC27 N8802 (2010) Biometric information protection. Final Committee Draft, 2010

Distinct Element Method Analysis of Retaining Wall Using Steel Frame and Fill Material

Sam Dong Jung, Jung Won Park, Jong Hwa Won, Jeong Soo Kim and Moon Kyum Kim

Abstract Retaining wall using steel frame and fill material is a kind of the cellular structure. The cellular structure using a fill material has a number of advantages such as excellent constructability in the short term, permeability, and decrease in effects of groundwater fluctuations after the end of construction. Because this structure is discontinuity, it is quite difficult to apply analysis in finite element methods. In addition, there are no evaluation techniques to establish its shear resistance clearly. In order to solve the issue, this paper presents that shear resistance analysis is performed by introducing displacement incremental analysis into the distinct element method. It enables not only to model continuous and discontinuous structure, but also to perform static and dynamic analysis. The results of analysis are compared with experimental results of the retaining wall with a steel frame and fill materials.

Keywords Cellular structure · Shear resistance · Fill material · Distinct element method

S. D. Jung (✉) · J. W. Park · J. H. Won · J. S. Kim · M. K. Kim
Department of Civil and Environmental Engineering, Yonsei University, Seoul,
South Korea
e-mail: sdjung@kictep.re.kr

J. W. Park
e-mail: wildflower@kictep.re.kr

J. H. Won
e-mail: exameve@yonsei.ac.kr

J. S. Kim
e-mail: coffee1210@yonsei.ac.kr

M. K. Kim
e-mail: applymkk@yonsei.ac.kr

1 Introduction

The cellular structure is composed of the cell body and fill materials. The cellular structure using a fill material has a number of advantages such as excellent constructability in the short-term, permeability, and decrease in effects of groundwater fluctuations after the end of construction. This structure is also economical because its fill material can be obtained around the construction site. For these reason, cellular structures are used widely, as the breakwater in particular.

As the load increases, the shear resistance occurs between fill materials in the cellular structure. Then, the shear resistance of fill materials increases the stiffness of the cellular structure. It is important to estimate discontinuous interaction of exterior steel frame between interior fill materials in a cellular structure. However, it is quite difficult to apply finite element methods to a cellular structure due to the fact that it is a typically discontinuous structure, and there are no clearly established evaluation techniques regarding its shear resistance. In recent years, as developing Distinct Element Method (DEM), discontinuous structures can be modeled and performed its dynamic analysis. This paper presents shear resistance analysis of cellular structure is performed by introducing displacement incremental analysis into the DEM. The results of analysis are compared with experimental results of the retaining wall with a steel frame and fill materials.

2 Mechanical Behavior of Cellular Structure

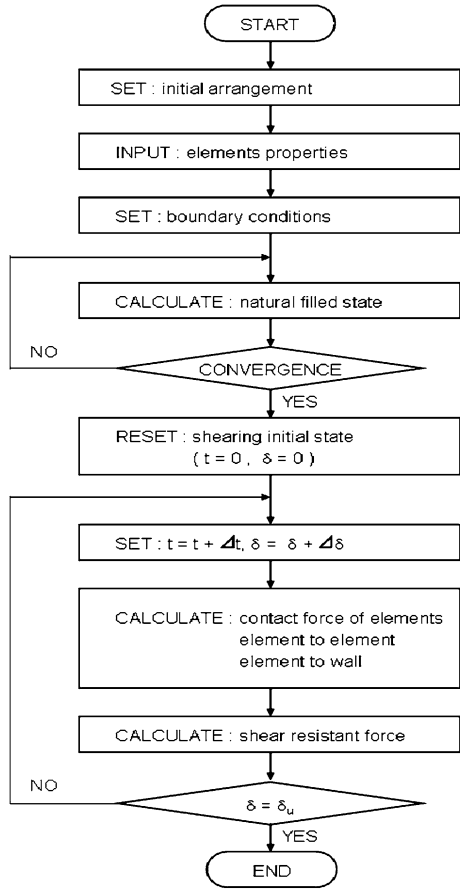
The theory of cellular structure's behavior was first proposed by Terzaghi. If the horizontal external force acts on the the cellular structure, the shear failure occurs along the vertical plane of the fill material of cellular structure. As the result, the cellular structure gradually changes from a rectangular shape to a parallelogram shape.

3 Structure Modeling By the Distinct Element Method

DEM is the numerical technique using the orientation and gap of the discontinuities of the rock mass separated by elements. Especially, DEM and Finite Element Method (FEM) differ in considering the interaction of adjacent blocks. Applying the method of steel check dam by suggested Katsuki, this study proposes the analysis method of shear resistance of Steel Frame Retaining Wall (SFRW). SFRW is a kind of the cellular structure. PFC2D, which is the universal program for DEM, is used to analyze the characteristics of cellular structure.

After modeling the cellular structure, the material properties and boundary conditions of the individual elements are assigned. Then, the external boundary

Fig. 1 The analysis procedure of PFC2D for cellular structure



conditions such as loading forces, moments are set. Finally, we calculate the force, moment, speed, location by repeating the calculation in each time step. The above procedure is presented in the Fig. 1.

3.1 Contact Bond Model and Parallel Bond Model

The SFRW is composed of the cell body(steel frame) and fill material. Each part show the behavior of continuum and dis-continuum, respectively. Regardless of whether it is a continuum or not, two parts are modeled by using then Ball model in the Fig. 2. However, it is possible to model the continuous or discontinuous characteristics by the use of contact bond model for discontinuous bodies and Parallel bond model for continuous bodies. Using the ball element, the model of cellular structure is generated in the PFC2D, as the shown Fig. 3.

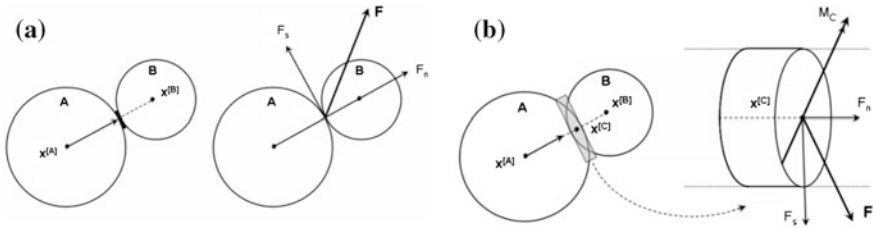


Fig. 2 Ball element model **a** Contact bond model **b** Contact bond model

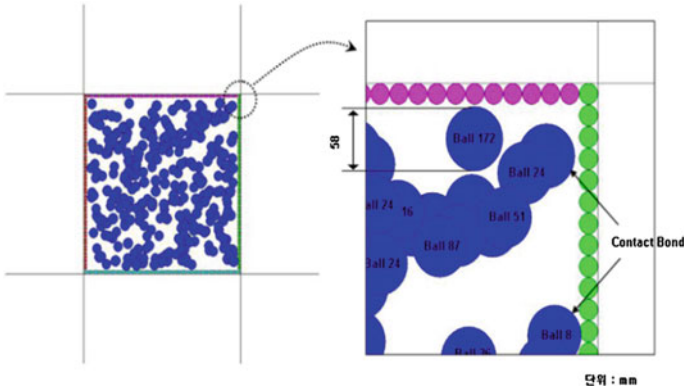


Fig. 3 Modeling of cellular structure using PFC2D

3.2 Shear Deformation

The shear deformation of fill material is influenced by its sedimentary process. To model the shear behavior of fill material considering its the sedimentary process, distinct elements for interior fill materials is randomly generated and rearranged by the gravitational force. The numerical analysis of this study is containing the above process. The Fig. 4 shows the example using PFC2D.

Fig. 4 Shear deformation considered natural filled state
a Natural filled state-PFC2D
b Natural filled state-PFC2D

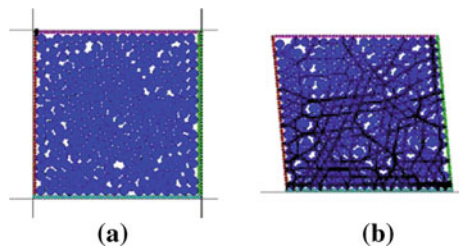


Fig. 5 Shear deformation experiments of SFRW **a** Steel frame **b** Horizontal Loading



Table 1 Properties of the fill materials

	Debris type	Grain size distribution (mm)	Average diameter (mm)	$C \dot{\theta}$	γ (kN/m ³)
Case 1	Gravel	25 ~	27	46	16.7
Case 2	Gravel	40 ~	44	47	16.4
Case 3	Gravel	55 ~	58	52	15.8

4 Shear Deformation Experiments of Steel Frame Retaining Wall

As the Fig. 5, shear deformation experiments are carried out to compare the computational results. The experiments are performed for three cases in the Table 1. The results are indicated as the Figs. 6 and 7. From the results, internal friction angle of fill material is most sensitive. Internal friction angle also grows as the shear resistance increases proportionally. As vertical loading was increased, the shear resistance was increased.

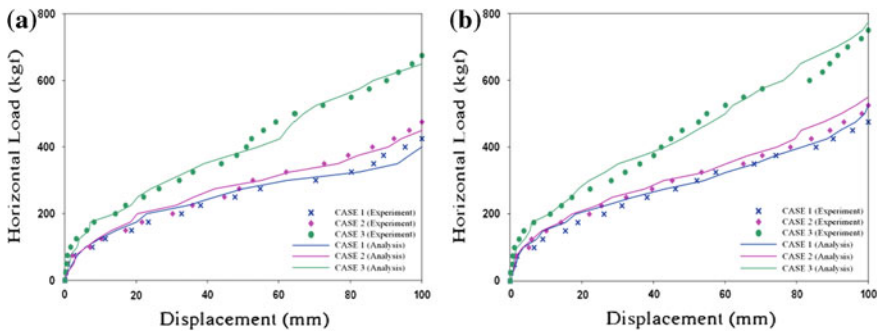


Fig. 6 The shear force–displacement relationship according to the overburdening **a** No overburdening **b** Overburdening: 10 kN

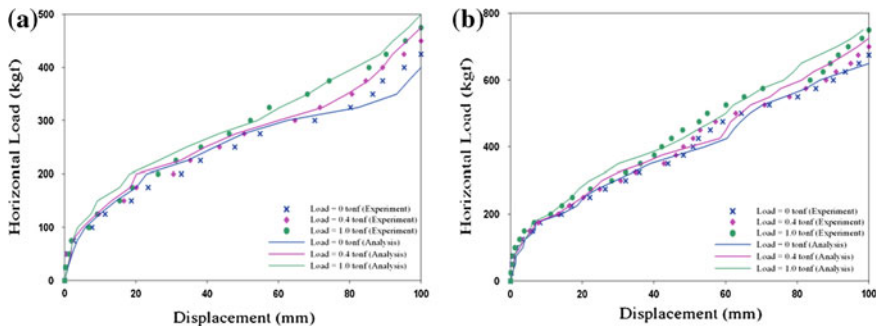


Fig. 7 The Shear force–displacement relationship according to the type of fill **a** Case 1 **b** Case 3

5 Conclusion

This paper presents an analysis approach on the shear resistance force of fill materials in a cellular structure. In order to estimate our approach, shear resistance analysis is performed by introducing displacement incremental analysis into the DEM and its numerical results are compared with the experimental results. The results show a good agreement with each other. It also led to conclusion that friction angle of fill material and vertical load are main factors influencing shear resistance, and both factors have a proportional relation to shear resistance.

References

1. Iwashita K, Oda M (1996) Distinct element method with the effect of moment transfer at the contacts. In: The third asian-pacific conference on computational mechanics, APCOM'96, pp 2187–2192
2. Terzaghi K (1945) Stability and stiffness of cellular cofferdams. ASCE 110

Full-Scaled Experiment for Behavior Investigation of Reinforced Concrete Columns with High-Strength Wire Ropes as Lateral Spiral Reinforcement

Kyu Won Kim, Jong Hwa Won, Sam Dong Jung, Jung Won Park and Moon Kyum Kim

Abstract This study performs the test with full-scaled models for each diameter to evaluate the behavior of the reinforced concrete columns using the spiral wire rope as a lateral reinforcement of the circular columns. This study performs the quasi-static test which induced binding shear destruction to review circular concrete columns bound with high strength wire rope. Three specimens of concrete columns with 700 and 800 mm diameter each were manufactured to evaluate the movement of reinforced circular columns with the wire rope. A hoop reinforced column and 2 spiral reinforced columns were manufactured to compare the bending history and the ductility reinforced with the wire rope to existing reinforced concrete columns. The peak strain of the longitudinal reinforcements is occurred at approximately 0.75 D from the bottom of the column. The spirally confined concrete columns with high-strength wire ropes showed that improve 2.5 % of the shear strength.

Keywords Wire ropes · Spiral · Ductility · Columns · Lateral reinforcements

K. W. Kim · J. H. Won · S. D. Jung (✉) · J. W. Park · M. K. Kim
Department of Civil and Environmental Engineering, Yonsei University,
Seoul, South Korea
e-mail: sdjung@kictep.re.kr

K. W. Kim
e-mail: kyu_won@yonsei.ac.kr

J. H. Won
e-mail: exameve@yonsei.ac.kr

J. W. Park
e-mail: wildflower@kictep.re.kr

M. K. Kim
e-mail: applymkk@yonsei.ac.kr

1 Introduction

The ductility of reinforced concrete columns is improved by the confinement force from lateral reinforcements including hoop or spiral reinforcements. The ductility of columns increases with an increase in the amount and strength of lateral reinforcements, but decreases with an increase in concrete compressive strength and axial load level. High transverse reinforcement ratio is not economic, causes dense rebar to make concrete construction difficult and becomes a reason of early spalling of the concrete cover due to planes of weakness between the core and concrete cover portions. Using high-strength stiffener subsequently decreases the volume ratio of the lateral reinforcements. In some cases, the transverse reinforcement plays a role in degrading the performance of the reinforced concrete columns. Meanwhile, the wire rope which is proposed as an alternative to the transverse reinforcement in the study has high-strength and high-flexibility. Generally, the wire rope with higher than 1500 MPa of the tensile strength has high ductility and makes it possible to form spiral structure of the stiffener, showing the effect of existing hoop and spiral lateral reinforcements without additional tension in case of bending the concrete columns. Therefore, the wire rope facilitates the array of structural steels in the column as a lateral reinforcement shortens the construction period and improves bending performance of the column. The lateral reinforcements with the spiral shape using the wire rope are expected to be effective in enforcing ductility of the column and preventing the longitudinal reinforcement from buckling than existing binding method.

The study performs the test with real models for each diameter to evaluate the movement of the reinforced concrete columns using the spiral wire rope as a lateral reinforcement of the circular columns, compares and analyzes the results to those of circular concrete columns with hoop-type lateral reinforcement.

2 Experimental Program

2.1 Specimen Specification

This study performs the quasi-static test which induced binding shear destruction to review circular concrete columns bound with high strength wire rope. Three pieces of concrete columns with 700 and 800 mm diameter each were manufactured to evaluate the movement of reinforced circular columns with the wire rope. One hoop reinforced column and 2 spiral reinforced columns were manufactured to compare the bending history and the ductility reinforced with the wire rope to existing reinforced concrete columns and the accuracy of the test and the test was performed under constant axial load and repetitive lateral load. The axial force ratio is fixed to 0.1 and Table 1 summarizes the features of the tested specimens.

Table 1 Specification of test specimens

Specimen	D	d	Longitudinal steel		Transverse steel			$\frac{P}{f_{ck}A_0}$
			Ratio (%)	Confinement	Material	Ratio (%)	Space (mm)	
SP700-W1	700	600	2.0	Spiral	Wire	0.16	150/250	0.1
SP700-W2								
HP700-S1				Hoop	Steel	0.33		
SP800-W1	800	640		Spiral	Wire	0.14		
SP800-W2								
HP800-S1				Hoop	Steel	0.29		

A stub with the top of $700 \times 700 \times 550$ mm and the bottom of $1000 \times 2400 \times 900$ mm was installed on the column with 700 mm diameter and 3000 mm height. The main rebar ratio is 2.11 % and 16 heterogeneous steel bars with 25 mm diameter are installed. A stub with the top of $800 \times 800 \times 550$ mm and the bottom of $1100 \times 2400 \times 900$ mm was installed on the column with 800 mm diameter and 3500 mm height, the main rebar ratio is 2.02 % and 20 heterogeneous steel bars with 25 mm diameter are installed. The hoop has 13 mm diameter and the wire rope has 9.53 mm nominal diameter. All the lateral reinforcements have the same interval and the stiffener ratios depending on the core volume ratio of the columns are 0.0045 for the hoop and 0.0025 for the wire rope. Specific tension is not introduced when applying lateral reinforcement with the wire rope. Figure 1 shows the cross-section of the test material and the rebar location of the stiffeners.

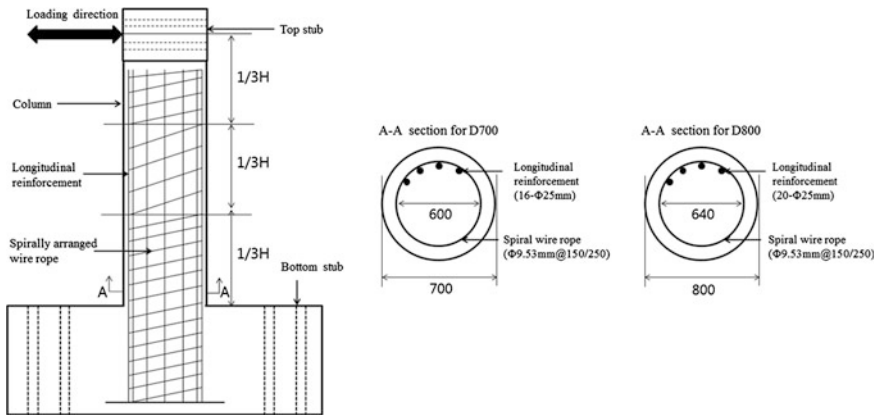


Fig. 1 Specimen details (mm)

2.2 Material Properties

The design strength of the concrete is 27 MPa and the maximum size of the coarse aggregates is 25 mm. Ready-mixed concrete is used to construct the concrete. The wire rope is GAC 7×19 (9.53 mm) with the tensile strength of 1200 MPa and the longitudinal reinforcements is a deformed bar with 25 mm diameter and 300 MPa of yield strength.

2.3 Test Procedure and Instrumentation

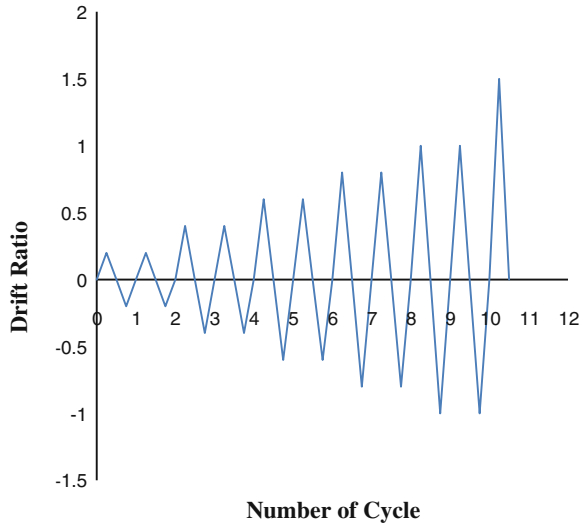
The columns are tested under constant axial load and repetitive lateral loads. The lower stub in the column is completely fixed by the steel angle blocks installed on both sides of the stub and 8 bolts which penetrated into the bottom sub to prevent the test material from moving under the influence of the lateral load. The compressive force is added by 2 oil jacks with the capacity of 1000 kN from the steel loading beam installed on the top stub and 2 frames installed on the side of the test material. The center of the steel loading beam is the same with the central axis of the test material. The lateral load is added by the oil jack operated by the air pump. The center of the actuator is the same with the center of the top stub and completely fixed by 4 bolts. The lateral load is actuated after applying the planned axial load and the test is completed when the buckling appears on the longitudinal reinforcement or the wire rope is fractured. Figure 2 shows the test specimen of the real model.

The lateral load is controlled by the displacement history shown in Fig. 3. The load pattern is under the displacement control method and the increment is determined by the draft ratio of the column and 2 cycles are repeated in the

Fig. 2 Test specimen of the full-scaled Model



Fig. 3 Lateral displacement



displacement increase. As shown in Fig. 4, the ratio increases by 0.2–1 % to specially observe the movement of the initial column and difference in the column movement for the same amplitude is compared and analyzed through repetitive loading as the fatigue movement as 2 cycles increasing 0.5 % of the ratio after 1 %.

The axial load is measured by the load controller and the lateral load and the displacement are recorded by the load cell installed between the oil jack and the hinge. The 200 mm displacement gauge is installed at 1/3H and 2/3H to measure the column displacement and the 5 mm displacement gauge is installed to check the movement of the bottom stub under the application of the lateral load. The distortion ratio of the longitudinal reinforcement is measured by the strain gauge at 0.1H, 0.75D, 0.3H, 0.5H and 0.8H.

3 Experiment Results

The Fig. 4 presents the strains on longitudinal reinforcement. The yielding occurs on the transverse reinforcement in 1 % of lateral displacement in the ratio of height/diameter with concrete cover spalling. Generally, the reinforcement began to buckle approximately at 0.75D from the bottom of the column. The buckling shape of the spirally reinforced column is larger than it of the column reinforced by existing method. Figure 5 shows the buckling shape of longitudinal reinforcement for each SP700-W1, SP700-W2 and HP700-S1.

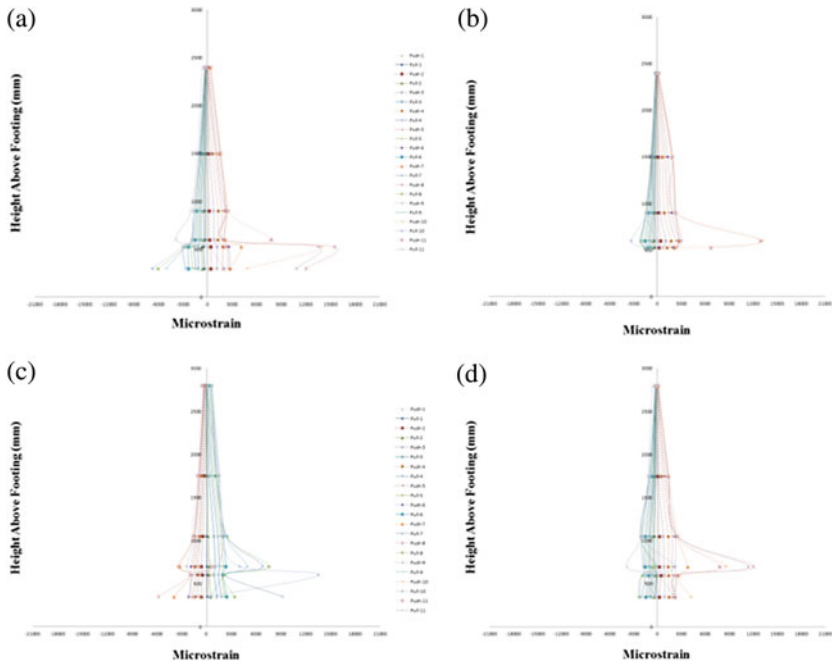
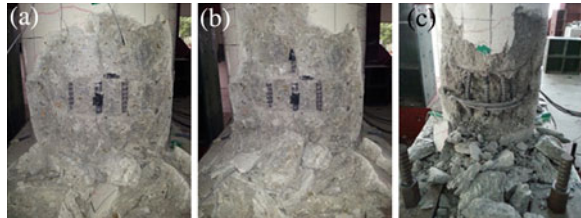


Fig. 4 Strains on longitudinal reinforcement. **a** SP700-W1 **b** HP700-S1 **c** SP800-W1 **d** HP800-S1

Fig. 5 Buckling shape of longitudinal reinforcement. **a** SP700-W1 **b** SP700-W2 **c** HP700-S1



4 Conclusion

In this study, the full scaled test is conducted for each diameter to evaluate the movement of the reinforced concrete columns using the spiral wire rope as a lateral reinforcement of the circular columns. A hoop reinforced column and 2 spiral reinforced columns were manufactured to compare the bending history and the ductility. The buckling on the longitudinal reinforcements arose at approximately 0.75D from the bottom of the column. The spirally confined concrete columns with high-strength wire ropes showed that improve 2.5 % of the shear strength.

References

1. Yang KH (2012) Flexural behavior of RC columns using wire ropes as lateral reinforcement. *Mag Concr Res* 63(3):269–281
2. Budek AM, Priestley MJN, Lee CO (2002) Seismic design of columns with high-strength wire and strand as spiral reinforcement. *ACI Struct J* 99(5):660–670
3. Priestley MHN, Park R (1987) Strength and ductility of concrete bridge columns under seismic loading. *ACI Struct J* 84(1):61–76

Local Deformed Diameter Analysis of a Pipe in Rigid Encasement for Water-crossings Application

Jong Hwa Won, Gun Kim, Sam Dong Jung, Jung Won Park,
Do Hak Kim and Moon Kyum Kim

Abstract This article presents a new type of stress assessment method for double-layered pressure vessel. Due to its robust strength, a steel–concrete composite pipe is generally installed in poor burial conditions. However, it is difficult to directly define the stress state on pipe sections. For the convenient stress estimation of pipes encased in non-circular concrete, this article suggests a Stress Index (SI), a function of the interface pressure and changed diameter, based on the interface pressure. The ovalization characteristics of plain pipes and of pipes encased by circular and rectangular concrete are examined. The resultant stresses are replaced by the non-dimensional ratio of diameter change; thus, the stress index has the advantage of application for stress assessment irrespective of pipe characteristics.

J. H. Won · S. D. Jung (✉) · J. W. Park · M. K. Kim
Department of Civil and Environmental Engineering, Yonsei University,
Seoul, South Korea
e-mail: sdjung@kictep.re.kr

J. H. Won
e-mail: exameve@yonsei.ac.kr

J. W. Park
e-mail: wildflower@kictep.re.kr

M. K. Kim
e-mail: applymkk@yonsei.ac.kr

G. Kim
School of Civil and Environmental Engineering, Georgia Institute of Technology,
Atlanta, GA, USA
e-mail: gunkim@gmail.com

D. H. Kim
GS E&C Research Institute, GS Engineering & Construction Corporation,
Yongin, South Korea
e-mail: dohkim@gsconst.co.kr

Keywords Pressured pipeline · Encasement · Stress index · Stress assessment · Interface pressure

1 Introduction

A pipeline is the most popular means of energy transportation, having economical, convenient and reliable merits for long-distance energy transportation such as gas or oil. Almost all pipelines, however, have fundamental maintenance problems because they are buried underground or located in offshore and rivers, making access difficult. Thus, in order to reduce maintenance costs and improve the long-term performance of pipeline structures, the construction of load-supporting or anti-corrosion structures is usually demanded.

In this study, a new stress assessment method is proposed for buried pipelines encased in rectangular concrete. And encasement-dependent load distribution characteristics of the pipe section are proposed according to the theory of elasticity. Finally, the procedure for estimating the stress of non-circular pipes is simplified by using the stress index obtained from the relationship of the diameter change ratio and the interface pressure.

2 Analytical Approaches to Stresses on a Multi-layered Pipeline

The pressure acting on a pipe's surface can be generally divided into internal and external loads for calculating the thickness in the design process. In this chapter, the stresses on simple and multi-layered pipes are derived by using the equation of Lamé's thick-walled cylinder.

2.1 Interface Pressure Analysis for a Multi-layered Pipeline

In order to estimate the stress distribution between the inner and outer pipes in a multi-layered pipeline, the interface pressure between members should be defined in advance. It is assumed that the external radius of the inner pipe is the same as the internal radius of the outer pipe. Therefore, the interface pressure acts equally on both the inner and outer pipes with opposite signs. Thus, the interface pressure at the contact surface between inner and outer pipes is defined as following Eq. 1 [1].

$$P_c = [E\delta(b^2 - a^2)(c^2 - b^2)] / [3b^2(c^2 - a^2)] \quad (1)$$

where, E = elastic modulus, δ = radial displacement, a and b = internal and external diameter of inner pipe and, b and c = internal and external diameter of outer pipe.

3 Numerical Analysis of Concrete-Steel Composite Pipelines

Korea Gas Corporation (KOGAS) has established pipeline design specifications for safe and efficient operation of pipeline networks. There are many types of classifications for materials, standard size, and operational pressure on guideline. Generally, API 5L X 65 is used for gas transportation pipelines with a 762 mm diameter and a 17.5 mm wall thickness.

In order to reasonably identify the characteristics of a double-walled pipe with a rectangular concrete encasement, a pipe encased in circular concrete and a plain pipe are also examined for comparison. The external diameter of the circular encasement is calculated as 1,212 mm in order that the sectional modulus of a circular double-walled pipe is same as that of a pipe with rectangular concrete encasement, through a modification of the moment of inertia. To describe the burial condition, the material properties of surface soil in Han River (Seoul, Korea) are used in this research. Figure 1 shows the representative pipeline sections for this research based on KOGAS design criteria. The material properties of the pipe, concrete and soil are summarized in Table 1.

In this study, the internal pressure and cover depth of the pipes were taken to be 0–7.84 MPa and 0–8 m, respectively, in consideration of the KOGAS operational guidelines. The DIANA solver based MIDAS/GTS ver. 2.0.2 was employed to characterize the soil-structure system.

The accuracy of the analytical solution as a function of the interface pressure was verified prior to detailed analysis for a specific loading condition. This process ensures the validity of further research in this paper. From the results of analytical and numerical studies, an error of 0.18–0.32 % was found, and most of the stress data from the numerical analysis show good agreement with the analytically obtained exact solution.

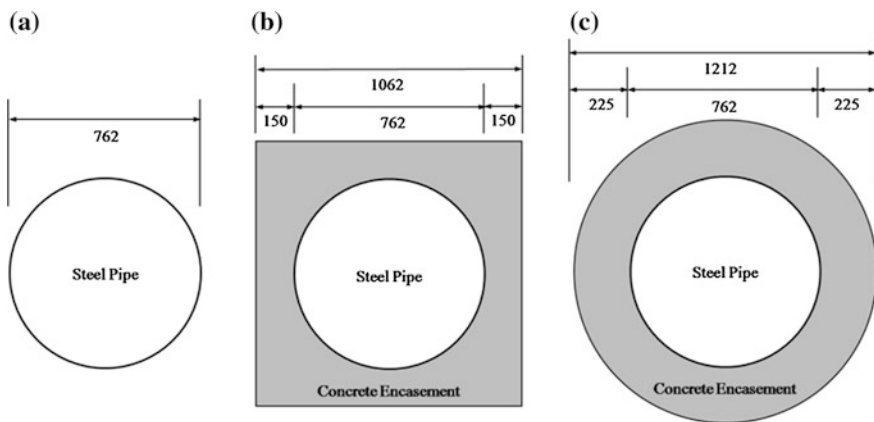


Fig. 1 Representative pipe section **a** plain pipe **b** pipe encased in rectangular concrete [2] **c** pipe encased in circular concrete

Table 1 Material properties of the pipe, concrete and soil

	$E(MPa)$	$\gamma(MPa)$	ν	$C(MPa)$	Remarks (cm)
Steel Pipe (API 5L X65)	2.1×10^5	7.7×10^{-6}	0.3	–	$t = 1.75$
Concrete	2.01×10^4	2.5×10^{-6}	0.2	–	–
Soil	5.9	1.7×10^{-6}	0.35	0	–

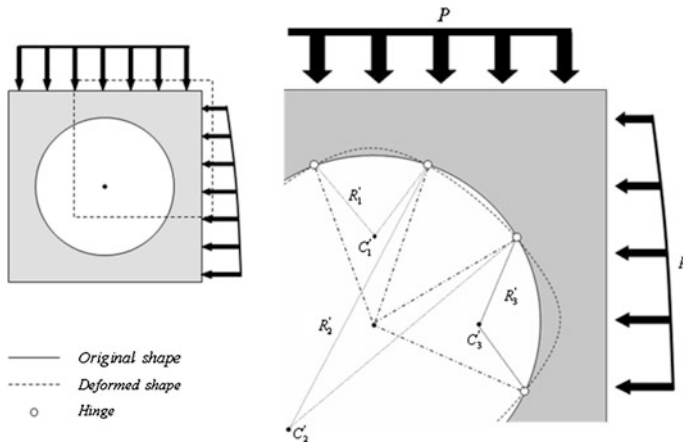


Fig. 2 Local deformation of encased pipe [1]

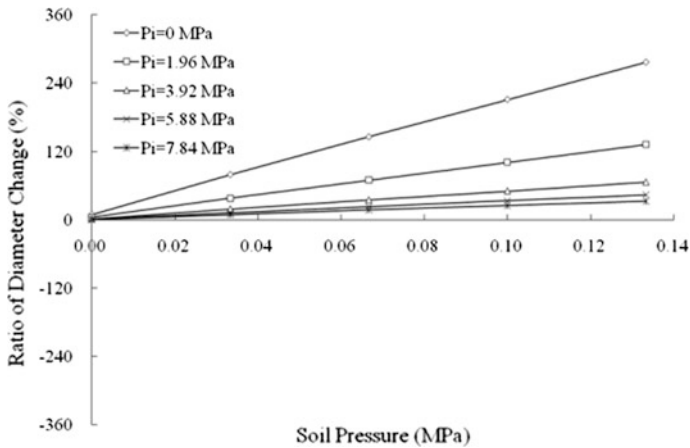


Fig. 3 Ratio of diameter change versus soil pressure at the top

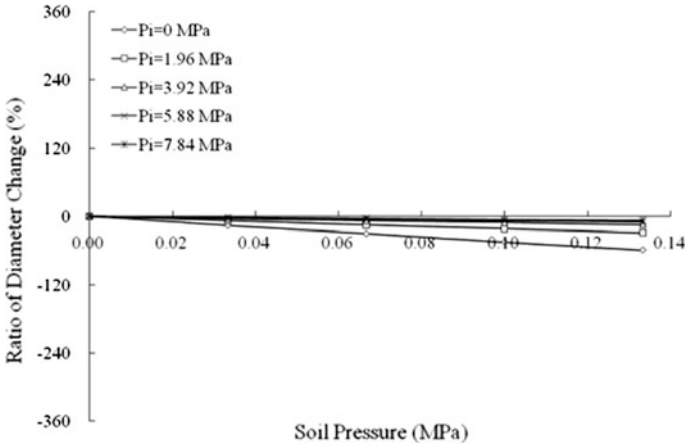


Fig. 4 Ratio of diameter change versus soil pressure at the *shoulder*

3.1 Behavior Characteristics of Bare Pipe and Encased Pipe

Typical flexible pipe in the soil has an elliptical shape due to compressive force acting at the top and bottom of the pipe. The behavior of a rectangular concrete covered pipe is presented in Fig. 2. Without external pressure, every part of the pipe has a Point C as a center of the pipe with radius R. However, as local deformation develops by external pressure, the radius for each part of the pipe has various values and it results in different local deformed diameter.

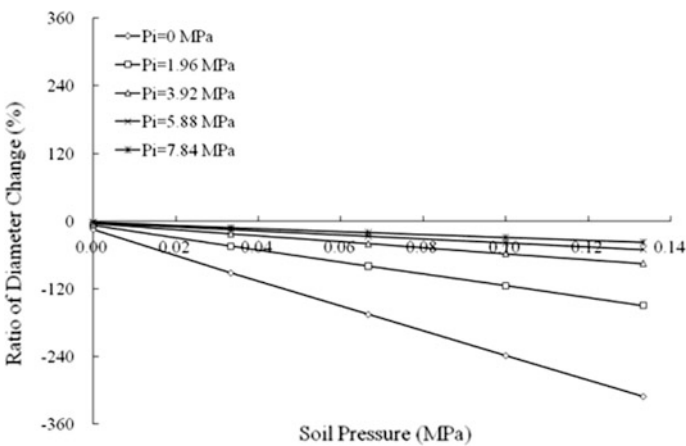


Fig. 5 Ratio of diameter change versus soil pressure at the *springline*

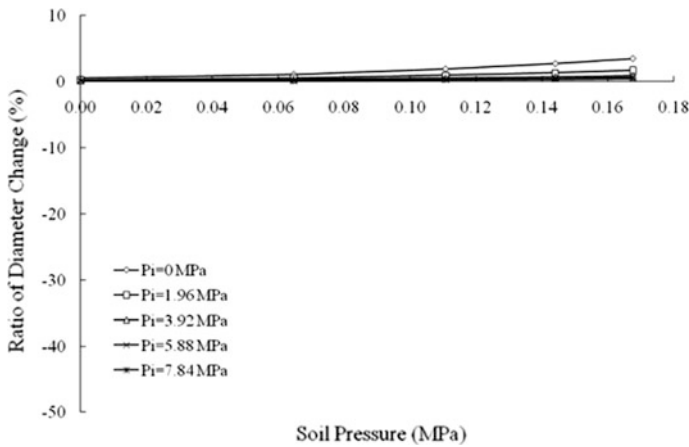


Fig. 6 Ratio of diameter change versus soil pressure at the *top*

4 Analysis of the Ratio of Local Deformed Diameter Change and the Stress Index

To consider all terms as non-dimensional variables, the shape of the encasement and the deformed diameter with respect to soil depth were replaced by the ratio of diameter change to the original diameter ($D_o = 762$ mm). Thus, with these simplification procedures, it is possible to reduce the units of the variables. In addition, by setting the ratio of diameter change as a variable, the degree of stress can be expressed as a function of the pressure acting on the contact surface and the non-dimensional ratio of diameter change, irrespective of the shape and type of the

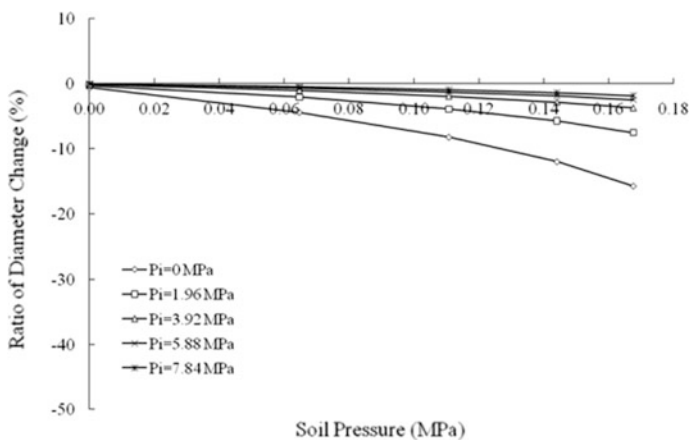


Fig. 7 Ratio of diameter change versus soil pressure at the *shoulder*

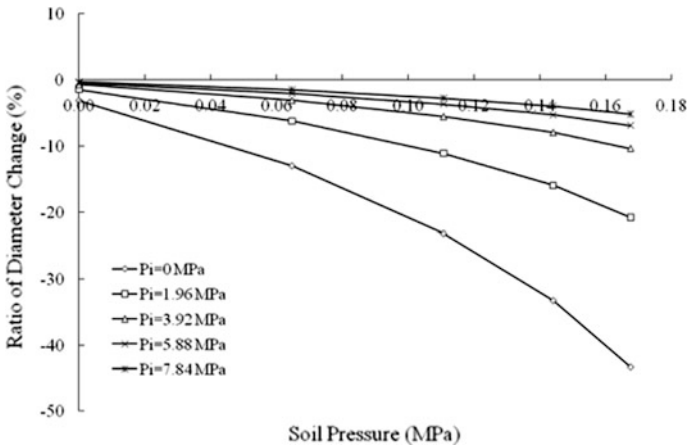


Fig. 8 Ratio of diameter change versus soil pressure at the *springline*

pipeline. The ratio of Local Deformed Diameter Change is presented in Figs. 3, 4, 5, 6, 7, 8, 9, 10 and 11. Figures 3, 4 and 5 is for a bare pipeline, Figs. 6, 7 and 8 is for a double layered pipeline and Figs. 9, 10 and 11 is for a pipeline encased in the rectangular concrete box (in the last page).

For a simple stress estimation of underground pipelines, this research proposes a Stress Index (SI) based on the findings of the present study. The SI is defined by Eq. 2, with units of force per area. Moreover, this is could be effective to obtain the circumferential stress as following Eq. 3.

$$SI = 0.5P_c[(100 + R)/100] \tag{2}$$

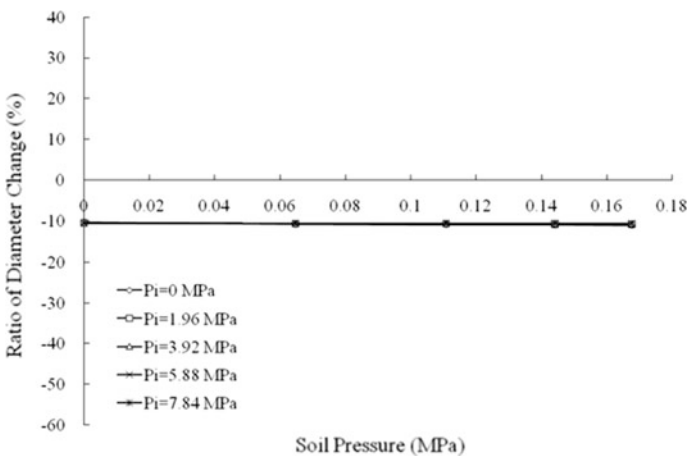


Fig. 9 Ratio of diameter change versus soil pressure at the *top*

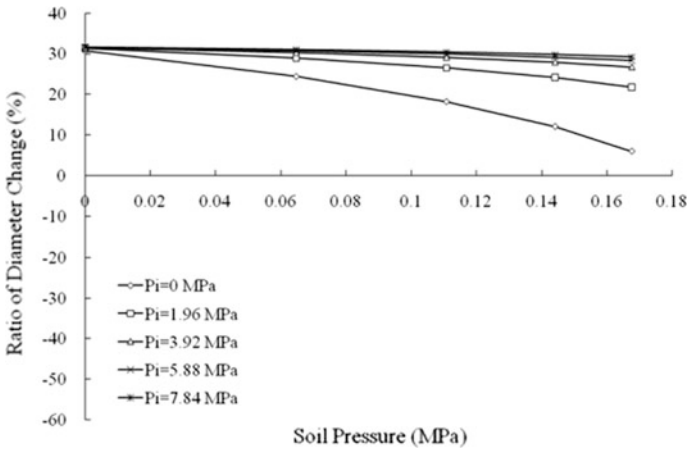


Fig. 10 Ratio of diameter change versus soil pressure at the *shoulder*

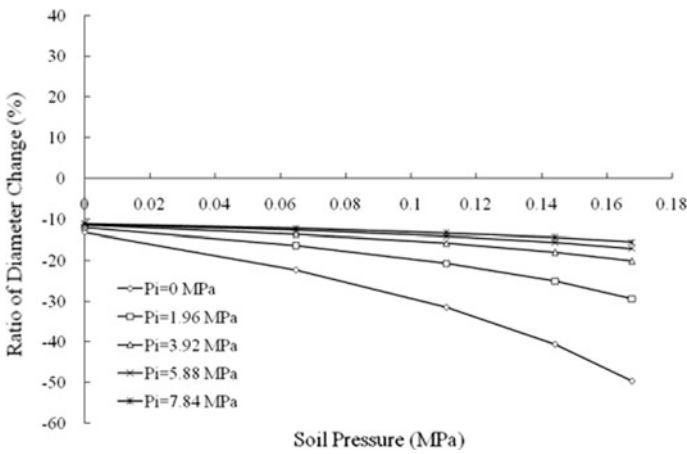


Fig. 11 Ratio of diameter change versus soil pressure at the *springline*

$$\sigma_c = D_o SI / t \tag{3}$$

where, SI = stress index, R = the ratio of diameter change and D_o = original diameter. When these factors are all known, the stress index can be determined based on Fig. 12.

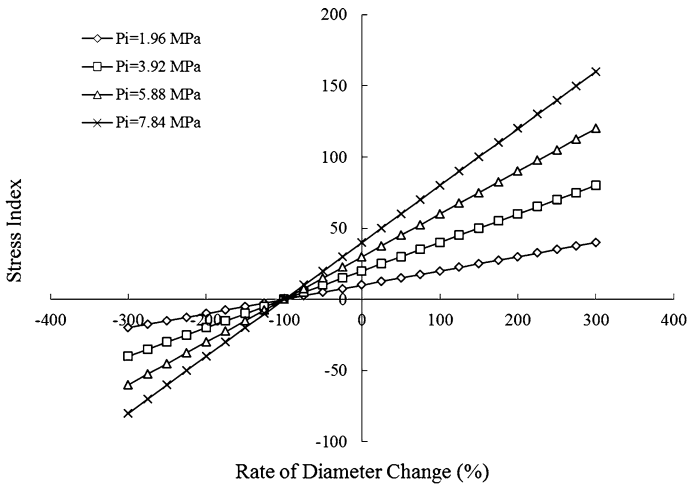


Fig. 12 Stress index for the ratio of diameter change

5 Conclusion

This study proposes a stress index for the stress estimation of a double-layered pipeline covered with a circular or a rectangular encasement. The stress index has the advantage of wide utility for stress assessment, irrespective of pipe characteristics. It is expected to realize the reasonable design of non-circular multi-layered pipeline.

References

1. Won JH, Kim MK, Ryu DH (2007) Characteristics of an encased gas transportation pipeline in offshore application. In: Proceedings of the 17th international offshore and polar engineering conference, ISOPE, III. Lisbon: ISOPE, pp 2700–2706
2. KOGAS (2006) The estimation of integrity assessment method for pipeline crossings over road and river. KOGAS, Korea

A Study on Management System Design of Swimming Exercise Prescription by Using Fussy ANP

Kyoung-Hun Kim, Won-Hyun Kim, Tae-Won Kyung,
Gyeng-Taek Yu and Chung-Sick Shin

Abstract It has become general common sense through numerous researches that exercise provides positive impacts on physical and mental health. And it has been reported that regular exercise adjusts obesity by reducing body fat and lipid levels found in the blood and ultimately, it improves human quality of life. In this study, indices for managing swimming exercise therapy were induced through prior researches and weighted value was measured by modelling correlations between indices by using fuzzy Analytic Network Process (ANP) technique. And patient management system was intended to be realized so that tailor-made management per patient can be established on real-time through mobile equipments such as portable phone, smart phone, notebook and etc.

K.-H. Kim (✉) · G.-T. Yu
School of Computer Information, GangDong University, 1, JangHoWon,
Icheon-Si, Gyeonggi-Do, South Korea
e-mail: iioii.net@gmail.com

G.-T. Yu
e-mail: rhyukt@gangdong.ac.kr

W.-H. Kim
School of Community Sports Major, DaeDuk University, 48, Jang-Dong,
Yuseong-Gu, Daejeon, South Korea
e-mail: whkim@ddu.ac.kr

T.-W. Kyung
Department of Patent Performance Management, R&D Patent Center,
Seoul, South Korea
e-mail: twkyung@rndip.re.kr

C.-S. Shin
School of Sports Major, Inha University, 402-751, inha-ro, Nam-gu,
Incheon, South Korea
e-mail: inhaelite@empal.com

Table 1 Whether or not of diseases per answerer of questions

Classification	People with experiences of musculoskeletal disease	People with experiences of cardiovascular disease	Others ^a	Total
C sport center located in Seoul city	19	7	11	37
T sport center located in Suwon city	16	6	8	30
J sport center located in Incheon city	17	8	4	29
Total	52	21	23	96

Keywords Fuzzy · ANP · Fuzzy ANP · Swimming exercise treatment

1 Introduction

It has become general common sense through numerous researches that exercise provides positive impacts on physical and mental health. And it has been reported that regular exercise adjusts obesity by reducing body fat and lipid levels found in the blood and ultimately, it improves human quality of life. Fields of exercise prescription can be largely divided into two fields, first, it is the field which general public can prevent diseases and rehabilitate through continuous health management. Second, it can be classified into the field of patient management and therapy by hospitals and physician's offices.

In this study, indices for managing swimming exercise therapy were induced through prior researches and weighted value was measured by modelling correlations between indices by using fuzzy Analytic Network Process (ANP) technique. And patient management system was intended to be realized so that tailor-made management per patient can be established on real-time through mobile equipments such as portable phone, smart phone, notebook and etc. Therefore, more systematic and fast patient tailor-made services become possible while data sharing is possible on real-time for both exercise therapists and patients.

Table 2 Effective answerers per age

Classification	People with experiences of musculoskeletal disease	People with experiences of cardiovascular disease	Total
Ages of 20 years old	2	0	2
Ages of 30 years old	4	2	6
Ages of 40 years old	19	6	25

2 Object and Method of Research

2.1 Research Object

Objects of this study were people with experiences of musculoskeletal disease and cardiovascular disease among adult male and female who are participating in swimming programs of sport centers located in the Seoul Metropolitan region and it was investigated using questionnaire and detail characteristics of the object people of the questionnaire are like Tables 1 and 2.

2.2 Research Method

As an investigation tool for this study, questionnaire was used and as for the questionnaire to measure impacts of swimming on mental health, we have re-standardized for conformity with this study and have used (BSI) of simple mental diagnosis investigation paper (SCL-90-R) which had re-standardized Symptom Check List (SCL-90) which was made by Derogatis et al. (1973) by developing Hopkins Symptom Check List (HSCL) in compliance with our situation by Gwang-Il Kim and Ho-Taek Won (1984) et al. And as tools for measuring psychological, physical, environmental and social satisfaction through swimming, the questionnaire developed by Ragheb and Beard (1980) and used by translation in researches of Mi Kim (1997) and Sung-Cheol Lee (1992) et al. was referred and we used the questionnaire which was developed through modification and supplementation in conformity with this study.

In this study, prior research data with regard to impacts of swimming on patients were analyzed and based on these data, indices were re-arranged like the following (Table 3) for designing exercise therapy management system for the object of patients utilizing swimming as therapy methods.

Table 3 Index for swimming exercise therapy management

Items	Index	Explanation
Environmental element	Cleanliness of facility and place (A1)	Exercise facility and place are clean and fresh
	Interior of facility and place (A2)	Exercise facility and place are well decorated
	Modernization of facility and equipment (A3)	Exercise facility and equipment must be the most recent ones
Mental health element	Anxiety symptoms (B1)	Physical symptoms related to anxiety such as tension, impatience, oversensitiveness and etc.
	Melancholia symptoms (B2)	Lower emotion such as lacking of motivation, lost of vitality, sense of frustration and etc.
	Somatizing symptoms (B3)	Symptoms appealing periodically about abnormalities of physical function
Physical element	Improvement of physical strength (C1)	Physical strength can be improved
	Recovery of physical vitality (C2)	Vitality will be restored physically
	Stress elimination (C3)	It becomes helpful for eliminating stress
	Test about physical capability(C4)	I can test my physical capability through the swimming program I am practicing
Psychological element	Emotional stability (D1)	Swimming program makes to have emotional stability
	Interest (D2)	Swimming program is very interesting
	Psychological confidence (D3)	Swimming program provides me with confidence
Social element	Consideration for other people (E1)	Swimming program enables to understand other people
	Relationship with peers (E2)	Other relationship with peers participating in swimming program
	Relationship with therapists (E3)	Relationship with therapists instructing swimming program

3 Weighted Value Analysis by Using Fuzzy ANP Technique

In this study, indices for designing therapy management system of swimming exercise were established for patient management and for measuring weighted values and priority of these indices, Fuzzy ANP technique was used.

In this study, fuzzy theory was applied to ANP technique and attempts were made to minimize absurdity in selecting object people of questionnaire.

Table 4 Indices for management of swimming exercise therapy

Items	Index	Weighted value	Priority
Environmental element	Cleanliness of facility and place	0.011285	12
	Interior of facility and place	0.001736	16
	Modernization of facility and equipment	0.007812	14
Mental health element	Anxiety symptoms	0.010749	13
	Melancholia symptoms	0.031134	8
	Somatizing symptoms	0.040392	6
Physical element	Improvement of physical strength	0.114614	5
	Recovery of physical vitality	0.238322	1
	Stress elimination	0.135614	3
	Test about physical capability	0.005414	15
Psychological element	Emotional stability	0.030115	9
	Interest	0.172728	2
	Psychological confidence	0.129707	4
Social element	Consideration for other people	0.019749	10
	Relationship with peers	0.019077	11
	Relationship with therapists	0.031551	7

In fuzzy ANP analysis processes, unique vector method of Saaty (1996) cannot be used as results of paired comparisons shall be handled as number of fuzzy. Therefore, in this study, extent analysis method on fuzzy AHP method of Chang (1996) was applied for handling number of fuzzy in ANP analysis processes.

4 Conclusions

In this study, weighted value and priority of elements providing impacts on exercise therapy of swimming participants were analyzed. ANP technique, which applied fuzzy theory, was used for analysis of weighted value and priority.

Table 4 are weighted values and priorities of 16 indices induced through ANP technique. Cleanliness of facility and place, Interior of facility and place, Modernization of facility and equipment, Anxiety symptoms, Melancholia symptoms, Somatizing symptoms, Improvement of physical strength, Recovery of physical vitality, Stress elimination, Test about physical capability, Emotional stability, Interest, Psychological confidence, Consideration for other people, Relationship with peers, Relationship with therapists

In this study, elements providing impacts on exercise therapy were established to design management system of effective swimming exercise therapy for patient management and through Fuzzy ANP technique, priority was induced through analysis of relationship between each index.

Figure 1 is the result of analyzing the entire priority with regard to 16 indices. An index which has obtained the highest weighted value was ‘Recovery of physical vitality’ which was included in item of “physical element” and obtained



Fig. 1 Realization of actual use environment in PC, portable phone and smart phone

the value of 0.2383. The second was ‘Interest’ index included in item of “psychological element” and has obtained the value of 0.1727. An index which has obtained the third highest weighted value was ‘Stress elimination’ included in “physical element”

It is judged that this study can find meanings in the following aspects.

First, the importance of exercise therapy was induced from the position of patients.

For establishing exercise therapy system, exercise was applied for the purpose of therapy with object of patients now going through or experienced musculo-skeletal disease or cardiovascular disease and in particular, elements people feel through swimming physically and mentally were re-arranged.

Second, level of importance (Relative importance) for each element was quantified.

Exercise therapy until now was carried out from the perspective of therapists. However, in this study, main interest indices were classified from the perspective of patients and level of importance of each index was quantified and its validity was verified. In particular, when calculating weighted values of indices, mutual relationship between each index was reflected using ANP technique.

References

1. Gang S-J, Gang M-O, Lee H-J, Lee H-Y, Jin Y-S (2003) Development of underwater exercise programs for improving cardiopulmonary capacity and physical strength for patients of coronary artery disease. *Korea Sport Med J* 21(2):151-160
2. Saaty TL (1996) Decision making with dependence and feedback: the analytic network process. RWS Publications, Pittsburgh

3. Zadeh LA (1978) Fuzzy sets as a basis for a theory of possibility
4. Establishment of mid- and long-term comprehensive plan for U-Healthcare invigoration Korea Health Industry Development Institute, Korea Health Industry Development Institute, Seoul (2008)
5. Beaufriere B, Morio B (2000) Fat and protein redistribution with aging: metabolic considerations. *Eur J Nutr* 54(Suppl 3):S48–S53
6. Meade L, Sarkis J (1999) Analyzing organizational project alternatives for agile manufacturing processes: an analytic network approach. *Int J Prod Res* 37(2):241–261

Intelligent Recommendation System for Automotive Parts Assembly

Jong-Won Ko, Su-Jin Baek and Gui-Jung Kim

Abstract This study proposed a method of developing an intelligent recommendation system for automotive parts assembly. The proposed system will display the detailed information and the list components which make up the relevant part that an user wants through the database using the ontology when selecting an automotive part that an user intends to learn or to be guided of. The intelligent recommendation system for parts is offered to users through determining the automatic recommendation order between parts using the weights. This study has experimented the principles of the recommendation system and the method of setting the weights by setting two scenarios.

Keywords Intelligent recommendation system · Automotive parts · Weight

1 Introduction

The intelligent recommendation system for automotive parts assembly is a system that enables us to know the importance and uses for the assembly of automotive parts easily without the need for special technical knowledge through expressing

J.-W. Ko · S.-J. Baek

Department of Computer Engineering, Kyunghee University, Deokyoungdaero, Giheung-gu, Yongin-si, Gyeonggi-do, South Korea

e-mail: jwko@khu.ac.kr

S.-J. Baek

e-mail: croso@khu.ac.kr

G.-J. Kim (✉)

Department of Biomedical Engineering, Konyang University, Nonsan-city, Chungcheongnam-do, South Korea

e-mail: gjkim@konyang.ac.kr

the components that have a close relationship with the part when selecting an arbitrary part based on the data generated from the automotive part that an user has hard time of approaching. Currently, the actual work site of the domestic companies is being operated as work and learning are separated. And since there is not sufficient support for atypical learning, the effect of training and learning is very low so that it is hardly linked to the achievement of business conduct. The technology related to the domestic real-time business support is mainly based on studies that focus on reviewing technologies at an early stage. Thus, in general, when having a technical problem in the workplace, the work will be supported by the manual document. In addition, there is not any method of work processes that can support the vast amount of the existing data and information and search, discovery, recommendation such as the complexity of knowledge that has been increasing in recent days. Thus, instead of ending in a one-dimensional visualization of knowledge to simply display the knowledge, it is imperative to display the knowledge that fits the context of the production, work and learning activities of an operator in progress and integrate search, recommendation and technology that is able to grasp the relation with the relevant components [1]. Thus, this study designed an intelligent recommendation system using hierarchical taxonomy that has been accomplished through the ontology. Each part is assigned a weight set by the expert knowledge of the vehicle; and the priority is set using the arithmetic product of the weights and times chosen as a user uses it directly. When using a system that uses the principle of an intelligent recommendation, it is possible to provide the appropriate information through the analysis using the user's behavior and professional view.

2 Background

The recommendation system technology refers to a method of predicting a user's interest automatically on the basis of the information obtained from many users [2]. This system is not limited to a particular user's information, the system is characterized in that it recommends while collecting information for multiple users together. For example, collaborative filtering or recommendation system for shopping predict the user's shopping preference using the list of items the user likes. This recommendation system has the advantage that users can easily find the information they need to support search in the context of mutual collaboration. In other words, this recommendation system shows that gathering knowledge properly cannot be simply done by artificial intelligence technology in the flood of information that increases exponentially [3].

It is necessary to have personalization, recommendation agents and data mining techniques for intelligent recommendation system [4]. Personalization can be defined as providing the information related to product or service to individual customers. Recommended agent is a recommendation system that analyzes customer preference to provide custom information. Data mining is a method for

extracting new and meaningful information from a large amount of data. This intelligent recommendation system is a technique to efficiently provide timely information to individual customers by accumulating and gathering information about each customer.

3 Design of Intelligent Recommendation System

3.1 Principle of Intelligent Recommendation

The principle of the intelligent recommendation system is that it displays the detailed information and the list components, which make up the relevant part that an user wants through the database using the ontology when selecting an automotive part that an user intends to learn or to be guided of [5]. But users do not know which part is the core part and more important unless they have expertise for automotive parts. The intelligent recommendation system automatically selects the recommendation order within the system and outputs this to users. The principle of determining this order is weight. Each part is assigned a weight set by the expert knowledge of the vehicle; and the priority is set using the arithmetic product of the weights and times chosen as a user uses it directly. Then, it is to make it possible to view easily which factor of which component is important through recording the arithmetic product value in the system through making users output by the order of priority. As for engine, the elements that form an engine are too many. However, users must not have deep knowledge about engine since they are not experts. Thus, they may get in trouble since they do not know how to inspect and learn engine when they do want to inspect and learn engine. However, the principle of this intelligent recommendation system will provide adequate information through the analysis using user behavior and experts view.

The recommended procedure that appears when selecting the engine is a result of that the weight of the user and the expert is reflected. The default weight with the intention of experts prevents confusion when an user uses the system for the first time and also has a significant impact even when storing and outputting data. Each time an user selects the specific feature of the engine, the weight of that feature is increased by the equation and stored in data to make an impact on the users who intend to use the system. Basically, the equation is represented by (the default weight value) \times (number of user selection) by experts. The most important feature of the principle of the intelligent recommendation system for automotive is that it is reliable. The reason is that the car absolutely needs expert opinion. Thus, placing a weight that reflects the opinions of professionals in the system increases the efficiency and reliability of the system.

3.2 *Intelligent Recommendation Scenario*

3.2.1 Scenario 1

If an user tries to use the intelligent recommendation system for assembly, when selecting a component, the system outputs the detailed process of assembly and disassembly of parts to the user. The system outputs appropriate and efficient information to users through retrieving information as to the parts based on the database established using the ontology upon outputting the relevant information. For example, it displays the list of the relevant recommended components through the database among various components that comprise the timing belt. As for the timing belt, water pump, anti-freezing liquid, tension bearing are shown as the list of recommendation. Each of the recommended components has a default weight respectively. A default weight is a value stored by experts; thereby, being reliable. Users can get help in learning the detailed information of the components through the list of recommendation to be displayed upon learning the basic assembly/disassembly course.

- ① An user selects the timing belt.
- ② Show the process of assembling/disassembling of the timing belt.
- ③ Get the information of composites of the timing belt stored in database.
- ④ Show the recommendation list by outputting the component parts based on the default weight.

3.2.2 Scenario 2

The intelligent recommendation system not only has information to be outputted at the database established by the classification of upper/lower based on the ontology. The configuration of the relevant components has basically the weight value that has reflected experts opinion. And the order of recommendation will be decided in accordance with the weight value. The value is determined through the arithmetic operation in the system. As a result, it is stored as data; thereby, having an impact on all users who use the system. The car absolutely needs expert opinion. Thus, the reliability of the information provided to users is very high and suitable and efficient for all users.

- ① Output in the order of default weight value which has been given by the expert.
- ② If using the system, data will be stored as being multiplied by the number of times it has been selected.
- ③ Data values stored is always delivered to the user.
- ④ Users are being able to obtain a more efficient and reliable data.

4 Conclusion

This study designed the intelligent recommendation system using a hierarchical taxonomy that has been accomplished through the ontology. The intelligent recommendation system for automobile parts assembly is a system that enables us to know the importance and uses for the assembly of automotive parts easily without the need for special technical knowledge through expressing the components that have a close relationship with the part when selecting an arbitrary part based on the data generated from the automotive part that an user has hard time of approaching. For example, if an user wants to inspect and replace the spark plug by the recommendation system for components since there is a problem with the spark plug among the components that make up the engine of the vehicle, it will recommend to check spark cable, oxygen sensor, crank sensor, etc. that are in the ignition device that is related to the spark plug. Each part is assigned a weight set by the expert knowledge of the vehicle; and the priority is set using the arithmetic product of the weights and times chosen as a user uses it directly. Then, it is to make it possible to view easily which factor of which component is important through recording the arithmetic product value in the system through making users output by the order of priority. It is possible to find the problems of the vehicle that are not yet discovered by the recommendation system and to ensure stability when an user operate the vehicle. In addition, it is possible for users, who do not have the special knowledge on the structure of the engine of the vehicle, to inspect, learn and manage the vehicle more easily and efficiently using the recommendation system.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0006911).

References

1. Suh HJ, Kim YH, Lee SW, Lee JS (2009) e-learning technology based on mixed reality. *Electron Telecommun Trends* 24(1)
2. Kim Y, Kim J (2011) Attack detection in recommender systems using a rating stream trend analysis. *J Korea Soc Internet Inf* 12(2):85–101
3. Nguyen NT (2007) Computational collective intelligence. Semantic web, social networks and multiagent systems. In: *ICWS 2007*, pp 1164–1167
4. Herlocker JL, Konstan JA, Riedl J (2000) Explaining collaborative filtering recommendations. In: *CSCW'00*, 2–6 Dec 2000, Philadelphia
5. Kim G-J, Han J-S (2012) Application method of task ontology technology for recommendation of automobile parts. *J Digit Policy Manag* 10(6):275–282

Model Transformation Verification Using Mapping Pattern and Model Transformation Similarity

Jong-Won Ko, Su-Jin Baek and Jung-Soo Han

Abstract Regarding the software development, MDA (Model Driven Architecture) of OMG can be regarded as the concept of making an independently-designed model according to the development environment and language and reusing it according to the desired development environment and language by expanding the reusable unit into the software model when developing software. The problem with these traditional research methods, but the first model, design model for checking the information with the model by defining a formal representation in the form of an abstract syntax tree, as you have shown how to perform validation of UML design model. Additional steps need to define more complex due to a software problem that is not the way to the model suitable for model transformation verification. In this paper, as defined in the verification based meta model for input and target model. And we also suggest how to perform model transformation verification using property matching based transformation similarity and mapping patterns.

Keywords MDA · Model transformation verification · Mapping patterns

J.-W. Ko · S.-J. Baek (✉)

Department of Computer Engineering, Kyung Hee University, Giheung-gu,
Yongin-si, South Korea
e-mail: croso@khu.ac.kr

J.-W. Ko

e-mail: jwko@khu.ac.kr

J.-S. Han

Division of Information and Communication, Baekseok University, Cheonan,
Chungnam-do, South Korea
e-mail: jshan@bu.ac.kr

1 Introduction

The initial research related to the model transformation technology for MDA has been executed by mainly focusing on such factors as the function of generating the source code supporting various converting formats or development languages and the expandability and applicability of the model-converting mechanism. Recently, some researches for the model transformation method for the optimization of the model by applying the specific metrics or design rules based on the given model and the analysis of the uniformity related to the generated model by executing the model transformation process in various points of view for the same model have been executed. Also, the research for the expandability of the model which can be used to generate an expanded model through the model transformation process automatically when the factors required for the expansion of the model from the base model are applied is being currently executed [1]. In regard to the researches which have been recently executed, some test issues have been applied to support the accurateness of the verification method or transformation rule. In such a case, the accurateness of the converting model can be largely considered in terms of the accurateness for structures and the one for meanings. The accurateness for structures represents the one for the generation of a good target model by following the transformation rule with the input of a good source model, while the accurateness for meanings represents the one for the transformation of the information desired by the developer into the target model by interpreting the subject meaning with various points of view in regard to the generated target model [2]. The previous research about the model transformation verification was executed by mainly focusing on the accurateness for structures. However, at the University of Alabama in Birmingham, USA, researches about the transformation of models by using the C-SAW model transformation engine based on the Aspect-Oriented Software Development for the development of the embedded system and the related testing framework have been executed [4]. The testing framework suggests the model-comparing algorithm which can be used to compare the differences between two graphs by expressing the structures of the input model and the target model with the nodes and edges of the graph. However, according to such a comparison, the software design model is simply expressed in the form of the version tree in the management of the software images. Also, there seems to be not much information for the transformation between different software models with various complicated points of view, while the verification between the two models is no more than the simple comparison of the graph structure. Also, when executing the transformation of the model for one input model with various points of view, it could be difficult to execute the test for uniformity between the software models. As another research for the model transformation verification, there is the research method which applies the graph patterns to the verification on the graph transformation process by independently developing the VIATRA model transformation engine [3]. According to such a research, it is possible to define the compare the input model and the target model and define the same graph pattern

with the sub-graph by expressing the input model with the graph and executing the model transformation process based on the graph patterns. Also, it is possible to execute the model verification by investigating whether the graph model has such a pattern or not [4]. Since such a method is also based on the execution of the model verification through the simple graph comparison, it is impossible to verify various kinds of information contained by the test for uniformity in regard to the previously-mentioned model or the design model.

In this study, in order to improve the problems shown by the model verification mechanism through the comparison of different models in regard to the property matching based transformation similarity which is mainly used for the research of the model transformation verification and the verification process based on the existence of specific property by considering the converting verification based meta information and graph comparison algorithm. The information for the property of the related model is added to the previous graph model to define the converting information required for the model transformation verification on the graph model. It is possible to verify the converting information containing the information for the property of the model by using the verification metamodel containing the property of aspects, actions and times for the comparison of various property contained by the software design model in regard to the simple structural graph in the form of the version tree shown by the previous graph model consisting of nodes and edges.

In addition to, we chose two different transformation type, such as UML to RDBMS model and UML to RT-UML model. So, we compared property matching based transformation similarity about two model transformation type with existing C-SAW verification method.

2 Model Transformation Verification Using Transformation Similarity and Mapping Patterns

2.1 Model Transformation Verification Process

We proposed model transformation verification process for UML Design model, as shown in Fig. 1, first step is definition for a meta-model for source model which expressed in Feature based verification model and target model which expressed in UML Class model. And then we also need to define the transformation rule for feature based verification model as source model, we need to definition meta-model elements those are defined through source and target model elements. Third Step is graph-based model transformation to perform the conversion between source model and target model. Last step is model transformation verification by the results of model transformation which performed automatically, and then we also proposed the graph comparison algorithm with model property to perform model transformation verification between source model and target model.

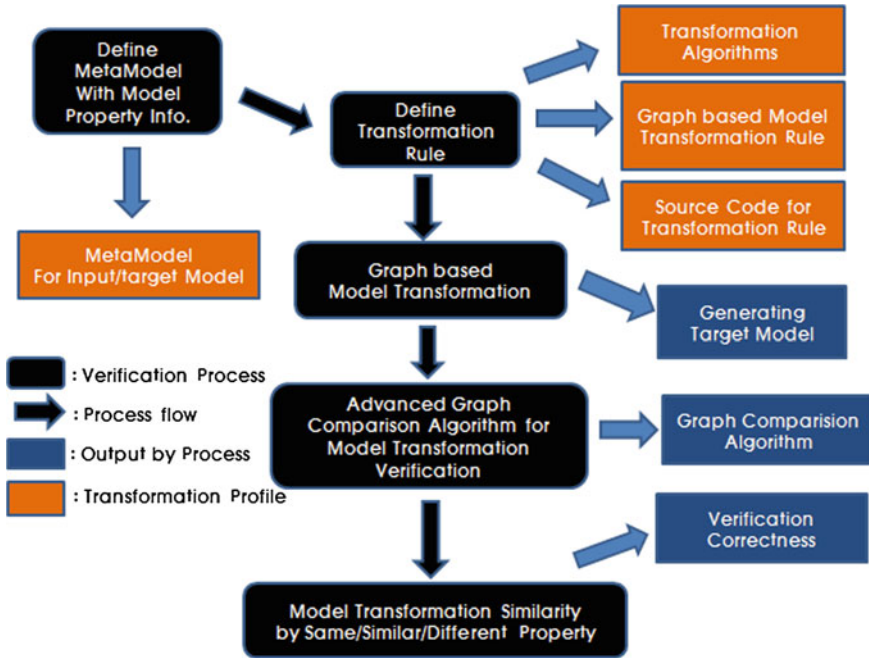


Fig. 1 Overview of model transformation verification process

And then we find same property, similar property, and different property in the mapping set and different set by graph comparison algorithm. And we also calculated transformation similarity using those information.

2.2 Meta Model Definition for Model Verification

In order to supplement existing model transformation verification researches in the model properties of the software static structural model and model General features a number of properties added information Feature based verification is defined meta-model elements. Model transformation verification research by C-SAW model transformation engine, used to a model transformation node and Edge was defined as a graph model. We proposed advanced graph model more node properties and edge properties such as the General Properties, Selection Properties, the Aspect Properties, Time Properties, that classified as you can define the parent node properties. Also the Edge between node and node, as well as the general properties (Edge_general) In addition to the edge node relationship properties to define the relationship between (Edge_relation) and metadata elements required to define classification properties (Edge_type) is described by adding the property

Table 1 Metamodel definition (node) for model verification

Verification based meta model	Sub property	Description	
General property	Node name	Node name of structure model	
	Node attribute	Node attribute of structure model	
	Node constraints	Node constraint of structure model	
	Classification	Relationship with parent node	
Selection property	Mandatory	Mandatory feature node	
	Optional	Optional feature node	
Time property	Capsule	Capsule node of RT-UML model	
	Port	Relation object between capsule of RT-UML model	
	Port constraints	Port Constraints of RT-UML model	
	Port behaviour	Port Behaviour of RT-UML model	
	WCET	Worst case execution time of node	
	Priority	Execution priority of node	
	Response time	Response time of node	
	Aspect property	Aspect name	Aspect name
		Joint point	Aspect joint point, method call
		Advice	Aspect advice, injection logic
Point cut		Aspect point cut	

Table 2 Metamodel definition (edge) for model verification

Edge property	Sub property	Description
Edge_general	Name	Name of edge
	Type	Type of edge
	Source	Node of source edge
	Desitination	Node of destination of edge
Edge type	Directional	Directional edge type
	Non-directional	Non-directional edge type
Edge_relation	Has_part	Generalization relationship between node and node
	Edge constraint	Constraints of Edge
	Part_of	Association relationship between node and node
	Dependency	Dependency relationship between node and node

information. More information about those node properties and edge properties, describe by Tables 1 and 2.

General Property has some sub properties such as node name, node attribute, node constraints, and classification, and Selection Property has mandatory feature node and optional feature node. Also Time Property has capsule, port, port constaint, port behavior, worst case execution time, priority, and response time. So we defined more model property information that to describe useful method about structural software model with some kind of information.

2.3 Model Verification Using Property Matching Based Transformation Similarity

The Same property is match property syntax and semantics such as convert a UML class model's class name to the name of the table in the RDBMS model, and the syntax of the attributes that characterize precisely the attributes and syntax and semantics in the same sense, because the accuracy of the mapping is done the same with each other.

In contrast, the Similar property is mismatch property syntax between the verification meta-model and class model and RDBMS model, or RT-UML model elements. But property semantics is match each other. The Different property is mismatch syntax and semantics between verification-based meta-model and class model and RDBMS model, or RT-UML model elements when transformed each other.

These properties are represented as the following:

Same Property

$$= (C \cap F), C_syntax = F_syntax, C_semantics = F_semantics$$

Similar Property

$$= (C \cap F), C_syntax \neq F_syntax, C_semantics = F_semantics$$

Different Property

$$= (C \cap F), C_syntax \neq F_syntax, C_semantics \neq F_semantics$$

We also define transformation similarity that compare same property, similar property and different property between input model and target model with meta information.

Transformation Similarity $M(TS)$

$$= \sum_{i=1}^n M(P_i) = \frac{\text{Num of } M(SP_i \dots SP_n) + \text{Num of } M(\text{Sim}P_i \dots \text{Sim}P_n) * 0.8}{\text{Num of } M(P_i \dots P_n)}$$

Transformation similarity calculated number of same properties and number of similar properties with weighted value in the both model. Table 3 shows property matching result about UML to RDBMS model and UML to RT-UML model.

Table 4 shows property matching transformation similarity between UML model and RDBMS model, and also between UML model to RT-UML model. As a result, transformation similarity value of RT-UML model that added time property in UML model greater than transformation similarity for RDBMS model transformation. Because RDBMS model transformation have more limitable properties and partial property matching.

Also, we compare transformation similarity result between model verification of existing research, C-SAW system and proposal method in Fig. 2.

Table 3 Property matching table for UML to RDBMS, RT-UML metamodel

	UML model	RDBMS model(exist method)	RDBMS model (proposal method)	RT-UML model (exist method)	RT-UML model (proposal method)
General property	Class	Table	Node type: table	Capsule	Capsule
	Attribute	Column	Node attribute: column	Port	Port
	Operation	x	x	x	Port behaviour
	Responsibility	x	x	x	x
	Constraints	x	Node constraints	x	Port constraints
	Stereotype	x	Node classification	x	x
	Relationship property	Association	Foreign key	Edge constraint: Fkey	Part of
Dependency		Foreign key	Edge constraint: Fkey	Has part	Dependency
Aggregation		Foreign key	Edge constraint: Fkey	Part of	Aggregation
Composition		Column	Node attribute: column	Part of	Composition
Generalization		Foreign key	Edge constraint: Fkey	Has part	Generalization

Table 4 Property matching based transformation similarity

Property matching based transformation similarity	Transformation similarity between UML model to RDBMS model				Transformation Similarity between UML model to RT UML model			
	Same property	Different property	Similar property	Ts value	Same property	Different property	Similar property	Ts value
	C-SAW verification	0.18182	0.36364	0.54545	0.54545	0.27273	0.36364	0.27273
Proposal method	0.36364	0.036364	0.18182	0.72727	0.63636	0.14545	0.18182	0.78182

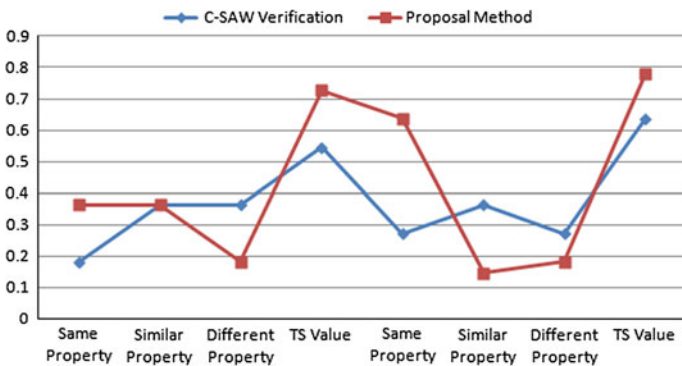


Fig. 2 Transformation similarity between C-SAW and proposal method

3 Conclusions

This paper proposes model transformation verification using verification meta information and transformation similarity by property matching. In addition, in order to support verification of the target model generated from the source model, we define verification meta model for UML model, RDBMS model and RT-UML model .

Moreover, if applying test issues at design stage as like that, there is an advantage to reduce modification cost of error comparing to test at the phase when source codes are almost completed. However, as perfect mapping of matching APIs for software application model transformation is actually difficult, it requires additional code complement works and it has a restriction that model transformation can be applied to only those domains suitable for a specific domain.

This paper has identified important challenges to make model transformations trustable. The first issue is to have well defined meta data that can use the source model and target model based on graph model, and the second issue is to make improve model transformation verification mechanism by other patterns those support mapping rule. In addition, it requires studies on scalability to apply transformation rules diversely and on improvement of comparison analysis between models, implementation of supporting tools to support it is also in progress. And we consider to apply model transformation verification that develop mobile application.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (No.: 2012-0003084).

References

1. Lin Y, Gray J (2007) A model transformation approach to automated model transformation, Ph.D thesis
2. Varro D (2003) Automated model transformation for the analysis of IT system, Ph.D thesis
3. Darabos A, Varro D (2006) Towards testing the implementation of graph transformation, GT-VMT'06
4. Csertan G, Varro D (2007) Visual automated transformations for formal verification and validation of UML Model, SAC'07
5. Czanecki K, Helsen S (2003) Classification of Model transformation approaches. In: OOPSLA'03 workshop on generative techniques in the context of model-driven architecture
6. Cabot J, Clariso R, Guerra E, de Lara J (2009) Verification and validation of declarative model-to model transformation through invariants. J Syst Softw (in press)
7. Zhao G, Kong J, Zhang K (2007) Design pattern evolution and verification using graph transformation. In: Proceedings of the 40th Hawaii international conference on system sciences
8. Varro G, Schurr A (2005) Benchmarking for graph transformation. In: Proceedings of the 2005 IEEE symposium on visual languages and human-centric computing
9. Varro D (2002) Automatic transformation of UML models, Budapest University of Technology and Economics
10. Varro D (2004) Towards formal verification of model transformations, Budapest University of Technology and Economics

Hierarchical Analysis of Steel House Material for 3D

Jung-Soo Han and Myeong-Ho Lee

Abstract This paper proposed a hierarchical architecture of steel house materials. The proposed architecture will include the detailed information of the components. For this, we made description about steel house design method and material change task ontology. Especially AFM is composited with total four steps. Using the method, the solution process of possible problems occurred in a steel house can be converted to an ontology. Also we made category about composite patterns of steel house materials.

Keywords Steel house · 3D · Material · AFM · Ontology

1 Introduction

This paper proposes composite pattern ontology for a steel house design allowing efficient architecture design in pattern unit with assembly of patterns of components composited with steel house architectural materials by adapting the steel house construction. The steel house construction is a representative dry method for easy assembly of wall panels made with panels and trusses. The steel house assembly construction technology is the prefabricated construction method

J.-S. Han (✉)

Division of Communication and Information, Baekseok University,
Cheonan, Chungcheongnam-do, South Korea
e-mail: jshan@bu.ac.kr

M.-H. Lee

Department of e-Commerce, Semyung University, Jecheon, Chungbuk-do,
South Korea
e-mail: mhlee@semyung.ac.kr

receiving attentions more than any other methods [1–3]. If architectural design contents are developed by combining this design method with IT, a house or a building can be designed easily not only by an architecture, but a user as well. Therefore, this study aims the automatic information creation of required patterns for material change and the changed architecture during a design by composing steel construction materials into patterns. Also, this study allows users design easily in owners' positions, and supports architectural design by providing not only cost estimation utilizing changed material information according to design changes but repository establishments based on pattern information composed with components.

2 Related Works

2.1 Steel House Design Method

The steel house construction technique using architecture design software extracts material components, builds a Meta data for each component and connects to Meta data of the composed patterns. At this time, the each component has a meaningful relation for pattern composition, and this is connected to the pattern's relation. The architecture design using patters allows new architecture design process with automatic re-composition of other related components according to the patterns. Also, to change an architecture built with assembly method, the patterns are supported according the selected the area to change as well as the needful information for re-building to designers through required material related material and pattern information analysis. In general, since an architecture is composited with several hundreds of parts, it is impossible for a designer to work to assembly on screen by selecting each part name separately. Therefore, for an efficient performance of the process, pattern type assembly building architecture method is proposed [4].

2.2 Material Change Task Ontology

The technology is for users'convenience in building steel houses by using AFM (Activity-First Method) methodology. AFM is task ontology. The task ontology is a definition of concept structure with the world targeted problem solving process, and its center concept is process and process target. Here, the subject of process is the subject of problem solving. The task ontology is a concept structure to describe various problems solving process in actual existence with using concepts of high generality undetected in the target area. It decompose a problem solving steps to principal action concepts, and simultaneously extracts and organizes limits to be

Table 1 AFM ontology steps

Step	Description
Step 1	<ul style="list-style-type: none"> • Divides the assigned work into blocks of which each block has exactly one step of the series of work in solving one problem • Extracts task units from each block • Establishes a flowchart connecting the extracted task units
Step 2	<ul style="list-style-type: none"> • Composes a task activity using concept representing the intrinsic action/operation • Organize a hierarchy with is-a relation
Step 3	<ul style="list-style-type: none"> • Acquires a typical task flow by generalizing the task flow • Describes the object task flow focusing on input and output flow of task activities and interconnection relations
Step 4	<ul style="list-style-type: none"> • Extracts concepts by abstraction and separates the role concept depended on the concept • Organize the extracted concept to the is-a hierarchy • Interprets the ontology acquired with the ontology editor

satisfied with the steps and required concepts to express the steps. As shown in Table 1, the task ontology is in composition of total four items including role, action, status and other limit. AFM is composited with total four steps. Using the method, the solution process of possible problems occurred in a steel house can be converted to an otology. For example, a process required to extend the roof or to increase the number of rooms in a design can be converted to an otology through the AFM method [5].

3 Customized Composite Pattern

This is the technology enables users design more easily to connect components to patterns. Since existing patterns depend on only simple inputs, the customized pattern composition is needed to connect patterns with other links after selecting the patterns. After the establishment, an architectural design becomes easier because the architectural design system indicates what patterns is connected as the next eliminating the need of the designer’s own selection for the pattern. In building walls and a room, the room is made using the walls so that the room uses two sides of the walls, only two sides are required. If up to this point is the existing pattern, the technology allows automatic inputs by storing converting doors, windows and areas required for the room, a component created by using the customized composite pattern technology. The pattern for each composition is shown in Table 2 [6].

Table 2 Steel house pattern

Pattern	
Foundation	<ul style="list-style-type: none"> • Outputs the materials for the foundation type (ex: when inputting a foundation without a basement <ul style="list-style-type: none"> – Stud, anchor bolt, insulation material, reinforcement, building paper, floor slab, track, sand bed, foundation wall, foundation plates are composited.) • Inputs the dimensions desired by the user (creates the foundation firs to the dimension)
Beams And Columns	<ul style="list-style-type: none"> • Beams and columns composition material and information output <ul style="list-style-type: none"> – Material information: web stiffener, beam joist, stud, track, shape steel, angle • Installation method selection—drop beam type, flush beam type
Wall Framework	<ul style="list-style-type: none"> • Common/basic material output and measurements input <ul style="list-style-type: none"> – Materials : stud, track, header, wall bracing, sheathing material – Measurements: stud flange width, hole width, length, gap • Selecting wall types according to weight <ul style="list-style-type: none"> – Bearing wall selection: supports vertical weight – Non-bearing wall selection: no weight support – Shear wall selection: supports parallel weight caused by wind loads • Wall's location selection <ul style="list-style-type: none"> – Interior/exterior wall selection: determined by stud's width • opening and header information output <ul style="list-style-type: none"> – Header's composition: track + king stud + screw + jack stud + structure shear panel + cripple stud
Floor Framework	<ul style="list-style-type: none"> • Floor framework composition material and information output <ul style="list-style-type: none"> – Material information: joist, rim track, joist bracing, structure substrate, ondol ground class, ondol • For floor type selection <ul style="list-style-type: none"> – Dry floor: dry ondol + base panel + structure plywood + joist + ceiling – wet floor1: wet ondol + aerated light weight concrete + deck plate/ structure plywood + joist + ceiling – Wet floor2: wet ondol + aerated light weight concrete + deck plate/ structure plywood + joist + ceiling – Fire resistant floor: wet ondol + aerated light weight concrete + deck plate + joist + ceiling • floor opening and stair information output <ul style="list-style-type: none"> – Floor opening: joist + header joist + trimmer joist + clip angle – Stair: joist hanger + stair framework + column + hanger + stair panel + stair side assembly girder
Roof Frame work	<ul style="list-style-type: none"> • Roof framework composition material and information output <ul style="list-style-type: none"> – Material information: rafter, howe truss, gable end, loft framework, surface material, jack truss • Roof characteristics and information output <ul style="list-style-type: none"> – Roof slope: 4/10 degree in general – Roof eaves: eaves type according to eaves ceil's availability – Roof shape: flat roof, pent house roof, Netherlands style roof, gemrel roof, hip gable roof, mansard roof • Roof framework selection <ul style="list-style-type: none"> – Rafter type: framework composition with single material – Truss type: truss composition with stud material • On roof selection, information output on the roof selection desired by the user

4 Material to Weight

Steel house material's text is separated to blocks in composition of several sentences. To be more specific, the separation is made one block includes exactly one level of a series of works in one construction material. From each block, task units are extracted. Task unit refers material + subject structure composited by exactly one material and its subject. After task unit extraction, establish a flow-chart. And, this process is called 'refinement task flow procedure'. And, later, perform the organization of task activities. The process eliminates the variety of terms by conceptualizing materials presented in task units, converts the task units using the concept representing original material information and composes the task activities. Then, organize the task activities in the 'is-a' class structure. Extract and conceptualize the role of the task activities' input and output. And, call this procedure, 'task activity role'. Now, analysis of task structure is required. By generalizing the refinement task flow, acquire the general task flow for construction material. Describe objective flow with focusing on the task activity's input and output flow and the interconnection relation of the general task flow. Task context role means the assigned role in the entire focused task.joist, rim track and joist bracing would be the examples for foot framework. Extract the domain material terms which actually are responsible for the task context role. In final, organization of construction material domain concept is required. Extract the concept by abstracting the domain terms for the construction material and separate the concept of role which depending on the domain concept clearly. Also, organize the extracted domain concept with is-a classes.

5 Conclusions

This paper aims to effective management of information regarding buildings, to efficient use of data and to develop an assembly architecture design system in the steel house assembly architecture design using composite patterns. First, to figure out relations of patterns regarding buildings, necessary materials are categorized through the Taxonomy based on objective properties. Then, Meta data are defined by extracting components regarding buildings and using upper Ontology. Through the pattern type assembly architecture design system, users become capable to assemble their desired constructions more easily.

The final goal is the development of an assembly architecture design system for steel house architecture design based on the composite patterns. Therefore, the complex relationships will be easily identified, searched and designed. Also, reusable parts and newly required materials of each component will be easily differentiated while automatic conversion for similar architecture will be possible. And, as the result, a 3D based assembly architecture design system will be developed.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0003084).

References

1. Kozaczynski W, Booch G (1998) Component based software engineering. *IEEE Softw* 15(5):34–36
2. Gamme E, Helm R, Johnson R, Vlissides J (1995) Design pattern: elements of reusable object-oriented software. Addison-Wesley, Boston
3. Park HS (2010) A study on module design of steel house structure wall. Master's thesis, KunKook University
4. Jung HS, Choi K (1998) An introduction to the construction of low-rise steel-framed housing in Korea, vol 12, no 1. RIST Pohang University, Korea
5. Jung HS, Lee PG (1999) Construction optimization through use of prefabricated light-gauge steel panels, vol 13, no 4. RIST Pohang University, Korea
6. Kim G-J, Han J-S (2012) Application method of task ontology technology for recommendation of automobile parts. *J Digit Policy Manag* 10(6):275–282

Multi-Faces Recognition Process

Jung-Soo Han and Jeong-Heon Lee

Abstract In this paper, we explained multi-faces recognition process. To construct a security framework of a CCTV-based face detection and recognition system, we suggested modeling method applied as 4 steps. For CCTV image recognition, we explained image recognition system composition, image recognition core module and face recognition system process. Especially, Face Recognition System is a system that extracts face from input images, and recognizes who is whom through similarity assessment process and characteristic data registered by extracting characteristic data for recognition in extracted face.

Keywords CCTV · Face recognition · Security · Image recognition

1 Introduction

To construct a security framework for safe CCTV-based face detection and recognition system that can be widely used the following three tasks are prerequisite to resolve. First, through development a framework for personal information protection in the image surveillance system using CCTV, system operation and technology plans are required to be established for personal privacy protection. Second, context-aware and privacy protection software is required to be

J.-S. Han (✉)

Division of Communication and Information, Baekseok University,
Cheonan-city, Chungcheongnam-do, South Korea
e-mail: jshan@bu.ac.kr

J.-H. Lee

National Information Society Agency, Seoul 641-773, South Korea
e-mail: opendori@gmail.com

developed for CCTV through suggesting context-aware and privacy protection plans for high performance face recognition technology by using CCTV. Third, through standardization development for personal information protection in an image surveillance system using CCTV, required is a standardized security framework of face detection and recognition system based on CCTV that are widely used through preparing standardized plans on personal information protection for CCTVs in ISO/IEC SC27 WG5 and extracting ISO/IEC JTC1 SC27 [1–3].

Resolving the problems will not only lead to countermeasures of privacy infringement resulting from positive application of the next generation biometric system through expanded support of the next generation biometric system, and international and national standardization, but also promote reinforced competitiveness related to international standardized technology such as ISO standardization by raising biometric providers' technology level, standardized technology development and revitalized domestic industry through national standardized security framework of CCTV-based face detection and recognition system, and prior occupation of the market of CCTV-based face detection and recognition system by securing source technology [4].

2 Modeling and Standard

To construct a security framework of a CCTV-based face detection and recognition system, the following modeling method can be applied as 4 steps. For first step, conducted is a security framework modeling of face detection and recognition system that can be widely applicable to a CCTV environment where privacy can be protected at maximum and criminals can be quickly arrested not to cause any more victims. To achieve this, through analysis on face detection methods including a knowledge-based method, characteristic-based method, template matching method, and exterior-based method, a framework for face detection system security is made to be suitable for a CCTV environment. Further, through analysis on face recognition methods including a sex distinction method using facial color, a telebio recognition method, a 2D/3D face recognition method, an IR-based face recognition method, a framework for face recognition system security is made to be suitable for a CCTV environment [5].

For second step, on important service (e.g. client ID management, criminal arrest, terror prevention, access management, etc.) using bio information based on the 1 step modeling, software is developed, which can be demonstrated. For third step, by extracting requirements for personal information protection in a image surveillance system such as CCTV, established is operation and acquisition plans for image information, considering personal information protection. Finally, for fourth step, based on a standardized plan for personal information protection of image surveillance system, conducted is operation on a new standardized plan and standardization on personal information protection based on CCTV environment.

Therefore, to construct a framework of CCTV-based face detection and recognition system security, a well-prepared applicable modeling and a standard plan will not only correspond to personal information infringement, resulting from positive application of biometric reference system in a CCTV system environment, but also contribute to reinforced competitiveness at home and abroad through prior domination of pioneering technology for international standardization such as ISO standardization on personal privacy protection plans in a CCTV system.

3 CCTV Image Recognition System

3.1 Image Recognition System Composition

Unmanned security surveillance robots can recognize authorizer/unauthorizer by real-time processing images input by HD-level camera, save a relevant image to transmit via control server, store transmitted image according to a policy, which enables a controller to provide a preview function after searching saved images through a search function. Unmanned security surveillance robot's image recognition system has the following composition elements.

It serves as a function to detect a face in a HD-level image, extract and record face characteristic data to compare with authorizer. With regular hour (daytime) fixed mode, it serves as an alert surveillance (HD-level CCTV function), and automatically changes into move mode in set time (e.g. night time, after school, holidays, etc.), alert surveillance is performed by moving in a set place and a surveillance poor site. It can be recharged in a fixed mode, and saved images can be transmitted to the control center in a move mode when an alert surveillance. In a move mode, when an unauthorizer is found, a warning message is transmitted to the unauthorizer by speaker warning and lamp alarm. Saved images during an alert surveillance are saved in the inside of the robot, and when it is recharged, they are transmitted to the control center to save. To compare information between authorizer/unauthorizer, a captured frame (or image) in a video clip should be solely extracted and transmitted to the center to be compare with DB. In case of lights-out in night time alert, detect a moving object by sensing with a built-in moving detecting sensor, and when detecting something, a lamp of the robot is lighted on to save a relevant image [6].

3.2 Image Recognition Core Module

It is a module composed of processors to obtain a pause image from images input in HD-level IP camera, and image recognition. A Robot_Agent function controls robot and camera by receiving setting parameters resulting from an integrated

control server policy. Further, robot motion is checked and managed real-time. This performs by receiving a setting parameter value from the integrated control server, which enables Robot_Agent to consistently perform during the robot performing a motion to communicate with an integrated control server, perform FTP_Uploader by order from the integrated control, and check a robot motion status by checking RTSP_Recorder. If RTSP_Recorder performance is complete, it should be performed again. Its implementation is interlocked with the integrated control server system to receive necessary data for robot operation.

IMG_Analyser performs a recognition process by analyzing a pause image captured by RTSP_Recorder. It is composed of a face detection module and a characteristic extraction module. Face recognition is much influenced by various environmental factors such as minute change in facial angle, brightness, and expression or images' complicated background, which contributes to in-depth research to overcome limitation factors such as whose face is his face, and what facial expression belongs to whom in an extracted condition of face, compared to recognition of other parts. Recently, face recognition technology has developed by transforming 2-dimensional face recognition into 3-dimensional face recognition, or front face recognition into non-front face (e.g. slanting face or sided face) recognition, and various face recognition techniques and systematic assessment methods are suggested.

Face recognition can be broadly classified into the following two aspects: first, how to express a face and second, how to classify the expressed face. Out of these, facial expression plays a key role in face recognition. The following methods are broadly used for research on face recognition: geometric registration that recognizes face by using a distance between locations, sizes of eyes, nose, mouth, etc., template matching that recognizes face by analyzing correlation by comparing face data with template images saved in database systems, a method using Artificial Neural Network (ANN), a method of Support Vector Machine (SVM), and a method of Hidden Markov Model (HMM) [4].

3.3 Face Recognition System Process

Face Recognition System is a system that extracts face from input images, and recognizes who is whom through similarity assessment process and characteristic data registered by extracting characteristic data for recognition in extracted face. To achieve this, it is composed of face detection module, characteristic extraction module and similarity assessment engine, a recognition module. An assessment module for face similarity is used as a PCA-specific extractor input based on Gabor-kernel for facial range image extracted through Adaboost learning algorithm. A face recognition system calculates an Euclidean distance between authorizer and unauthorizer registered and factors of characteristic data extracted, which maps face in the closest distance. It is a module that calculates similarity

between registered faces, recognizing who is out of authorizers through face recognition, and it serves as a function of grouping as unauthorizers, otherwise.

4 Conclusions

In conclusion, this paper was conducted for a method to protect personal privacy at maximum at a CCTV environment following the trend of international and domestic standardization, and a framework modeling method that analyzes face detection and recognition system security widely applicable, which finds an implementation method to develop a standard model that can be demonstrated on important service (client ID management, criminal arrest, terror prevention and access management, etc.) based on information protection and biometric reference standard development, and delineates standardized plans for personal information protection on requirements for related bio information and reference management server, mechanism, secured system of database, and management system, image surveillance system and CCTV-based image surveillance system.

To achieve this, by manufacturing an unmanned security surveillance robot, images input through HD-level camera are processed real-time to identify authorizer/unauthorizer, save a relevant image to transmit to the control server, and keep transmitted recorded images according to a policy, which offers a person in charge to preview them after saved image search through search function. It is a module to upload images real-time to the integrated control system by saving transmitted module and image real-time that enables monitoring at the integrated control system application. The face recognition system is a system that detects face from input images, and recognizes who is whom through similarity assessment process between characteristic data registered by extracting characteristic data to recognize extracted face.

Acknowledgments This work was supported by the Security Engineering Research Center, granted by the Korea Ministry of Knowledge Economy.

References

1. Cootes T, Taylor C, Cooper D, Graham J (1992) Training models of shape from sets of examples. In: Proceeding international conference British machine vision, pp 9–18
2. Spagnolo P, Caroppo A, Leo M, Martiriggiano T, D’Orazio T (2006) An abandoned/removed objects detection algorithm and its evaluation on PETS datasets. In: Proceeding international conference video, signal based surveillance, pp 17–21
3. Yu and E, Aggarwal J (2009) Human action recognition with extremities as semantic posture representation. In: Proceeding international conference computer vision, pattern recognition, pp 1–8
4. TI (2008) TMS320C64x + DSP image/video processing library programmer’s guide. Texas instruments, Dallas, Texas

5. Stanley TB, Sriram R (2005) Spatiograms versus histograms for region- based tracking. CVPR 2:1158–1163
6. Nummiaro K, Koller-Meier E, Van Gool L (2002) A color-based particle filter. In: European conference on computer vision, pp 53–60

Software Performance Test Automation by Using the Virtualization

Gwang-Hun Kim, Yeon-Gyun Kim and Seok-Kyu Shin

Abstract In this paper, we propose a method on software performance test automation by using the virtualization. In general, most test engineers use the public performance testwares such as Load Runner and Silk Performer to validate the performance efficiency of their own systems. In case that they cannot use the performance testwares due to some technical limitations in the testwares, the testers should perform the testing in manually. As waste of computer and human resources is resulted from the situation, we need to propose the test automation scheme by using the virtualization technology to prevent the dissipation in the test environment which has limited resources. The system architecture considered efficient usage of computer resources and test automation to reduce human acts are addressed mainly in this paper. Finally, a number of experiments show that the proposed schemes allow offering the possibility for automated software performance testing by using the virtualization.

Keywords Software testing · Software performance engineering · Performance testing · Test automation · Virtualization

G.-H. Kim (✉) · S.-K. Shin

Software Quality Evaluation Center (SQEC), Telecommunications Technology Association (TTA), 267-2 Seohyeon-dong, Bundang-gu, Seongnam, Gyonggi, South Korea
e-mail: nuly17@tta.or.kr

S.-K. Shin

e-mail: skshin@tta.or.kr

Y.-G. Kim

The 7th R&D Institute-3, Agency for Defense Development (ADD), Yuseong, 35-7Daejeon, South Chungcheong, South Korea
e-mail: yg_kim@add.re.kr

1 Introduction

As IT industry grows up dramatically, people not only consider that software should be operated normally, but also software should have sufficient performance. Therefore, people focus on software performance test because it is possible to accomplish company's reliability through high performance service and help us prevent from unexpected performance degradation and system failures. A recent survey about performance [1] found that half of software companies had encountered performance problems with at least 20 % of the application they deployed. Especially, the performance of web application software that processes a number of user's requests has high priority and performance test for web application should be carried out sufficiently.

The importance of performance test results in existence of a lot of performance testwares on the market such as Load Runner [2] and Silk Performer [3]. The performance testwares emulate a number of clients, which are virtual users, to generate load to server components. With the performance of those testwares, we can save computer and human resources because the testwares generate sufficient virtual users and it is automated.

However, the testwares cannot cover some scenarios. For example, we cannot use the testware in case that server actively sends information to a number of clients. Also, we cannot use the testware in case that it is not able to analyze communication protocol between servers and clients. Therefore, in these situations, we need a number of human and computer resources to perform software performance test. A number of computers should execute client applications and a number of people should operate client applications on the computer to measure system performance.

To solve the problems, we propose a method on software performance test automation by using virtualization technology to prevent the waste of computer and human resources. Proposed method can reduce computer resources because virtualization technology can make a number of virtual computers with small number of physical computers. Also, the proposed method can reduce human resources due to the management computer to control a number of virtual computers automatically. A number of experiments have been conducted in order to verify the efficiency of the proposed methods. Our results show that the proposed schemes allow offering the possibility for automated software performance testing by using the virtualization.

2 Related Works

SPE represents the entire collection of activities and related analysis to meet software's performance requirements [4]. Any SPE process includes some or all of the following activities.

First one is to identify and define requirement. We identify qualitative performance factors of target software affecting performance goal [5]. And then we define detail performance factors like workload intensities, delay and throughput requirement, and scenario describing software behavior [6]. Second one is to predict performance. We estimate measured value of performance factors with three kinds of method like simulation-based prediction, profile-based prediction, and modeling-based prediction [7, 8]. Third one is performance testing. We make load on part or all of system and measure resources of target system and performance factors. There are two categories in performance testing like load generating, for supplying the workload to a system under test, and monitor, for gathering data as the system executes. Last one is total system analysis. After making report about software performance, we analysis the report and compare predicted performance to actual performance.

Among these SPE process, we focus on load generating of performance testing. There are many load generating testwares such as Load Runner [2] and Silk Performer [3]. However, the testwares cannot cover every testing scenario. For example, 1) In server/client environment, client sends information to server generally. However, server may send information to client. Existing testwares cannot support that scenario. 2) If testware does not understand communication protocol, it cannot generate load. 3) If software uses new technique that testware does not support, it cannot generate load. In those situations, we cannot use existing load generating testware and we use many real computers and people operate software on that computer. Therefore, we want to reduce computer resource and human resource using virtualization technology.

Virtualization is the creation of a virtual version of something such as a hardware platform, operating system, a storage device or network resources [9].

Virtual version is logically created from physical device. A single physical device can present many individual logical devices and many physical devices can appear to function as a single logical unit. Virtualization is divided into several categories as purpose such as server virtualization, network virtualization, desktop virtualization, storage virtualization, application virtualization and so on.

Virtualization has many advantages and the biggest advantage of virtualization is that it can increase the efficiency. For example, most servers with no virtualization are in use only 5–15 % of the time they are powered on [10]. However, Server Virtualization which has multiple virtual servers on one physical server can increase hardware utilization. Therefore, it can reduce server hardware cost and power consumption.

There are a lot of virtualization products such as VMWARE [11], Virtual PC [12], Parallels [13], Virtual Box [14] and so on. We select VMWARE product to construct virtualized performance test environment because it is the most frequently used product, it uses computing resource efficiently, and it has high compatibility with various computer hardwares and operating systems.

3 Proposed Approach

3.1 System Architecture

The system architecture is illustrated in Fig. 1. Firstly, we will describe specifications for the used components of our system and approaches to the performance measurement. And then we will describe how to make the virtual machines on physical machines.

As shown in Fig. 1, Performance Test Target Server (PTTS) with server applications installed on is the main server for performance testing and makes a response to the request of their clients. As a role of giving a load to PTTS, the virtual machine operates client applications. We also need a number of virtual machines for appropriate loads on the server application. The manager is a controller of the virtual machines. The test automation performed between the manager and agent will be explained by showing how the manager works with the virtual machines in the Sect. 3.2. If the manager makes a command to start their actions for each virtual machine simultaneously, the virtual machines send their

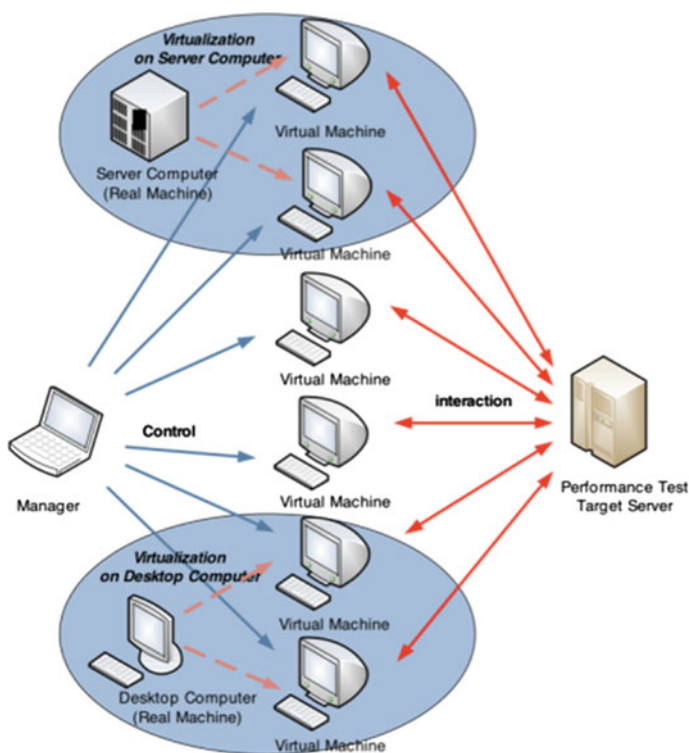


Fig. 1 System architecture

request to PTTS. And then a tester measures the resource usage on PTTS with some performance measurement tools such as Team Quest [15] and Perfmon [16]. The CPU, memory, disk I/O, network I/O are main units as measured resources for performance testing in this paper.

This research does focus on using the virtualization technology to reduce computer resource consumption for performance testing. The virtualization tool like the VMware can help us to reduce the number of physical computers for performance testing because it can generate a number of virtual machines on a physical machine. As illustrated in Fig. 1, we suggest two ways to generate virtual machines. Generating virtual machines on desktop computer will be dealt with and then on the server computer. Various physical machines can be selected to operate virtual machines with respect to performance test environments. According to the test environment which includes only desktops or servers, each physical machine makes virtual machines properly for them. We can generate virtual machines on desktops and servers and run the performance testing with no regards to what the physical machines are in test environments because all of the physical machines are available for the testing. As virtual machines are appropriately used for variable test situations, we can use the computer resources effectively for the performance testing. In general, servers can operate more virtual machines than desktops due to the relative superiority of computer resources. The issues on what is the resource difference between desktops and servers will be dealt with in Sect. 4.

3.2 Test Automation

The system architecture using the virtualization can help us to reduce computer resource consumption as mentioned in previous section. However, we still have a problem considering the waste of human resources in order to control each virtual machine for load generation to servers, even though computer resources can be reduced by the virtualization. Therefore, we propose the test automation to reduce human resources.

Figure 2 shows the use case diagram between a tester and a test scenario recorder. The test scenario recorder which is sub-function of manager application has a function to record keyboard and mouse events such as ‘move mouse to (point x, point y)’, ‘click mouse left button’, and ‘push enter key’ in order to control client application. The tester clicks ‘start button’ on the Test Scenario Recorder and operates client applications to record the test scenario. After the tester finishes operating client applications, the tester clicks ‘stop button’ and then test script file which contains test scenario with XML format is generated as a result of the Test Scenario Recorder. Test script files will be dispatched to each agent when the tester executes performance testing.

Figure 3 shows the relation between the manager and agent. The agent is installed on the virtual machine to control the virtual machine automatically. The

Fig. 2 Use case of tester and test scenario recorder

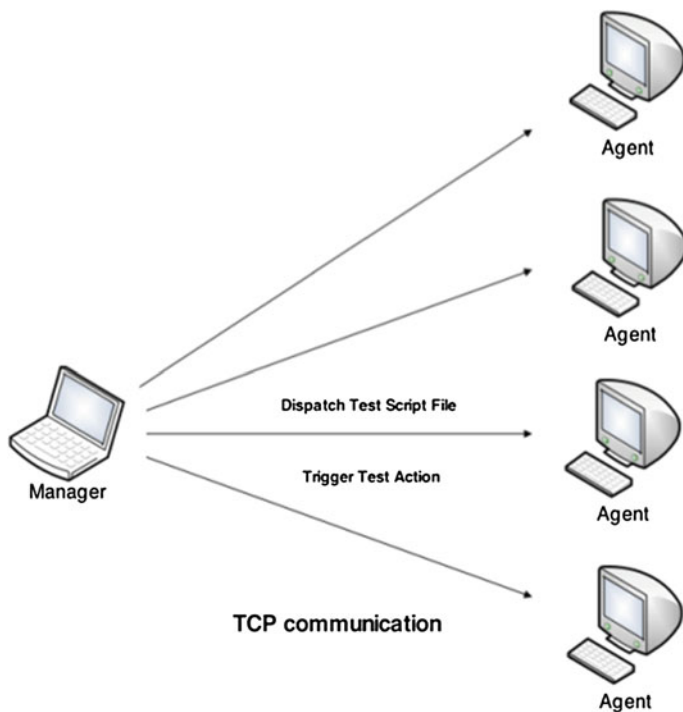
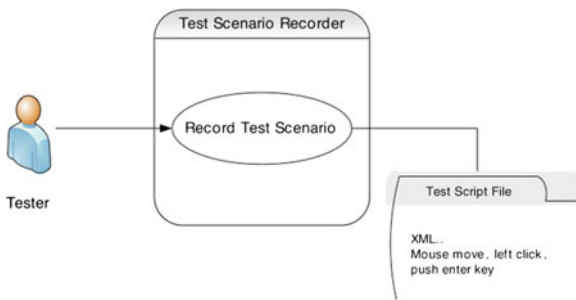


Fig. 3 Relation between manager and agent

manager is operated by performance test engineers to centrally manage and control virtual machine. The tester installs the agent to each virtual machine and registers all virtual machines' IP addresses with IP address transmission function of agent. If PTTS is ready to measure performance, the tester dispatches the test scenario script which is already recorded to the registered agent and commands the agent execute the test scenario script. Then each agent executes the test scenario script and it imposes a load on PTTS.

Table 1 The performance test metric

Metric	Unit	Description
Average response time	s	The average response time to complete the download in the client
Average bitrate	MB/s	The average bitrate for download

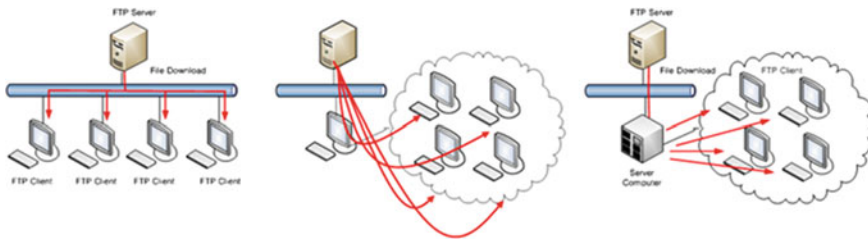


Fig. 4 Three test environment of physical machine, desktop machine, server machine

If the agent on the virtual machine is installed, the virtual machine can be controlled by the manager. By establishing test automation, we can reduce human resources to operate computer which installs client application. In the next section, we show a number of experiments to verify the efficiency of the proposed methods.

4 Experiment

In this section, the comparison of the performance test results between the physical and the virtual machine when running the same testing on each machine. The performance test scenario, metrics, and test environment are as follows.

The performance test scenario—the FTP client tried to download the 100 MB file from the FTP server (the number of the concurrent download clients: 1 person, 2 people, 4 people, 8 people, 16 people).

The metrics shown in Table 1 were measured with the performance test scenario for the virtual machines on the desktop and server. The performance test environment for the physical and virtual machines are as follows.

As shown in Fig. 4, the physical machines connected to the network download some files from the FTP server with the FTP protocol. The virtual machines on the desktop computer and server computer are illustrated in the middle and right of Fig. 4. The desktop computer and server computer operating the virtualization manages several virtual machines, which download the files from the FTP server.

Figures 5 and 6 shows that the bitrate and response time are changed by increase in the number of clients. The bitrate results were measured independently on the physical machine, desktop virtual machine, and server virtual machine. In the case of the desktop virtual machine, 6 desktop computers should be prepared to

Fig. 5 Bitrate comparison between the physical machine and virtual machines according to the number of client

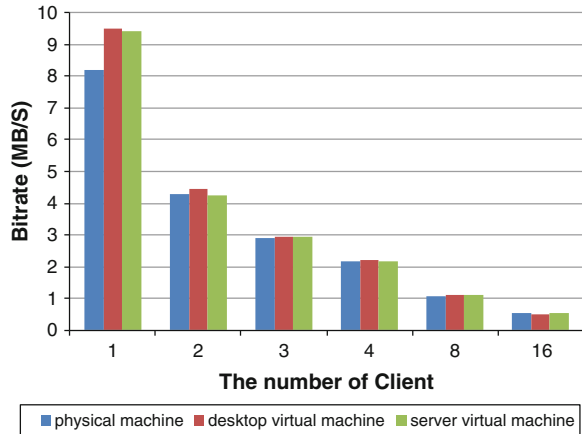
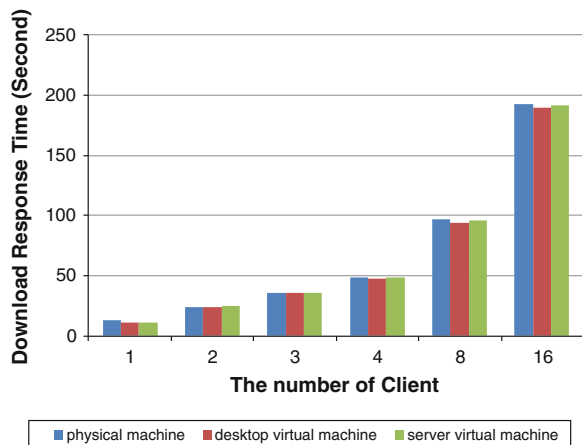


Fig. 6 Download response time comparison between the physical machine and virtual machines according to the number of client



operate 16 virtual machines because the number of the virtual machines which can be installed on one desktop is 3. Figure 5 illustrates that the bitrate was decreased linearly as increased in the number of clients, and the results were similar to the physical machine and the virtual machines in the desktop or server. Figure 6 shows that the response time was increased dramatically for more clients and the results were approximately same to the physical machine and the virtual machines in the desktop or server.

5 Conclusion

We proposed the method on software performance test automation by using the virtualization in order to overcome the technical limitations in the commercial load

generation testwares. We used the virtualization technology and developed the test automation scheme to reduce the waste of computer physical and human resources. The virtual machines were laid out in order that the desktop and server virtualization technology can be used compatibly for the test environment.

According to the experimental results, the amount of the virtual machines which can be made by the physical machine was highly depended on the memory use rather than the other factors. The desktop which had 2 GB memory was enabling to generate about three virtual machines, and the server which had 32 GB memory was possible to install about 120 virtual machines. The comparisons of the changes about the resource usage commensurate with increase in clients from the test scenario show that the experimental results were significantly similar with each machine such as physical machine, desktop virtual machine, and server virtual machine.

References

1. Compuware (2006) Applied performance management survey, Oct 2006
2. https://h10078.www1.hp.com/cda/hpms/display/main/hpms_content.jsp?zn=bto&cp=1-11-126-17%5E8_4000_100__
3. <http://www.borland.com/us/products/silk/silkperformer/index.html>
4. Woodside M, Franks G, Petriu D (2007) The future of software performance engineering. In: SOSE'07, pp 171–187
5. Chung L, Nixon BA, Yu E, Mylopoulos J (2000) Non-functional requirements in software engineering. Kluwer, Dordrecht
6. Barber S (2004) User community modeling language for performance test workloads. <http://www-128.ibm.com/developerworks/rational/library/5219.html>
7. Zheng G, Wilmarth T, Jagadishprasad P, Kalé LV (2005) Simulation-based performance prediction for large parallel machines. *Int J Parallel Prog* 33:2–3
8. Balsamo S, DiMarco A, Inverardi P, Simeoni M (2004) Model-based performance prediction in software development. *IEEE Trans Softw Eng* 30:295–310
9. Turban E, King D, Lee J, Viehland D (2008) Building E-commerce applications and infrastructure. *Electronic commerce a managerial perspective*, 5th edn. Prentice-Hall, New Jersey, p 27
10. Lee H (2008) Server virtualization overview and related solution areas. *Commun Korean Inst Inf Sci Eng* 26(10):5–13
11. <http://www.vmware.com/>
12. <http://www.microsoft.com/Windows/products/winfamily/virtualpc/default.aspx>
13. <http://www.parallels.com/>
14. <http://www.virtualbox.org/>
15. <http://www.teamquest.com/>
16. http://www.microsoft.com/resources/documentation/windows/xp/all/proddocs/en-us/nt_command_perfmon.aspx?mfr=true

Requirement Analysis for Aspect-Oriented System Development

Seung-Hyung Lee and Hyun Yoo

Abstract Recently the software system is becoming little by little complexity termination. Aspect-oriented Programming is support a crosscutting concern. Crosscutting concerns are responsible for producing scattered and tangled representations those are difficult to maintain and evolve. Aspect-Oriented Software Development aims at providing means to identify, modularize, specify and compose crosscutting concerns. Requirements engineering techniques that explicitly recognize the importance of clearly identifying and treating crosscutting concerns are called Aspect-oriented Requirements Engineering Approaches. Aspect-oriented requirements engineering approaches improve existing requirements engineering approaches through an explicit representation and modularization of concerns that were otherwise spread throughout other requirements. Aspect-oriented requirements engineering approaches adopt the principle of separation of concerns at the analysis phase. This approach provides a representation of crosscutting concerns in requirements artifacts.

Keywords System requirement analysis · Aspect-oriented system development

S.-H. Lee (✉)

Department of Computer Engineering, Kyung Hee University, 1732 Deokyoungdaero, Giheung-gu, Yongin-si, Gyeonggi-do, South Korea
e-mail: shlee7@khu.ac.kr

H. Yoo

Department of Internet Information, Osan University, 45 Cheonghakro, Osan-si, Gyeonggi-do, South Korea
e-mail: hyoo@osan.ac.kr

1 Introduction

The recent software system is becoming more complex, and the complex software is interrupting the software development paradigm. Object-oriented programming has difficulties with modularization of the crosscutting concerns such as persistency, distribution, exception handling, and security, since it does not provide crosscutting concern fully enough. Also, tangle and dispersion occur in object-oriented development engineering because of the repeated code. [8, 4].

Aspect-oriented development engineering supports crosscutting concern during the lifecycle of software development. Aspect is a development engineering which can be used for the developers in system to apply modularization, analysis, and trace for the crosscutting concern. The advantages of aspect-oriented development engineering are the improved understanding of complex system, lessening of intricacy, reuse, customizing, easiness of testing, and improved maintenance of final system.

At the moment, aspect-oriented system development is based on the idea of object-oriented programming; however, aspect-oriented system development is insufficient. Aspect-oriented system development is for the whole development stages such as requirement specification and architecture design. [5, 6] This thesis suggests a systematic method which is used for aspect crosscutting concern in the programming stage, and deals with how crosscutting concern can be understood from requirement step. This development method can help the clear understanding of system requirement and can be applied to system development by the quick replacement of the aspect's component modules that are used in other existing system. [7].

2 Requirement Engineering of Aspect-Oriented Programming

Separation and encapsulation of crosscutting concern is not enough for object-oriented development programming. Since the application of crosscutting concern through system development lifecycle is difficult, the method for the separation of the crosscutting behavior from the design and code, and the encapsulation is provided. To attain this object, it is necessary to perform aspect-oriented system design including the independence of performing language, assembly of design level, and the assembly utilizing the existing design. There are some difference between requirement and design, design and code. Aspect-oriented programming is expanded based on modeling during the aspect weaving to ultimate implementation model in the designing process.

The expansion from all of the stage of system development cycle can be done by using specified aspect. Design and implementation aspect is declared in the designing stage since it can be traced from the requirement through the source code as design language to provide aspect-oriented design environment. [7, 9].

High standard design model is required for complex structure and software system. This model should clearly specify the principle and guideline for the system structure. Practically, the developer tends to depend on the documented design model and guideline of the system which is implemented to be separated from the model. This method can discourage the understanding of the whole system and because of this; using aspect programming is the object to enhance the quality and maintenance while minimizing cost. This principle cannot be limited to a single module. The method of crosscutting architecture should be found in the system. [2, 3].

3 Importance of Requirements in Aspect-Oriented System Development

The identifying of what maps and influences the requirement level aspect enables the tracing of requirement and constraint that are formed through development, maintenance, and expansion of system. The relation might give flexibility to changes that don't fulfill the suitability of field, such as banking, remote communication, and e-commerce. Improved module and traceability can be obtained from quick identification of crosscutting concern. Improved module and traceability can be obtained from quick identification of crosscutting concern. A general model is an aspect-oriented requirement engineering model, and can be concretized through XML specification. The focus of this model lies on modulation, and the construct of requirement is the concern which crosscuts other requirement. Various crosscutting creates intricately entangled code from understanding and maintenance. The examples of crosscutting concern relation are unencapsulatable suitability, usability, and security requirement. Figure 1 is the process to which aspect is applied in general lifecycle. Following are the contents that should be checked essentially in aspect-oriented system development.

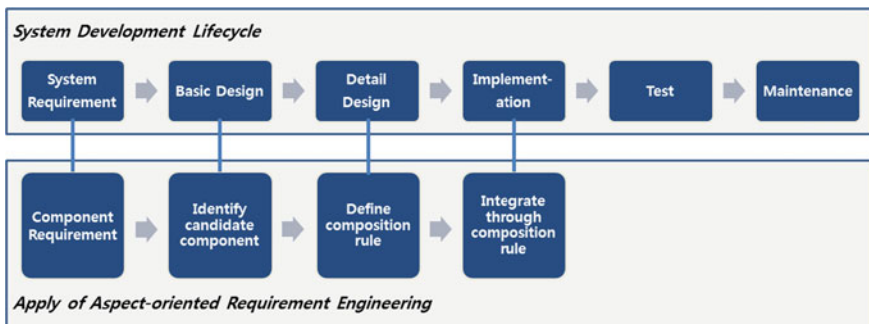


Fig. 1 System development lifecycle for aspect-oriented requirement engineering

3.1 Component Requirement

Define aspect from the requirement document can increase the reusability of codes as well help supporting automation tools. And it is also a process which requires the guideline that system requirement stage. Requirement document can help reduce time and efforts of development for application of aspect in automation tool support.

3.2 Identify Candidate Component

When aspect of requirement stage is identified, specification which affects other requirement in the system should be composed. It should be specified that what relation requirement holds.

3.3 Define Composition Rule

Requirement engineering selects specific architecture in the solution domain to get better idea of problematic field. The developer can recognize that he/she is using aspect-oriented system development technique through lifecycle. Also, he/she can analyze the transaction between aspect modules before the architecture is created.

3.4 Integrate Through Composition Rule

Developers assume that aspect is inaccurate at requirement stage and at design and embodiment stage. They use tracing and mapping in order to solve this problem.

4 Aspect-Oriented System Development Architecture Design

As in software architecture which emphasizes the relations between the components that are forming the software, there is a need to clarify the aspect relations in aspect-oriented architecture design. This is not ease because it must be assumed to be related aspect. Overlapping between different aspects causes new problem that it is not perceptible in a clear language to which module each language segments belongs. This is the major reason for the complexity while composing and maintaining aspect.



Fig. 2 Requirement reflected AORE process

To solve this problem the process shown in Fig. 2 is used. It is a process applying crosscutting concern at the earlier stage of system development lifecycle. Since crosscutting concern and the requirement of developer can be used and applied at requirement stage, there is an advantage of clear identification of candidate aspect.

A conceptual model called aspect structure is to provide aspect-oriented perspective in software structures. Major aspect-oriented development approach possessing clearly defined aspect relation is the aspect-oriented component engineering which supports aspect concern through the lifecycle of specification, design, substantial, and development that are in software component domain.

Following example is showing an online book purchasing order. For the first step, identify the requirement of the developer in order to apply crosscutting concern. Specify as XML as in Fig. 3 to identify candidate aspect regarding the requirement of the developer.

Specification of crosscutting concern and distribution of the developer’s requirement depend on the role of the interaction between the requirement analyst and the developer. In this case, it is possible to use requirement relation for the relation by using matrix. In Fig. 4, it is visible that which crosscutting concern crosscuts to the modules encapsulating the requirement of developer and which is suitable for candidate aspect.

```

<?xml version="1.0" ?>
<viewpoint name="BusinessService">
  <Requirement id="1">
    BusinessService function for on-line books order in necessity.
    <Requirement id="1.1">
      customer leads and viewer searches orders the books. Also, conceives inquires.
      <Requirement id="1.1.1"> The system provide searches, order, to delivery
        information to customer.
      </Requirement>
    </Requirement>
  </Requirement>
  <Requirement id="1.2">
    administrator with the order which is safe conceives for delivery does a security
    function in necessity.
    <Requirement id="1.2.1"> The system when the problem occurs in security, draws
      up the report in administrator
    </Requirement>
  </Requirement>
</Requirement>
</Viewpoint>
  
```

Fig. 3 Specification of the requirement for developer

	Administrator	Customer	Registration	Order	Delevery
View (UI)	○	○			
Member query	○		○		
Book query		○			
Delivery query	○	○			
Member management	○		○		
Book management		○		○	
Delivery management	○	○			○
Security	○	○			

Fig. 4 The requirement of developer and crosscutting relation

Following are the subjects that should be confirmed in aspect-oriented architecture.

4.1 Structural Aspect Identification

Structural aspect identification is an aspect component and is not an ease process because it is not an aspect expression of simple requirement level. New aspect is an aspect defined before the redefinition or redistribution is possible, and can be identified during the designing of architecture. Additionally, new aspect can occur during the architecture definition.

4.2 Expression of Aspect-Oriented Architecture

The expression of aspect-oriented architecture includes not only the general specification and crosscutting interface but also the relation between architecture components and the clear expression of connector. Crosscutting interface is different from a normal interface. Later, it provides service to a different component. Crosscutting interface specifies when and how it will affect the architecture aspect.

The selecting of candidate aspect selects the aspect that first satisfies the requirement of developer and displays the relation of identified aspect as in Fig. 5 based on the requirement of developer and crosscutting concern. Each candidate aspect displays the function its component possesses (provide) as '+', and the function that should be offered by other component (require) as '-'. In this way, system can be established through the identification of aspect based on the requirement of the developer.

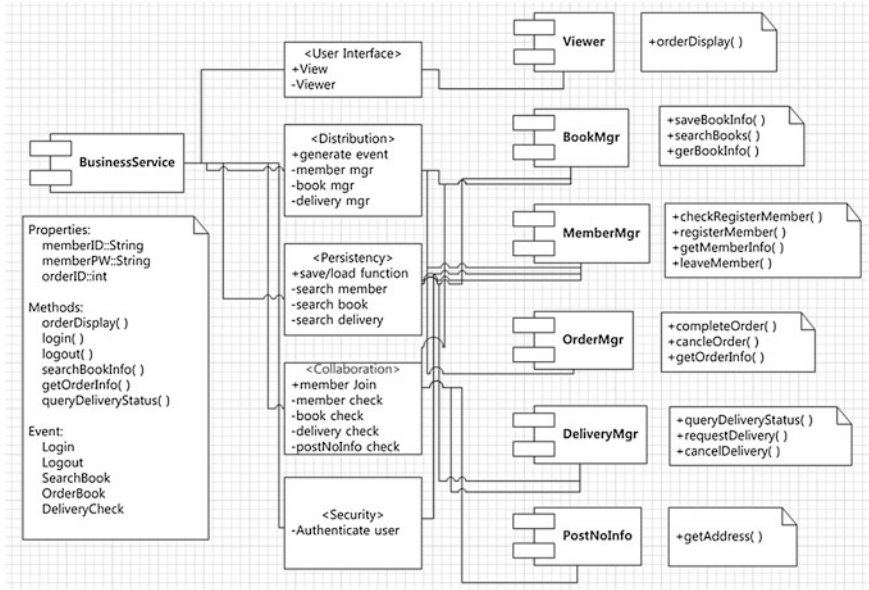


Fig. 5 Aspect structure creation and development

Once the deficient relation between crosscutting concern and the developer is formed and the substitute aspect is identified, the next stage defines detailed composition rule. This rule is a way to encapsulate both requirement and module. After composing the requirement of the candidate aspect and the developer using composition rule, check the recognition and the solution for discordance among candidate aspect.

5 Detailed Design of Requirement and Architecture

In addition to specification of aspect structure, the method of demonstrating the tracing of aspect component and possessing discernible method to substantiate a detailed design is necessary.

5.1 Expressing Design Aspect

Expressing aspect in design requires internal method, specification of internal factor of design aspect of same property, and a modeling language which clearly supports crosscutting interface factors such as advice, point cut, and internal type declaration.

5.2 *Aspect Design Evaluation*

It is necessary for the mechanism which helps the software developer in aspect design valuation. It expresses the framework to access the reuse and the maintenance of aspect-oriented design. The center of framework should be placed in relation principle and in other software properties, the methods of empirical software engineering such as cohesion, coupling and size is used. Quality model requires a method which estimates the probability of reuse and maintenance based on certain regulations.

The final stage of the process is identification of aspect features. Aspect can be mapped on system features and functions (method, object or component), decision (decision made for architecture), design, and substantial aspect. And this is the reason why we decided to call aspect from candidate aspect at the requirement stage despite the crosscutting features at the final stage.

6 Conclusion

Aspect-oriented programming focuses on the abstraction of modularization, encapsulation, and crosscutting concern. It is not easy to apply aspect to the normal object-oriented development engineering, which is why the process that can cover the shortcomings of object-oriented programming has been proposed.

In this thesis, requirement approach is used to apply aspect-oriented software development to normal system lifecycle. We proposed the method applying crosscutting concern at an earlier stage of software development lifecycle. Aspect can be clearly identified since crosscutting concern and the developer's requirement in requirement identifying stage. Using XML, Developer's requirement specific supports extension of functional/nonfunctional crosscutting concept. By using requirement approach at requirement identifying stage, we are able to apply the strong points of aspect including understanding of complex system, decrease of complexity, reuse, customizing, easiness of testing, and enhancement of the ultimate system maintenance.

References

1. Aspect-Oriented Requirements Engineering and Architecture Design, Workshop Report, 2004
2. Clarke S, Walker RJ (2002) Towards a standard design language for AOSD. AOSD, ACM, New York, pp 113–119
3. Glinz M, Wieringa R (2007) Stakeholders in requirements engineering. *IEEE Softw*, *IEEE Comput Soc* 24(2):18–21
4. Kiczales G, Lamping J, Mendhekar A, Maeda C, Lopes C, Loingtier J-M, Irwin J (1997) Aspect-oriented programming. In: *Proceedings of ECOOP '97*, Springer

5. Kiczales G et al (2001) Aspect-oriented programming
6. Nuseibeh B (2001) Weaving together requirements and architectures. *IEEE Comput* 34(3):115–117
7. Rashid A, Sawyer P, Moreira A Araujo J (2002) Early aspects: a model for aspect-oriented requirements engineering. *IEEE joint international conference on requirements engineering, 2002*, IEEE CS Press, pp 199–202
8. Rashid A, Moreira A, Araujo J (2003) Modularisation and composition of aspectual requirements, *AOSD 2003*. ACM, New York, pp 11–20
9. Sommerville I, Sawyer P (1997) *Requirements engineering—a good practice guide*. Wiley, New Jersey

System Analysis and Modeling Using SysML

Muzaffar Iqbal, Muhammad Uzair Khan and Muhammad Sher

Abstract In software engineering, Unified Modeling Language (UML) is considered as the de-facto standard for modeling Object Oriented Systems. On the other hand when it comes to system engineering, then UML is believed to be not as good. More precisely, UML is not effective when modeling system's (Non-functional) requirements, linking these non-functional requirements with other artifacts of the system and defining constraints on the system, in an effective manner to define the system architecture. OMG (Object Management Group) released SysML (a UML-profile) to overcome such limitations of UML, when applied to system engineering. This paper is an effort to show how system's structure, its constraints and (non-functional) requirements can be effectively modeled and linked with each other in SysML with help of a case study.

Keywords Unified modeling language (UML) · System modeling language (SysML) · System engineering · Software engineering

M. Iqbal (✉) · M. Sher
Department of Computer Science and Software Engineering,
International Islamic University, Islamabad, Pakistan
e-mail: muzaffar.iqbal123@yahoo.com

M. Sher
e-mail: m.sher@iiu.edu.pk

M. U. Khan
Department of Computer Science, National University of Computer
and Emerging Sciences, NUCES-FAST, Islamabad, Pakistan
e-mail: uzair.khan@nu.edu.pk

1 Introduction

UML is considered the best choice while modeling software systems but is considered inadequate for modeling system engineering [1–3, 4]. Considering the inefficient constructs of UML for system engineering, OMG (Object Management Group) has released a new Domain Specific Language (DSL) called SysML [5, 6]. SysML (System Modeling Language) is designed to model complex systems specifically modeling system requirements and parametrics.

SysML is based on UML 2. Some diagrams of UML 2 (e.g. Object Diagram, Communication Diagram, Components Diagram, Deployment Diagram, Interaction Overview Diagram, Profile Diagram) are not included in SysML [7]. SysML has introduced some diagrams which were not present in UML 2 (e.g. Requirement Diagram and Parametric Diagram). Some diagrams of UML 2 are reused without any modifications and some are reused after some modifications.

Figure 1 shows the broad relationship between UML 2 and SysML.

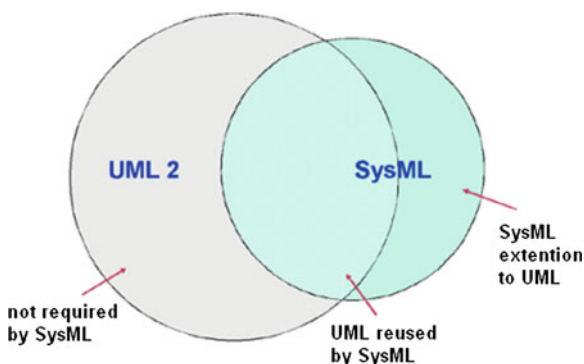
Figure 2 below depicts the various diagrams of SysML. It also shows the diagrams reused from UML 2 with or without modifications. The diagrams new to SysML are also shown in the following figure.

The primary structural unit in SysML is «block». SysML uses this structural unit to represent different artifacts of the system including hardware, software, personnel or any other facility. Block Definition Diagram (BDD) and Inter Block Diagram (IBD) are used for defining the basic system structure [5, 8].

SysML introduces a diagram known as Requirement Diagram. SysML Requirement Diagram can be used to model functional as well as non-functional requirements and their links with each other and other artifacts of the system to show that how they are satisfied.

In this paper, we present the case study of traffic controller to discuss the structure of the system using SysML. We first draw the system's requirement diagram to show the functional and non-functional requirements of the system and show the connection of different requirements with different parts of the system. We present the requirements of the system by using SysML requirement diagram.

Fig. 1 Relationship between UML and SysML



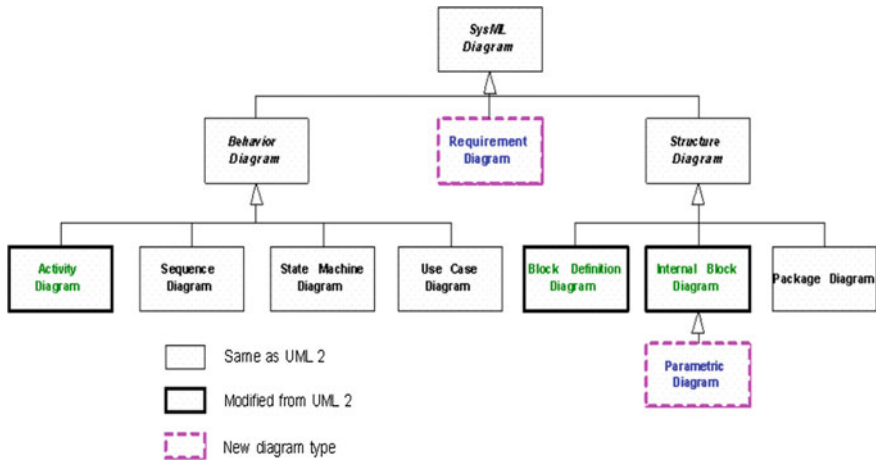


Fig. 2 SysML diagrams (courtesy [5])

After representing the external entities and their relationship with the system, the system’s structure will be constructed by using SysML Block Definition Diagram and Parametric Diagram.

Here is the paper’s structure. Section 2 discusses the selected case study. Section 3 discusses system’s requirements and SysML Requirement Diagram. Section 4 introduces SysML Block Definition Diagram (BDD) and its application to our case study. Section 5 describes SysML Parametric Diagram and its related application to our selected case study. Finally we provide conclusion in Sect. 6.

2 Case Study

The selected case study for this work is based on [9]. This case study is about a traffic controller. The traffic controller is responsible to control traffic signals at a traffic junction as shown in Fig. 3. Traffic lights 1 and 3 are required to always show the same signal as shown by lights 2 and 4 at a particular time.

One traffic cycle is Green–Orange–Red and another cycle is Green–Orange–Red. The safety requirement is that one pair of traffic lights must be red for 30 s and another pair of traffic lights must be green at any given time. The controller is responsible for ‘change direction’. The currently red signals must be set to green, and the currently green signals to red in response to ‘change direction’.

We will be using OMG SysML to model some of the major components of traffic controller discussed in this case study. The focus of this paper is using Requirement Diagram and Parametric Diagram of SysML to represent system requirements and its constraints respectively. Additionally we will show the structure of traffic controller by using Block Definition Diagram (BDD).

Fig. 3 Layout of traffic lights (reproduced from [8])

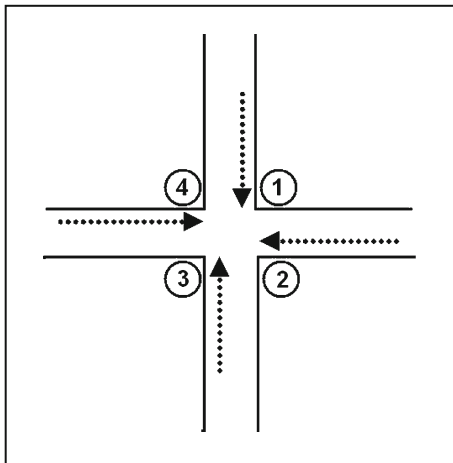
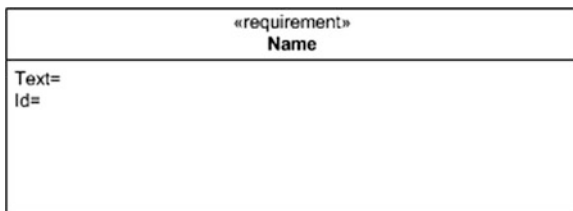


Fig. 4 Requirement diagram



3 SysML Requirement Diagram

UML is considered inefficient to model non-functional requirements [4]. To remove this inefficiency of UML, SysML introduces Requirement Diagram which can model and represent non-functional requirements as model element of the system structure.

Requirement Diagram of SysML has requirement name, requirement text and optional unique ID as shown in the Fig. 4.

Requirements and their relationships can be represented in a number of ways while using Requirements Diagram [6]. The relationships among different requirements of the system and their hierarchy can be easily shown. The relationships: hierarchy, satisfy, refine, verify, trace etc. can be used while requirements are being modeled. In SysML, requirements can also be represented in a tabular form [5] but is beyond the scope of the paper.

In the case study we have selected, different requirements (both functional and non-functional) exist. Some of the requirements are as below.

3.1 Software Requirements

The intended system to be developed in our selected case study is a traffic controller at a road junction (Fig. 3). Pairs of traffic lights (1, 3) and (2, 4) are required to show the same signal indication for a specified interval. There are two light cycles Green–Orange–Red and Red–Orange–Green.

There are two more (non-functional) requirements: safety and reliability. In order to ensure safety, it is required that one pair must be red at certain given time.

The reliability requirements ensure that:

- Lights of all directions must not be red at the same time
- Lights of all directions must not be green at the same time
- Continuous power supply is provided.

The above stated requirements can be modeled using SysML Requirement Diagram as shown in Fig. 5.

In Fig. 5 we can see that how flexible and powerful is the SysML Requirement Diagram. It enables us to show the requirements as a standardized model. The figure shows that how the traffic controller reliability (NFR) is contained by other

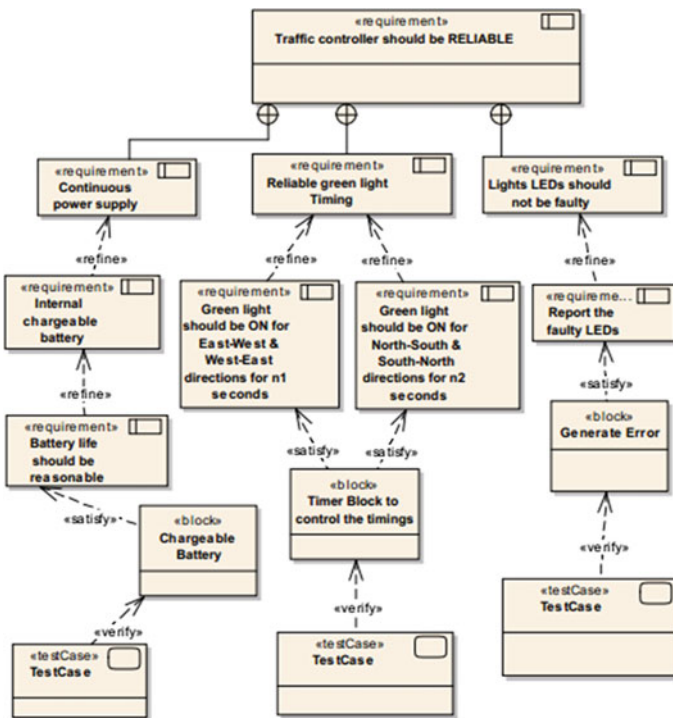


Fig. 5 Requirement diagram of the case study

requirements. It means that the system will be called reliable when all the following (contained) requirements are satisfied.

Figure 5 also illustrates some of the relationships like satisfy and refine. For example, ‘Reliable green light Timing’ requirement is refined by two requirements. On the other hand, ‘TrafficController’ block satisfies some of the requirements.

4 SysML Block Diagram

4.1 Block

In object orientation, classes are the central elements while in SysML classes are replaced by the blocks [10]. Blocks have the capability for modeling different type of systems and its features including both structural and behavioral [5]. Through the concept of block, SysML describes the static structure of systems. A block might be physical or logical unit of a system [10].

4.2 Block Definition Diagram (BDD)

The SysML Block Definition Diagram (BDD) in SysML is the simplest way to describe the system’s structure. It also provides a number of features like associations, dependencies and generalizations among blocks. Just like UML class, SysML BDD also uses properties and operations in order to define a static feature of the system [5].

The general structure of SysML Block Definition Diagram (BDD) is given in Fig. 6. BDD is equivalent to Class diagram in UML but offers more powerful features than class diagram of UML. The Block Definition Diagram (BDD) of our case study is given in Fig. 7. This BDD depicts the composition of a block by relating blocks with one another using the composition relationship. The figure shows that the Traffic Controller is composed of a number of subsystems; Power, Processing and TrafficLight. Each subsystem is modeled by a separate block while each block is subsequently decomposed in sub-blocks.

5 SysML Parametric Diagram

SysML has introduced another diagram which was not present in UML called ‘Parametric Diagram’. Parametric diagram works with association with a diagram called ‘Constraint Block’ [11].

Fig. 6 SysML block definition diagram (BDD) (courtesy [5])

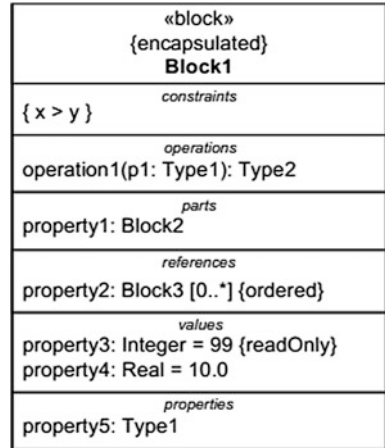
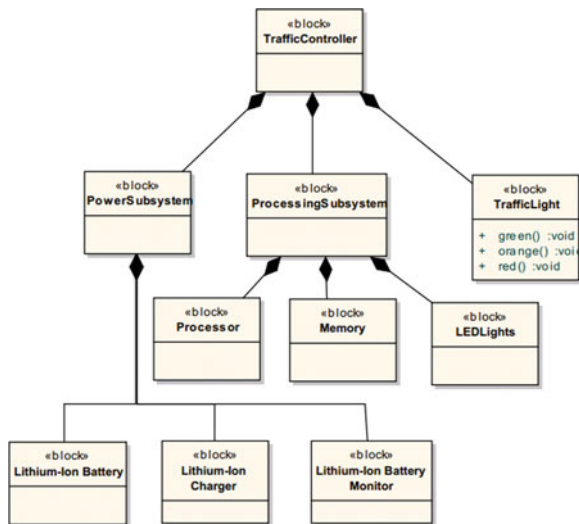


Fig. 7 Block definition diagram of traffic controller



5.1 Constraint Block

Constraint blocks can be used to present a way for integrating different system engineering analysis e.g. performance and reliability models with other SysML models [5]. By using Constraint block we can define a number of constraints and rules (based on these constraints) that must be conformed by the system.

Formulated constraints are regular UML constraints [9] and are usually written by using OCL which is recognized and evaluated (automatically) by almost all the tools that support SysML.

A constraint block, as shown in Fig. 8, defines the parameters used by the constraints like attributes. The notation used by constraint block is very similar to

that for blocks with the difference that stereotype «constraint» is used instead of «block» . A constraint block also contains another compartment used to write constraints using the parameters defined in the ‘parameters’ compartment. Figure 8 shows the constraint block of our case study.

Figure 8 shows that how constraint’s parameters and constraint’s formulae can be defined by using SysML constraint block. The list of constraints can be large but the figure shows some of the primary constraints related to our case study.

5.2 Parametric Diagram

Once constraints are defined then they can be applied to the system by using parametric diagram. A parametric diagram is a special kind of SysML Internal Block Diagram (IBD) that shows the usage of constraint blocks along with the properties they constrain within a given context [5, 9]. Parametric diagrams are based on one or more constraint blocks. Figure 9 shows the parametric diagram of our case study based on the constraint block shown in Fig. 8.

A parametric diagram could be very complex and as simple as demonstrated in Fig. 9. The use of a constraint block on a parametric diagram can be shown as round-cornered rectangles as shown in Fig. 9 above. Small rectangles at the inside edge of the parametric diagram represent constraint parameters. They also provide connection points while linking them to other constraints or parts.

It is important to note that although we can define constraints (in form of OCL) using constraint block but some SysML supported tools still encourage to write the OCL constraint in form of scripts or comments. This allow us to write more complex constraints which would be difficult to write or show using constraint block.

Once constraints are defined then most SysML supported tools allow us to draw parametric diagram automatically based on the available scenario (constraint block).

Fig. 8 Constraint block of traffic controller

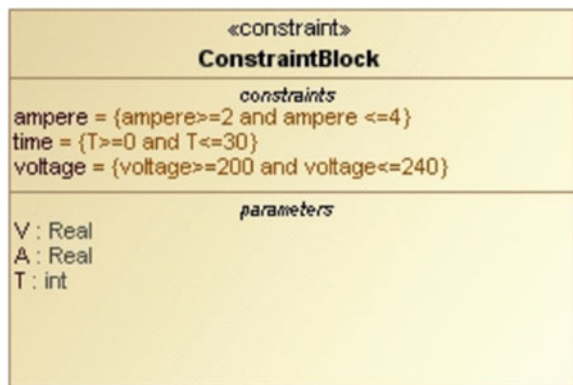
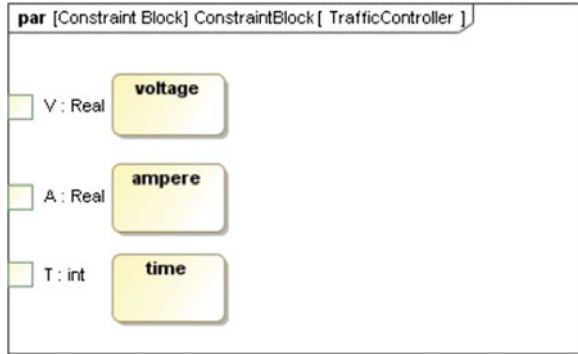


Fig. 9 Parametric diagram of traffic controller



The generated parametric diagram can be further customized to meet the requirements.

It is important to note that after construction of parametric diagram some code is required to write to accomplish the task. This code can be written in an appropriate language like VHDL, Verilog etc. depending upon the system requirements and the available facilities of the language.

6 Conclusion

In this paper, we have shown how SysML Requirement Diagram, BDD (Block Definition Diagram) and Parametric Diagram can be exercised to model the system's structure. SysML Requirement Diagram is one of the powerful SysML diagram valued by both software and system engineers. As we have seen that Requirement Diagram has made the life of engineers very simple while modeling non-functional requirements as system construct and linking them with other parts of the system. It also helps us to analyze whether a particular requirement is satisfied or not. At the other hand, BDD is used to model the system's structure just like class diagram is used in UML. The paper also illustrated that how Parametric Diagram can be used in association with constraint block to define constraints of the system. Since there are only a few case studies available on SysML, this effort will provide a basis of understanding of SysML for modeling complex systems.

References

1. Cantor M (2003) Panel: Extending UML from Software to Systems Engineering. In: Proceedings of the 10th IEEE international conference and workshop on the engineering of computer-based systems, IEEE, Huntsville, Alabama
2. www.omg.org/syseng/

3. Pettit RG (2004) Lessons learned applying UML in embedded software systems designs. In: IEEE
4. Vanderperren Y (2005) UML 2 and SysML: an approach to deal with complexity in SoC/NoC design. In: Proceedings of design, automation and test in Europe
5. OMG Systems Modeling Language, The Official OMG SysML Specification Site. www.sysml.org/
6. www.omgsysml.org/
7. <http://www.sysmlforum.com/faq/>
8. Fonoage M (2008) SysML, UML, and SDL: a comparison. Center for systems integration. Florida Atlantic University, Boca Raton
9. Lano1 K, Evans A (1999) Rigorous development in UML. In: Proceedings of the second international conference on fundamental approaches to software engineering, Springer, London
10. Weilkie T (2007) Systems engineering with SysML/UML: modeling, analysis, design. The MK/OMG Press, USA
11. Holt J, Perry S (2008) SysML for systems engineering. Professional applications of computing series 7. The Institution of Engineering and Technology, London, UK

Part X
Green Convergence Services

Handover Latency Reduction Scheme for Railroad Communications in 4G Mobile Networks

Ronny Yongho Kim and Baik Kim

Abstract In the design of 4G cellular mobile networks, also known as, IMT-Advanced systems, new air interface enhancements such as multi-carrier support and interference mitigation further require handover protocol to be scalable and flexible to support various 4G deployments. This paper presents the state-of-the-art handover schemes designed for IEEE 802.16 m based 4G mobile networks (next generation WiMAX), approved by ITU as an IMT-Advanced technology and provides discussion on technical challenges of railroad communications for high speed trains. Based on the discussion, handover latency reduction scheme for railroad communications is proposed. Various advanced handover procedures accepted in IEEE 802.16 m specification are explained in details in order to derive technical challenges and handover reduction scheme of railroad communications.

Keywords Handover latency · Railroad communications · IEEE 802.16 m

1 Introduction

In order to meet the requirements of growing high speed mobile computing users and large capacity, International Telecommunication Union Radio Section (ITU-R) has commenced the process of developing ITU-R Recommendations for the terrestrial components of the International Mobile Telecommunications-Advanced

R. Y. Kim · B. Kim (✉)
Department of Railroad Electrical and Electronics Engineering,
Korea National University of Transportation, 157 Cheoldo Bakmulgwan-ro,
Uiwang, Gyeonggi, South Korea
e-mail: whitek@ut.ac.kr

(IMT-Advanced) radio interface [1]. IMT-Advanced systems are also known as the 4th-Generation (4G) mobile systems. Emerging broadband wireless air interface specification such as IEEE 802.16 m [2, 3], which provides enhanced link layer amendment to the legacy IEEE 802.16 system [4, 5], is designed to meet and in many cases exceed IMT-Advanced requirements [1]. One of requirements for 4G mobile systems is mobility support moving at very high speed of 300 km/h. However, since such a requirement for high speed vehicle is a minimum requirement, system performance is only functional status. Since handover has the most stringent latency requirement on service interruption time compared to other mobility related operations, handover schemes for high speed vehicles need to be enhanced in order to provide good performance for vehicles.

In this paper, state-of-the-art handover schemes designed for IEEE 802.16 m based next-generation WiMAX [6] are presented and then their technical challenges for high speed vehicles, especially high speed train, are discussed. Based on the discussion, a handover latency reduction scheme is proposed.

2 Handover Procedures

2.1 Basic Handover Procedures

The handover procedure of IEEE 802.16 m consists of four steps that are very similar to IEEE 802.16e hard handover: network topology acquisition, handover preparation, handover execution, handover completion. Major Medium Access Control (MAC) layer enhancements have been done to improve IEEE 802.16 m handover to be a more seamless operation than IEEE 802.16e handover. In the remaining of this section, we will use the call flow shown in Fig. 1 to describe the details of improved IEEE 802.16 m handover procedure. During the Network Topology Acquisition phase, a serving BS broadcasts system information of neighbor Base Stations (BSs) for Mobile Stations (MSs) in its coverage. By using the information in the neighbor advertisement, the MSs can perform scanning efficiently since typically MS will only performs radio quality measurement and skip reading neighbor BS's system information. After scanning neighbor BSs, MS reports the scanning result to the BS. During the Handover Preparation/Initiation phase, both mobile-initiated handover and BS initiated handover are possible. IEEE 802.16 m handover is hard handover and network controlled, for which the network decides one or more candidate target BSs for the MS to handover to, while some flexibility is allowed for the MS to perform target selection if multiple targets are provided. Serving BS negotiates with one or more candidate target BSs for handover preparation by sending a handover request message to each candidate BS. The negotiation between serving BS and target BS during the Handover Preparation stage via R8 interface is of particular importance to enable the enhancement of "seamless handover" option over legacy system. The serving BS will request MAC

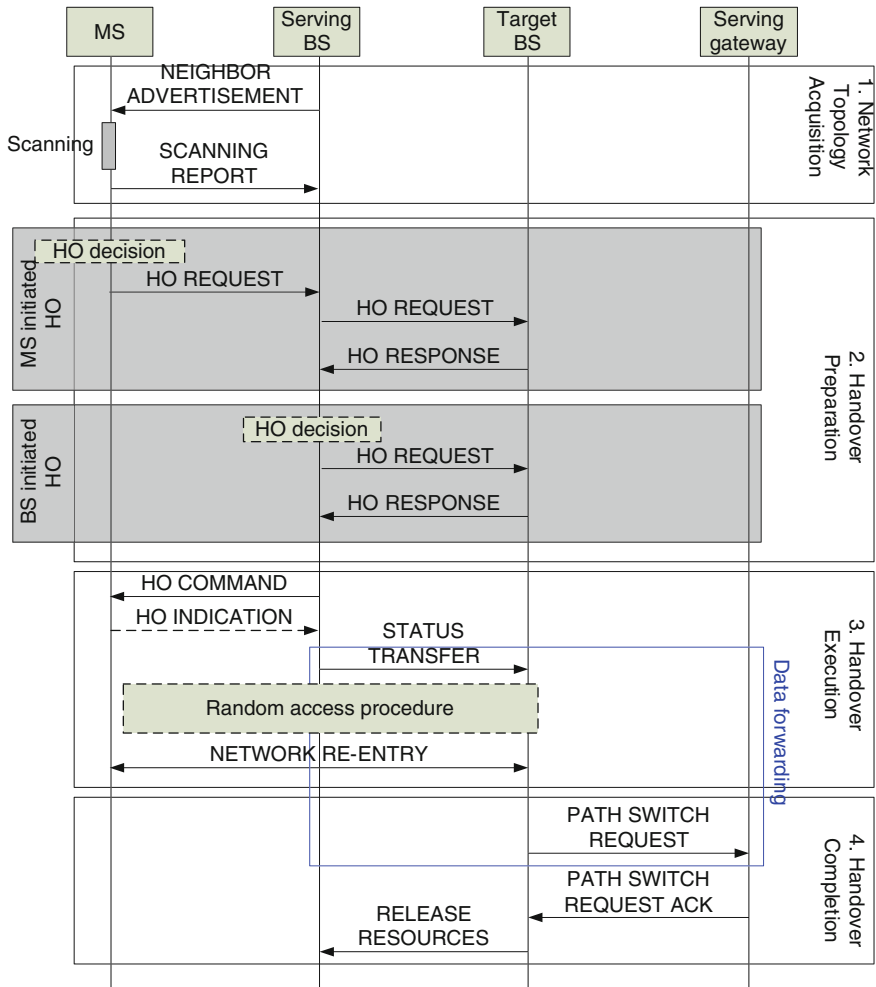


Fig. 1 An overall handover procedure of IEEE 802.16 m system

context pre-establishment, including the MAC station identifier (ST-ID) and all the ongoing service flows, as well as dedicated random access code for MS to access the target BS. The benefits of such negotiation become clear in Handover Execution phase as we discuss in the sequel. During the Handover Execution phase, the serving BS transmits handover command to the MS, in which a Disconnect Time from the serving BS, an Action Time for starting network re-entry, which is the handover procedure for MSs to be re-connected to the network, at target BS, and all the MAC pre-establishment configuration based on serving/target BS negotiation will be included. Upon reception of the handover command, the MS disconnects with the serving BS and performs network re-entry to the target BS specified in the handover command. The network re-entry procedures in IEEE 802.16 m include

downlink/uplink synchronization, request for uplink grants via random access channel, and security key update and mutual authentication. The MS may use the dedicated random access code or opportunity reserved by the target BS during handover preparation to avoid collision in random access. Dedicated ranging resource also allows BS to quickly identify incoming MS upon receiving the random access code. With ST-ID pre-assignment during handover preparation phase, the MS has already obtained a valid ST-ID for accessing the target BS, without need for additional signaling during network re-entry. Subsequently control messages for network re-entry will be exchanged between the MS and the target BS to mutually verify procession of valid security context and complete the protocol. During the Handover Completion phase, upon the completion of the MS network re-entry at the target BS, the target BS initiates a data path registration request to the data path anchor (ASN-GW) to request for data path switching. The data path switching tears down the data forwarding path to the previous serving BS and starts new data path to forward packets for the MS via the new serving BS.

2.2 Multi-Carrier Handover Procedures

Multi-carrier support is a key physical enhancement to boost peak throughput in 4G systems. The multi-carrier handover (MCHO) in IEEE 802.16 m is defined as the handover procedure which involves multiple radio carriers, by leveraging MS multi-carrier radio capability to further reduce handover interruptions. Two kinds of MCHO procedures are supported by the IEEE 802.16 m system, multi-carrier Entry Before Break (EBB) handover and secondary carrier pre-assignment, and they can be performed together or individually depending on the MS capability. In case the MS supports multi-carrier capabilities, the EBB handover can be done by performing network re-entry to the target BS on one carrier while maintaining data transmission with the serving BS on another carrier, and hence the handover re-entry interleaving interval is not required by EBB handover with multi-carrier support. In the operation of multi-carrier, a primary carrier of an MS is the carrier through which control packets between the MS and the BS are exchanged. Secondary carriers are assigned to exchange data packets between the MS and the BS and each secondary carrier has separate physical layer processing including the hybrid automatic repeat request (HARQ) function. Figure 2 illustrates an example of the multi-carrier EBB handover procedure, in which the MS maintains its data communication with the serving BS on its original primary carrier, while performing network re-entry to the target BS on a different carrier frequency. The target carrier frequency may be either previously served as a secondary carrier at the serving BS, or a new carrier not being assigned to the MS at the serving BS. Similar to the single-carrier EBB handover, the radio link between the MS and the serving BS is disconnected upon the expiration of the Disconnect Time defined by the serving BS, or upon reception of the handover completion notification from the target BS.

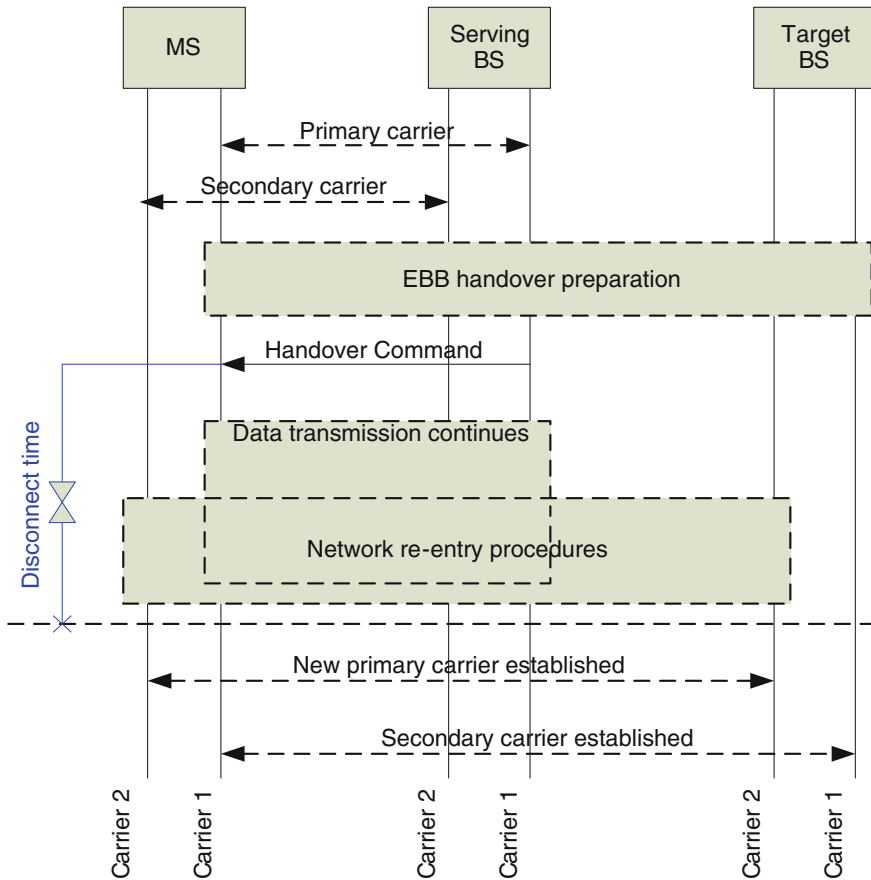


Fig. 2 A basic overall multi-carrier handover procedure of IEEE 802.16 m system

3 Handover Latency Reduction Scheme for Railroad Communications

There are several important technical challenges in order to utilize current handover schemes for high speed vehicles such as high speed train. The most important handover related issue is service interruption due to handover. There exist two kinds of handover latency: layer 2 (L2) handover latency and layer 3 (L3) handover latency.

- L2 Handover Latency
 - Fast Handover Trigger: Fast moving on board mobile terminal is required to trigger handover quickly.

- Fast Channel Measurement: Since channel measurement could not be useful for fast moving on board mobile terminal, channel measurement and report are required to be fast.
- New Station ID (L2 Address) establishment: Since change of base stations is very quick due to fast movement, a new station ID (L2 mobile address) is required to be established very quickly.
- L3 Handover Latency
 - L3 Address Change Detection: Since L2 and L3 are not tightly coupled and L3 has its own way of mobility management, L3 is required to detect L3 handover and necessity of L3 handover.
 - New IP Address (L3 Address) Establishment: In order to continue communication after L2 handover, L3 address should be re-established as quickly as possible after or at the same time of L2 handover.

Railroad communication network has its unique network configuration since train moves along with fixed railroad. Therefore, in case of railroad communications, a simple and efficient handover solution can be derived by taking advantage of its unique network configuration. In order to overcome technical challenges, zone based handover scheme can be considered where several BSs along railroad are grouped as a same handover group and within a same handover zone, same L2 ID can be used for expedited handover. Within a same handover zone, an MS only transmits simple notification to a new serving BS in order to change data path to the MS via the new serving BS. Handover zone can be easily configured by exploiting unique characteristics of railroad communications. L2 handover zone can be tightly coupled with L3 handover zone to further reduce L3 handover latency. Therefore, in case of railroad communication, by utilizing zone based handover concept, handover latency can be substantially reduced.

4 Conclusion

In this paper, the state-of-the-art handover schemes designed for IEEE 802.16 m based 4G mobile networks approved by ITU as an IMT-Advanced technology have been presented and technical challenges for railroad communications in 4G mobile networks have been discussed. Based on the discussion on the technical challenges, handover zone based handover latency reduction scheme is proposed. In the proposed scheme, a novel handover zone concept simplifying handover procedure is employed.

Acknowledgments This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012R1A1A1014610).

References

1. IEEE 802.16 m-09/0034r2, IEEE 802.16 m System Description Document, Sept. 2009
2. IEEE Standard for Local and metropolitan area networks Part 16: Air Interface for Broadband Wireless Access Systems Amendment 3: Advanced Air Interface, IEEE Std 802.16 m-2011, May 12 2011
3. IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Broadband Wireless Access Systems, Amendment 2: Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands and Corrigendum 1, IEEE Std 802.16e-2005, Feb. 28 2006
4. IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Broadband Wireless Access Systems, IEEE Std 802.16-2009 (Revision of IEEE Std 802.16-2004), May 29 2009
5. ITU-R M.2134, Requirements Related to Technical System Performance for IMT-Advanced Radio Interface(s) [IMT.TECH], draft new report, Nov. 2008
6. WiMAX End-To-End Network System Architecture, Stage 3: detailed protocols and procedures, WiMAX Forum, Aug. 2006

Neo Energy Storage Technology: REDOX Flow Battery

Sunhoe Kim

Abstract A REDOX flow battery as a large energy storage technology was reviewed in the paper. REDOX flow battery is a promising technology for the large scale of energy storage and can be expected as a new technology for energy storage device which can replace former secondary batteries. The recent research data and results concerning the REDOX flow battery technologies were summarized in the paper. Among many kinds of REDOX flow batteries the vanadium REDOX flow battery was mainly reviewed in this paper. The REDOX flow battery can be combined with the renewable energy systems, such as solar cell and wind power.

Keywords REDOX flow battery · Energy storage · Secondary battery

1 Introduction

Energy storage has been one of the most serious issues for human life. Various kinds of energy storage devices have been shown in forms of batteries. Among those batteries REDOX flow batteries were introduced in the aspects of both large scale and automotive applications. A redox flow battery is an electrochemical energy storage device. The electro-active species are stored externally and these reactants are circulated through cell or stack as required flow rates. There are several types of redox flow batteries by the electro-active species couple for both anode and cathode. Among them the most promising type is vanadium RFB (VRFB) which uses the vanadium as an electro-active species for both negative and positive sides.

S. Kim (✉)

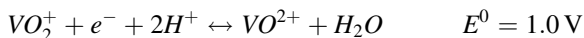
Department of New Energy and Resource Engineering, Sangji University,
83 Sangjidae-gil, Wonju, South Korea
e-mail: sunhoekim@sangji.ac.kr

A VRFB operates on an electrochemical couple based on two different reactions of vanadium ions in acidic aqueous solution. This is made possible because vanadium ions are stable in an unusually high number of valence states: vanadium can be found in +2, +3, +4 and +5 valence states. All four of these valence states are used in a VRFB. A VRFB consists of an assembly of electrochemical cells, each consisting of two half-cells. Separate reactions occur in each half-cell. During discharge, electrons are produced in the reaction in the negative half-cell and are consumed in the reaction in the positive half-cell, forming the basis for an electric current.

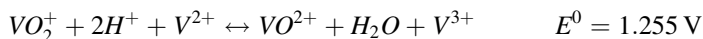
In the negative side during discharge, vanadium (2+) ions in solution are converted to vanadium (3+) ions, with the loss of an electron which is available for conduction:



In the positive side during discharge, vanadium (+5) ions are converted to vanadium (+4) ions, gaining an electron in the process.



The overall equation is: [1]



2 Parts of a VRFB

2.1 Ion Exchange Membrane Electrolyte

The cell of negative and positive side is separated by using a proton exchange membrane (PEM). The membrane physically separates the two vanadium-based electrolyte solutions, preventing self discharge. Several membranes can be used for VRFBs (Fig. 1). Commonly used products are Nafion[®] of DuPont or Asahi Glass Company's CMV, AMV and DMV [2–6].

2.2 Electrodes

The electrodes used in VRFB are composed of high-surface area carbon materials. These materials operate across a wide range of voltage potentials with minimal hydrogen and oxygen evolution, are chemically stable with respect to the acidic electrolytes at both the anode and cathode of a cell, and are available at reasonable costs. Commercial vanadium battery electrodes are usually made with carbon felt.

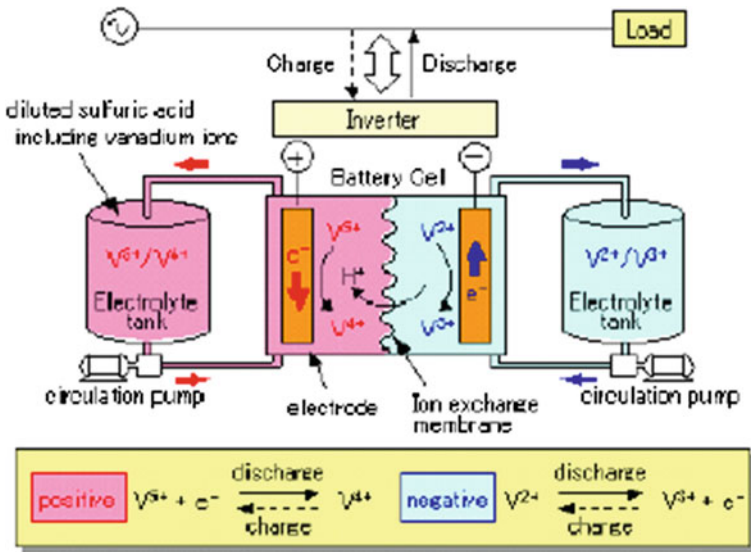


Fig. 1 Single stack flow circuit to describe the principles of operation, main components of a V-RFB system and parts to establish the power and energy rating of the system (Courtesy Sumitomo Electric Industries)

2.3 Separator Plates

The cells in a VRFB stack are separated by bipolar plates, which physically separate each cell from the next, while connecting the two electrically. The plate must be highly conductive and chemically stable in the highly acidic environment of the cell, with the ability to connect with the electrode material with low contact resistance. The tanks must be composed of materials which are resistant to corrosion in the very low pH environment.

2.4 Other Balance of Plants

Other Balance of Plants (BOPs), such as pumps, valves, pipes and other connecting components must be corrosion resistant and stable in acid environments. For this reason, pumps using plastic impellers are used in most installations. Similarly, valves must be rated for low pH environments. Developers usually use PVC material for piping for that reason, which is also strong cost competitive.

3 Performance and Degradation of a VRFB

3.1 Over-Charge and Over-Discharge

Like most of battery electrochemical chemistry, overcharge of the VRFB can have fatal effects on most of the cell components. Most VRFB designs incorporate controls that ensure that the battery as a whole is not overcharged, by watching the open circuit voltage in a reference cell outside the power generating stack. On the other hand, the VRFB is highly tolerant of over-discharge. Because the two electrolytes are identical in the discharged state, it is possible in principle to over-discharge the battery to a point at which the cell polarities are completely reversed, turning the negative electrolyte into the positive electrolyte, and vice versa. In practice, the ability to perform this reversal is dependent on the construction of the cell and cell components, which may not be optimized for such operation.

3.2 Electrolyte Crossover

The unique chemistry of the VRFB also makes it very tolerant of the movement of reactive species from one half-cell to the other. In almost every types of flow battery design, ion exchange between the half-cells is accomplished through the transfer of protons across an ion exchange membrane [7, 8]. For an ideal membrane that conducted only protons, the use of different chemical species in the two electrolytes would present no difficulty. However, real ion exchange membranes are not, and cannot be, perfectly selective. Other chemical species also travel from one electrolyte to the other, albeit at a slow rate. This phenomenon is often called crossover.

For most of battery systems crossover can be a serious contamination issue, leading to irreversible degradation in capacity and performance. In VRFBs, however, crossover does not result in contamination, since both electrolytes are based on the same chemical constituents. Crossover does result in temporary self-discharge, which is easily corrected through charging. This characteristic is another way in which VRFBs show their relative robustness over similar technologies.

3.3 Maintenance

Several VRFB systems have been in operation for the last several years, and some maintenance requirements have been established. For new installations, monthly visual inspections of piping and tanks are required, with detailed inspection at 6 month intervals. Pumps and HVAC systems require inspection every 6 months.

Pump bearings and seals may require replacement at 5 year intervals. Electronic parts such as boards, sensors, relays, and fuses, may require replacement as necessary.

The VRFB cell stacks have a life estimated at over 10 years. They require minimal maintenance over their lifetime, amounting to a visual inspection every 6 months and an exterior cleaning and bolt torque check every year. As the technology matures, intervals between regular maintenance may become longer.

The vanadium electrolyte does not degrade and does not require changing. Early research suggested that electrolyte rebalancing between the positive and negative sides is necessary to cancel the effect of water migration across the ion exchange membrane, but recent investigations have indicated that such rebalancing is probably not necessary. The addition of water may be required to replace water lost due to electrolysis during charging.

3.4 Lifetime

The crucial system component in determining the life of a VRFB system is the cell stack, which can degrade in performance over running time. The key component in the battery stacks is the membrane. Although degradation can happen in the electrodes, especially by overcharge, operation algorithm can easily minimize it. With the assumption of about 1,000 charge/discharge cycles per year, the stack's life time is expected 10–15 years. With replacement or refurbishment of stacks and pump components, a VRFB system can be expected to operate for more than 20 years.

4 Conclusions

Although environmental and cost considerations make traditional solutions more difficult to implement, however, storage is becoming a more attractive option.

As noted earlier, VRFB technology is considered to be one of the large-scale energy storage technologies closest to being commercially available for widespread use. The technical performance of this technology has benefited from continuing research and development, along with real-life experience from a number of demonstration projects conducted in Australia, Japan, and the U.S. over the last few years. These projects have also demonstrated that VRFB do have the capability to address utility problems.

References

1. Gattrell M, Park J, MacDougall B, Apte J, McCarthy S, Wu CW (2004) A study of the mechanism of the vanadium 4+/5+ redox reaction in acidic solutions. *J Electrochem Soc* 151(1):A123–A130
2. Chieng SC, Kazacos M, Kazacos MS (1992) Preparation and evaluation of composite membrane for vanadium redox battery application. *J Membr Sci* 39:11
3. Grossmith F, Llewellyn P, Fane AG, Kazacos MS (1988) Evaluation of membranes for all-vanadium redox cell. *Proc Electrochem Soc Symp* 88: 363 (Honolulu)
4. Larsson R, Folkesson B (2005) A catalytic oxidation of sugar by vanadium (IV). *J Mol Cat Chem* 229(1–2, 29): 183–190
5. Hwang GJ, Ohya H (1996) Preparation of cation exchange membrane as a separator for all-vanadium redox flow battery. *J Membr Sci* 120:55
6. Hwang GJ, Ohya H (1997) Crosslinking of anion exchange membrane by accelerated electron radiation as a separator for the all-vanadium redox flow battery. *J Membr Sci* 132:55
7. Mohammadi T, Chieng SC, Kazacos MS (1997) Water transport study across commercial ion exchange membranes in the vanadium redox flow battery. *J Membr Sci* 133:151
8. Mohammadi T, Kazacos MS (1997) Evaluation of the chemical stability of some membranes in vanadium solution. *J Appl Electrochem* 27:153

An Effective Interfacing Adapter for PRAM Based Main Memory via Flexible Management DRAM Buffer

Mei-Ying Bian, Su-Kyung Yoon and Shin-Dug Kim

Abstract An interfacing adapter is required between cache layer and PRAM based main memory to cover the shortcomings of PRAM. Thus this research is to design a flexible DRAM buffer (FDB) structure, which can improve performance by prefetch candidate data into FDB to reduce the miss penalty, and extends PRAM lifetime by filtering a large portion of write back data upon eviction from last level cache. Our results show that FDB can effectively minimize the access latency to achieve similar performance to the case of DRAM main memory and reduce a certain degree of write count to PRAM, thus limited endurance can get some respite thereby, typically extend their life expectancy.

Keywords Memory hierarchy · PRAM · Buffer management

1 Introduction

PRAM has emerged as a promising candidate for main memory, due to its exceptionally lower leakage power and higher density relative to the case of DRAM. Nevertheless, there are also several challenges required to be overcome before PRAM can be replaced by DRAM as main memory. For example PRAM is wearable

M.-Y. Bian (✉) · S.-K. Yoon · S.-D. Kim
Computer Science, Yonsei University, 50, Yonsei-road, Seodaemun-gu,
Seoul, South Korea
e-mail: jojo04592@gmail.com

S.-K. Yoon
e-mail: mypioioi@naver.com

S.-D. Kim
e-mail: sdkim@yonsei.ac.kr

such that only about $10^6 - 10^8$ writes can be performed. Moreover, PRAM read latency is $2X-4X$ slower than DRAM, write latency is about an order of magnitude slower than read latency and consumes more energy on writes than DRAM [4, 5].

To mitigate these shortcomings, we organize PRAM main memory with a small flexible DRAM buffer (FDB) which is designed as an interfacing buffer module between the last level cache (LLC) and PRAM main memory. The buffer could be a DRAM buffer, in order to provide performance similar to that of DRAM based main memory. The FDB improves performance by caching highly accessed data, and extends PRAM lifetime by filtering a large portion of victim data upon eviction from LLC.

The rest of this paper is organized as follows. The related work and Approach will be explained in Sects. 2 and 3, respectively. And its evaluation is provided in Sect. 4. Finally, Sect. 5 concludes this research.

2 Related Work

In order to overcome PRAM shortcomings, such as longer access latency and limited lifetime, there has been much subsequent work on chip technology. A lazy-write organization when a page fault is serviced, where the page fetched from the hard disk is written only to the DRAM storage, if it is evicted from the DRAM storage, only dirty pages will be written back to PRAM main memory [5]. However, once the system is suspended, previous execution stats will be lost without using any advantage of non-volatile main memory. A PRAM based hybrid memory using superblock-based adapting buffer located between LLC and main memory, when a miss occurs, a set of page units from main memory will aggressively prefetch to the spatial locality superblock buffer, where PRAM main memory maintains all the execution stats [3]. However, this technique requires a large amount of DRAM buffer space it is not conducive to save power.

Thus, to support low power consumption, and low cost, the goal of this research is use a small amount of DRAM buffer space, still providing same target access performance relative to DRAM based conventional memory.

3 Approach

3.1 Fundamental Idea

In order to compensate for slow reads and writes in PRAM, when PRAM replaces DRAM as main memory, we take into account the structure of the hierarchical DRAM and PRAM hybrid structure, a small DRAM buffer which is designed as an interfacing buffer module between the LLC and PRAM main memory. In our

approach, DRAM plays both as write buffer for main memory and also as a prefetch storage space for LLC, PRAM works as the large main memory exploiting to the benefits of larger capacity and low standby power.

To improve the temporal locality, thereby to reduce the number of write operations, we design the DRAM buffer management in a more flexible way. Especially dirty blocks being evicted from the LLC will first be written back to the FDB, where write latency of the slower PRAM main memory can be hidden while increasing the probability of buffer hits to be accessed by LLC. Furthermore, in order to take advantage of spatial locality, when a miss occurs in the LLC, the requested block is fetched into any candidate slot of the LLC and simultaneously its next one block ahead is directly prefetched into FDB upon eviction from the LLC.

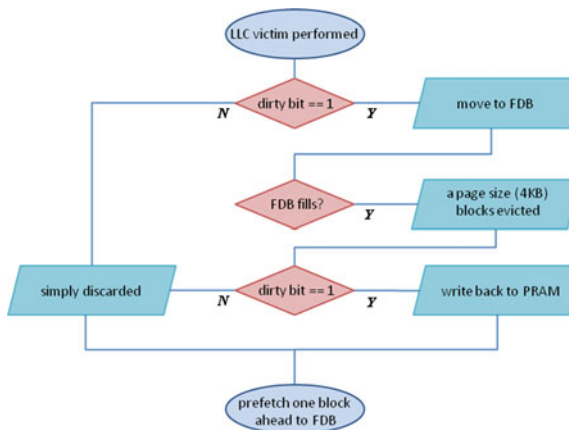
The basic concept of DRAM is used as a buffer of PRAM that has three purposes. The first purpose is to reduce the write count to PRAM, thus limited endurance can get some respite thereby, typically extend their life expectancy. Second purpose is to emphasize spatial locality, when a miss occurs in the LLC, the requested block is fetched into the LLC and simultaneously its next one block ahead is prefetched into FDB. This is to pre-collect and maintain those blocks that might be accessed at later time, thereby to reduce the miss penalty. Finally, for write requests to the PRAM, where write latency of the slower PRAM main memory can be hidden while increasing the probability of buffer hits accessed by LLC.

3.2 Operational Flow of FDB

Basic operational flow of the proposed FDB is provided as shown in Fig. 1. When a data request from LLC and also a victim process is performed, our proposed architecture is performed as follows:

- Step 1 A block request is generated from the LLC simultaneously a block is evicted from the LLC
- Step 2 First check the dirty bit, and look over if the dirty bit is set. If the dirty bit is not set, it is simply discarded from the LLC, otherwise continue to the next step
- Step 3 Check the FDB, find out if there is a mapping block in the FDB. If the evicted block exists in the FDB, go to Step 6, otherwise continue to the next step
- Step 4 Check the FDB, if the FDB has been filled with blocks a FIFO page will be evicted then, continue to the next step
- Step 5 Write back the dirty blocks to the main memory, otherwise simply discarded upon eviction from the FDB, go to Step 6
- Step 6 Transmitting the requested block to the LLC, transmit the next one block ahead to the FDB

Fig. 1 Operational flow of buffer management



4 Performance Analysis

We used a trace driven simulation as our evaluation methodology. We used five scientific applications, e.g., **bzip2**, **mcf**, **soplex**, **sjeng**, **wrf** from SPEC 2006 benchmark [2]. We used GEM5 full system mode to generate address traces of CPU requests. As a basic system configuration chosen in the experiment, L1 instruction and L1 data caches consist of 32-KBytes, 4-way set associative, with a 64 Byte block size. L2 cache memory is configured as a unified 1-MBytes, 8-way set associative cache, with a 128 Byte block size. The proposed FDA is configured 1-MByte, with a 128 Byte block size and managed as in FIFO. Also, for the baseline system the main memory is configured as 1-GByte, 4096 Byte page and 256 Byte block size for comparison. We implemented our memory simulator and added CPU of the GEM5.

4.1 Impact on Access Latency

For comparison, we take conventional DRAM main memory structure as a baseline. Because PRAM access latency is longer than DRAM thus, an effective buffering method to minimize the access latency is required to achieve similar performance to the case of DRAM main memory. However, as demonstrate by [1], read and write latencies for PCM are approximately 2X and 7X longer than those for DRAM.

In this experiment, DRAM access latency is normalized as one. Figure 2 shows the relative DRAM access latency, when we simply replace DRAM with PCM can increase the average access latency by 2.06X. However, a case of unified L2 cache with FDB can perfectly hide PCM longer access latency as in Fig. 2. Average PCM access latency relative to that of DRAM can be reduced up to 1.20X.

Fig. 2 Benchmark access latency when using PCM as a DRAM replacement

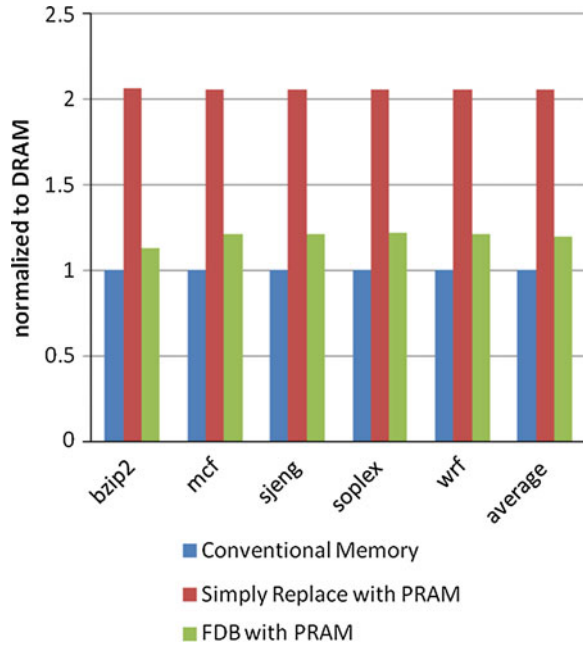
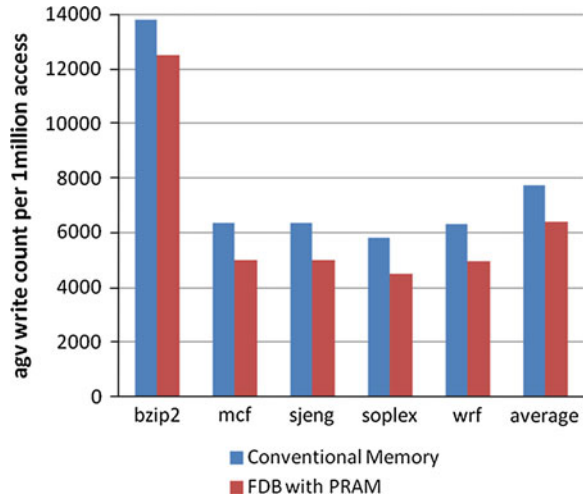


Fig. 3 Comparison of average write-back count



4.2 Impact on PRAM Lifetime

Because of the limited endurance of PRAM, we exploit the FDB as a flexible write buffer to collect dirty blocks that might be accessed again at later time, in this way to extend PRAM life expectancy.

Figure 3 shows the number of writes to the PRAM main memory, where the y-axis gives the average write count per 1 million access requests. As shown in Fig. 3, a combination of the L2 cache and FDB decreases the write count by 17.20 %.

5 Conclusions

To avoid large performance losses due to long memory access delays, we proposed a flexible DRAM buffer to bridge the gap between processor and memory speeds. When any cache miss occurs, the requested block is fetched into the last level cache and simultaneously its next one block ahead is directly prefetched into FDB upon eviction from the last level cache. In order to hide PRAM long write latency for write requests to the last level cache blocks are moved to the FDB first, where write latency of the slower PRAM memory can be hidden. Specifically, this technique may also reduce write back count to PRAM while increasing the probability of buffer hits accessed by L2 cache.

Our results show that FDB can reducing the misses by 90.96 %, comparing with the case of conventional L2 cache (with no buffer). And FDB not only can effectively minimize the access latency to achieve similar performance to the case of DRAM main memory and also reduce a certain degree of write count to PRAM, thus limited endurance can get some respite thereby, typically extend their life expectancy.

References

1. Dhiman G, Ayoub R, Rosing T (2009) PDRAM: a hybrid PRAM and DRAM main memory system. Design automation conference—DAC, ACM, New York, pp 664–469
2. Henning JL (2006) SPEC CPU2006 Benchmark Descriptions. ACM SIGARCH Comput Archit News 34(4):1–17
3. Jung K-S, Park J-W (2011) A superblock-based memory adapter using decoupled dual buffers for hiding the access latency of non-volatile memory. In: World congress on engineering and computer science pp 802–807, 19–21 Oct 2011
4. Lee BC, Ipek E, Mutlu O, Burger D (2009) Architecting phase change memory as a scalable dram alternative. ACM SIGARCH Comput Archit News 37(3):2–13
5. Qureshi MK, Srinivasan V, Rivers JA (2009) Scalable high performance main memory system using phase memory technology. ACM SIGARCH Comput Archit News 37(3):24–33

Erratum to: IT Convergence and Security 2012

Kuinam J. Kim and Kyung-Yong Chung

**Erratum to: K.J. Kim and K.-Y. Chung (eds.),
IT Convergence and Security 2012,
DOI [10.1007/978-94-007-5860-5](https://doi.org/10.1007/978-94-007-5860-5)**

The online version of the original book can be found at DOI [10.1007/978-94-007-5860-5](https://doi.org/10.1007/978-94-007-5860-5).

K. J. Kim

Convergence Security, Kyoung-gi University, Suwon, Gyeonggi-do, Republic of South Korea

K.-Y. Chung

Department of Computer Information Engineering, Sangji University, Wonju-si, Gangwon-do, Republic of South Korea

K. J. Kim and K.-Y. Chung (eds.), *IT Convergence and Security 2012*,
Lecture Notes in Electrical Engineering 215, DOI: [10.1007/978-94-007-5860-5_149](https://doi.org/10.1007/978-94-007-5860-5_149),
© Springer Science+Business Media Dordrecht 2013

E1

Page	Title	Changes
335	Implementation of Improved DPD Algorithm Using the Approximation Hessian Technique and an Adaptive Filter	Second author name “Gyong-Hak Lee” should read as “Kyong-Hak Lee” and “G.-H. Lee” should read as “K.-H. Lee”. In addition, the address for K.-H. Lee should be changed from “School of Mechanical Engineering, Hankook University, 5-4 Sinsa-dong, Seoul, Kangnam-gu, South Korea, e-mail: Kkim@hankook.ac.kr” to “ Industry-Academic Cooperation Foundation, Namseoul University, 91 Daehak-ro Seonghwan-eup Sebuk-gu Cheonan-si Chungcheongnam-do, South Korea, e-mail: khlee@nsu.ac.kr”
487	Unified Performance Authoring Tool for Heterogeneous Devices and Protocols	Acknowledgement is added: This research is supported by Ministry of Culture, Sports and Tourism(MCST) and Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Research & Development Program 2012.
514	Color Coding for Massive Bicycle Trajectories	Acknowledgement is added: This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2012.
865	Modeling Student’s Handwritten Examination Data and Its Application Using a Tablet Computer	Acknowledgement is added: This research is supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2012.
993	Efficient Detection of Content Polluters in Social Networks	“+1” in display equation $\sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} p_i + 1 - p_i }$ should read as $\sqrt{\frac{1}{n-1} \sum_{i=1}^{n-1} p_{i+1} - p_i }$