Jeng-Shyang Pan · Vaclav Snasel
Emilio S. Corchado · Ajith Abraham
Shyue-Liang Wang   *Editors*

Intelligent Data Analysis and Its Applications, Volume II

Proceeding of the First Euro-China
Conference on Intelligent
Data Analysis and Applications,
June 13–15, 2014, Shenzhen, China

Springer

Advances in Intelligent Systems and Computing

Volume 298

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

About this Series

The series "Advances in Intelligent Systems and Computing" contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing.

The publications within "Advances in Intelligent Systems and Computing" are primarily textbooks and proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

Jeng-Shyang Pan · Vaclav Snasel
Emilio S. Corchado · Ajith Abraham
Shyue-Liang Wang

Intelligent Data analysis and Its Applications, Volume II

Proceeding of the First Euro-China Conference on Intelligent Data Analysis and Applications, June 13–15, 2014, Shenzhen, China

Editors

Jeng-Shyang Pan
University of Applied Sciences
Kaohsiung
Taiwan

Vaclav Snasel
Department of Computer Science
Faculty of Elec. Eng. & Comp. Sci.
VSB-Technical University of Ostrava
Ostrava
Czech Republic

Emilio S. Corchado
Departamento de Informáca y Automática
Facultad de Biología
University of Salamanca
Salamanca
Spain

Ajith Abraham
Machine Intelligence Research Labs (MIR Labs)
Scientific Network for Innovation and Research Excellence
Washington
USA

Shyue-Liang Wang
Department of Information Management
National University of Kaohsiung
Kaohsiung
Taiwan

Preface

This volume composes the proceedings of the First Euro-China Conference on Intelligent Data Analysis and Applications (ECC 2014), which was hosted by Shenzhen Graduate School of Harbin Institute of Technology and was held in Shenzhen City on June13–15, 2014. ECC 2014 was technically co-sponsored by Shenzhen Municipal People's Government, IEEE Signal Processing Society, Machine Intelligence Research Labs, VSB-Technical University of Ostrava (Czech Republic), National Kaohsiung University of Applied Sciences (Taiwan), and Secure E-commerce Transactions (Shenzhen) Engineering Laboratory of Shenzhen Institute of Standards and Technology. It aimed to bring together researchers, engineers, and policymakers to discuss the related techniques, to exchange research ideas, and to make friends.

113 papers were accepted for the final technical program. Four plenary talks were kindly offered by: Ljiljana Trajkovic (IEEE SMC president), C.L. Philip Chen (IEEE Fellow, University of Macau), Jhing-Fa Wang (Tajen University, Taiwan), and Ioannis Pitas (University of Thessaloniki, Greece).

We would like to thank the authors for their tremendous contributions. We would also express our sincere appreciation to the reviewers, Program Committee members and the Local Committee members for making this conference successful. Finally, we would like to express special thanks for the financial support from Shenzhen Municipal People's Government and Shenzhen Graduate School of Harbin Institute of Technology in making ECC 2014 possible.

June 2014

Jeng-Shyang Pan
Vaclav Snasel
Emilio S. Corchado
Ajith Abraham
Shyue-Liang Wang

Conference Organization

Honorary Chairs

Han-Chieh Chao   National Ilan University, Taiwan

Advisory Committee Chairs

Tzung-Pei Hong   National University of Kaohsiung, Taiwan
Bin-Yih Liao   National Kaohsiung University of Applied
                                  Sciences, Taiwan

Conference Chairs

Jeng-Shyang Pan   Harbin Institute of Technology Shenzhen Graduate
                                  School, China
Vaclav Snasel   VSB-Technical University of Ostrava, Czech
                                  Republic

Program Committee Chairs

Emilio S. Corchado   University of Salamanca, Spain
Ajith Abraham   Machine Intelligence Research Labs, USA
Shyue-Liang Wang   National University of Kaohsiung, Taiwan

Invited Session Chairs

Wu-Chih Hu   National Penghu University of Science and
                                  Technology, Taiwan
Kuo-Kun Tseng   Harbin Institute of Technology Shenzhen Graduate
                                  School, China

Electronic Media Chairs

Jiun-Huei Ho	Cheng Shiu University, Taiwan
Tsu-Yang Wu	Harbin Institute of Technology Shenzhen Graduate School, China

Local Organizing Chairs

Yanfeng Zhang	Harbin Institute of Technology Shenzhen Graduate School, China
Chun-Wei Lin	Harbin Institute of Technology Shenzhen Graduate School, China
Chien-Ming Chen	Harbin Institute of Technology Shenzhen Graduate School, China

Publication Chairs

Shu-Chuan Chu	Flinders University, Australia

Finance Chairs

Linlin Tang	Harbin Institute of Technology Shenzhen Graduate School, China

International Program Committee

Abdel hamid Bouchachia	University of Klagenfurt, Austria
Abd. Samad Hasan Basari	Universiti Teknikal Malaysia Melaka, Malaysia
Abraham Duarte	Universidad Rey Juan Carlos, Spain
Akira Asano	Kansai University, Japan
Alberto Alvarez	European Centre for Soft Computing, Spain
Alberto Cano	University of Cordoba, Spain
Alberto Fernandez	Universidad de Jaen, Spain
Alberto Bugarin	University of Santiago de Compostela, Spain
Alex James	Indian Institute of Information Technology and Management – Kerala, India
Alexandru Floares	Cancer Institute Cluj-Napoca, Romania
Alma Gomez	University of Vigo, Spain
Amelia Zafra Gomez	University of Cordoba, Spain
Amparo Fuster-Sabater	Institute of Applied Physics (C.S.I.C.), Spain
Ana Lorena	Federal University of ABC, Brazil
Anazida Zainal	Universiti Teknologi Malaysia, Malaysia
Andre Carvalho	University of Sao Paulo, Brazil
Andreas Koenig	Technische Universitat Kaiserslautern, Germany

Anna Bartkowiak	University of Wroclaw, Poland
Anna Fanelli	Universita di Bari, Italy
Antonio Peregrin	University of Huelva, Spain
Antonio J. Tallon-Ballesteros	University of Seville, Spain
Anusuriya Devaraju	Forschungszentrum Julich GmbH, Germany
Aranzazu Jurio	Universidad Publica de Navarra, Spain
Ashish Umre	University of Sussex, United Kingdom
Ashraf Saad	Armstrong Atlantic State University, United States
Ayeley Tchangani	University Toulouse III , France
Aymeric Histace	Universite Cergy-Pontoise, France
Azah Kamilah Muda	Universiti Teknikal Malaysia Melaka, Malaysia
Bartosz Krawczyk	Politechnika Wrocławska, Poland
Beatriz Pontes	University of Seville, Spain
Brijesh Verma	Central Queensland University, Australia
Carlos Barranco	Pablo de Olavide University, Spain
Carlos Cano	University of Granada, Spain
Carlos Fernandes	GeNeura Team, Spain
Carlos Garcia-Martinez	University of Cordoba, Spain
Carlos Lopezmolina	Universidad Publica de Navarra, Spain
Carlos Morell	Universidad Central Marta Abreu de Las Villas, Cuba
Cesar Hervas-Martinez	University of Cordoba, Spain
Chang-Shing Lee	National University of Tainan, Taiwan
Chao-Chun Chen	Southern Taiwan University, Taiwan
Chia-Feng Juang	National Chung-Hsing University, Taiwan
Chien-Ming Chen	Harbin Institute of Technology Shenzhen Graduate School, China
Chin-Chen Chang	Feng Chia University, Taiwan
Chris Cornelis	Ghent University, Belgium
Chuan-Kang Ting	National Chung Cheng University, Taiwan
Chu-Hsing Lin	Tunghai University, Taiwan
Chun-Wei Lin	Harbin Institute of Technology Shenzhen Graduate School, China
Coral del Val	University of Granada, Spain
Crina Grosan	Norwegian University of Science and Technology, Norway
Cristina Rubio-Escudero	University of Sevilla, Spain
Cristobal Romero	University of Cordoba, Spain
Cristobal J. Carmona	University of Jaen, Spain
Dalia Kriksciuniene	Vilnius University, Lithuania
David Becerra-Alonso	ETEA-INSA, Spain
Detlef Seese	Karlsruhe Institut of Technology (KIT), Germany
Edurne Barrenechea	Universidad Publica de Navarra, Spain
Eiji Uchino	Yamaguchi University, Japan

Eliska Ochodkova	VŠB Technical University of Ostrava, Czech Republic
Elizabeth Goldbarg	Federal University of Rio Grande do Norte, Brazil
Emaliana Kasmuri	Universiti Teknikal Malaysia Melaka, Malaysia
Enrique Herrera-Viedma	University of Granada, Spain
Enrique Yeguas	University of Cordoba, Spain
Eulalia Szmidt	Systems Research Institute Polish Academy of Sciences, Poland
Eva Gibaja	University of Cordoba, Spain
Federico Divina	Pablo de Olavide University, Spain
Fernando Bobillo	University of Zaragoza, Spain
Fernando Delaprieta	University of Salamanca, Spain
Fernando Gomide	University of Campinas, Brazil
Fernando Jimenez	University of Murcia, Spain
Francesc J. Ferri	Universitat de Valencia, Spain
Francesco Marcelloni	University of Pisa, Italy
Francisco Fernandez Navarro	University of Cordoba, Spain
Francisco Herrera	University of Granada, Spain
Francisco Martinez-Alvarez	Pablo de Olavide University, Spain
Francisco Martinez-Estudillo	University Loyola Andalucia, Spain
Frank Klawonn	University of Applied Sciences Baunschweig, Germany
Gabriel Luque	University of Malaga, Spain
Gede Pramudya	Universiti Teknikal Malaysia Melaka, Malaysia
Giacomo Fiumara	University of Messina, Italy
Giovanna Castellano	Universita di Bari, Italy
Giovanni Acampora	University of Salerno, Italy
Girijesh Prasad	University of Ulster, United Kingdom
Gladys Castillo	University of Aveiro, Portugal
Gloria Bordogna	CNR IDPA, Italy
Gregg Vesonder	AT&T Labs Research, United States
Huiyu Zhou	Queen's University Belfast, United Kingdom
Ilkka Havukkala	Intellectual Property Office of New Zealand, New Zealand
Imre Lendak	University of Novi Sad, Serbia
Intan Ermahani A. Jalil	Universiti Teknikal Malaysia Melaka, Malaysia
Isabel Nunes	UNL/FCT, Portugal
Isabel S. Jesus	Instituto Superior de Engenharia do Porto, Portugal
Ivan Garcia-Magarino	Universidad a Distancia de Madrid, Spain
Jae Oh	Syracuse University, United States
Jan Martinovic	VŠB Technical University of Ostrava, Czech Republic
Jan Plato	VŠB Technical University of Ostrava, Czech Republic

Javier Perez   University of Salamanca, Spain
Javier Sedano   Technological Institute of Castilla y Leon, Spain
Jesus Alcala-Fdez   University of Granada, Spain
Jesus Serrano-Guerrero   University of Castilla-La Mancha, Spain
Jitender S. Deogun   University of Nebraska, United States
Joaquin Lopez Fernandez   University of Vigo, Spain
Jorge Nunez Mc Leod   Institute of C.E.D.I.A.C, Argentina
Jose Luis Perez de la Cruz   University of Malaga, Spain
Jose M. Merigo   University of Barcelona, Spain
Jose-Maria Luna   University of Cordoba, Spain
Jose Pena   Universidad Politecnica de Madrid, Spain
Jose Raul Romero   University of Cordoba, Spain
Jose Tenreiro Machado   Instituto Superior de Engenharia do Porto, Portugal
Jose Valente De Oliveira   Universidade do Algarve, Portugal
Jose Villar   Oviedo University, Spain
Juan Botia   Universidad de Murcia, Spain
Juan Gomez-Romero   Universidad Carlos III de Madrid, Spain
Juan Vidal   Universidade de Santiago de Compostela, Spain
Juan J. Flores   Universidad Michoacana de San Nicolas de
                                     Hidalgo, Mexico
Juan-Luis Olmo   University of Cordoba, Spain
Julio Cesar Nievola   Pontificia Universidade Catolica do Parana, Brazil
Jun Zhang   Waseda University, Japan
Jyh-Horng Chou   National Kaohsiung First Univ. of Science and
                                     Technology, Taiwan
Kang Tai   Nanyang Technological University, Singapore
Kaori Yoshida   Kyushu Institute of Technology, Japan
Kazumi Nakamatsu   University of Hyogo, Japan
Kelvin Lau   University of York, United Kingdom
Kubilay Ecerkale   Turkish Air Force Academy, Turkey
Kumudha Raimond   Karunya University, India
Kun Ma   University of Jinan, China
Leandro Coelho   Pontificia Universidade Catolica do Parana, Brazil
Lee Chang-Yong   Kongju National University, Korea
Leida Li   University of Mining and Technology, China
Leon Wang   National University of Kaohsiung, Taiwan
Liang Zhao   University of Sao Paulo, Brazil
Liliana Ironi   IMATI-CNR, Italy
Luciano Stefanini   University of Urbino "Carlo Bo", Italy
Ludwig Simone   North Dakota State University, United States
Luigi Troiano   University of Sannio, Italy
Luka Eciolaza   European Centre for Soft Computing, Spain
Macarena Espinilla Estevez   Universidad de Jaen, Spain
Manuel Grana   University of Basque Country, Spain

Manuel Lama	Universidade de Santiago de Compostela, Spain
Manuel Mucientes	University of Santiago de Compostela, Spain
Marco Cococcioni	University of Pisa, Italy
Maria Nicoletti	Federal University of Sao Carlos, Brazil
Maria Torsello	Universita di Bari, Italy
Maria Jose Del Jesus	Universidad de Jaen, Spain
Mariantonietta Noemi La Polla	IIT-CNR, Italy
Maria Teresa Lamata	University of Granada, Spain
Mario Giovanni C.A. Cimino	University of Pisa, Italy
Mario Koeppen	Kyushu Institute of Technology, Japan
Martine De Cock	Ghent University, Belgium
Michael Blumenstein	Griffith University, Australia
Michal Kratky	VŠB Technical University of Ostrava, Czech Republic
Michal Wozniak	Wroclaw University of Technology, Poland
Michela Antonelli	University of Pisa, Italy
Mikel Galar	Universidad Publica de Navarra, Spain
Milos Kudelka	VŠB Technical University of Ostrava, Czech Republic
Min Wu	Oracle, United States
Noor Azilah Muda	Universiti Teknikal Malaysia Melaka, Malaysia
Norberto Diaz-Diaz	Pablo de Olavide University, Spain
Norton Gonzalez	University of Fortaleza, Brazil
Nurulakmar Emran	Universiti Teknikal Malaysia Melaka, Malaysia
Olgierd Unold	Wroclaw University of Technology, Poland
Oscar Castillo	Tijuana Institute of Technology, Mexico
Ovidio Salvetti	ISTI-CNR, Italy
Ozgur Koray Sahingoz	Turkish Air Force Academy, Turkey
Pablo Villacorta	University of Granada, Spain
Patrick Siarry	Universit de Paris, France
Paulo Carrasco	Universidade do Algarve, Portugal
Paulo Moura Oliveira	University of Tras-os-Montes and Alto Douro, Portugal
Pedro Gonzalez	University of Jaen, Spain
Philip Samuel	Cochin University of Science and Technology, India
Pierre-Francois Marteau	Universite de Bretagne Sud, France
Pietro Ducange	University of Pisa, Italy
Punam Bedi	University of Delhi, India
Qieshi Zhang	Waseda University, Japan
Qinghan Xiao	Defence R&D Canada, Canada
Radu-Codrut David	Politehnica University of Timisoara, Romania
Rafael Bello	Universidad Central de Las Villas, Cuba
Ramin Halavati	Sharif University of Technology, Iran

Ramiro Barbosa	Instituto Superior de Engenharia do Porto, Portugal
Ramon Sagarna	University of Birmingham, United Kingdom
Richard Jensen	Aberystwyth University, United Kingdom
Robert Berwick	Massachusetts Institute of Technology, United States
Roberto Armenise	Poste Italiane, Italy
Robiah Yusof	Universiti Teknikal Malaysia Melaka, Malaysia
Roman Neruda	Institute of Computer Science, Czech Republic
S. Ramakrishnan	Dr. Mahalingam College of Engineering and Technology, India
Sabrina Ahmad	Universiti Teknikal Malaysia Melaka, Malaysia
Sadaaki Miyamoto	University of Tsukuba, Japan
Santi Llobet	Universitat Oberta de Catalunya, Spain
Satrya Fajri Pratama	Universiti Teknikal Malaysia Melaka, Malaysia
Saurav Karmakar	Georgia State University, United States
Sazalinsyah Razali	Universiti Teknikal Malaysia Melaka, Malaysia
Sebastian Ventura	University of Cordoba, Spain
Selva Rivera	Institute of C.E.D.I.A.C, Argentina
Shang-Ming Zhou	University of Wales Swansea, United Kingdom
Shyue-Liang Wang	National University of Kaohsiung, Taiwan
Siby Abraham	University of Mumbai, India
Silvia Poles	EnginSoft, Italy
Silvio Bortoleto	Federal University of Rio de Janeiro, Brazil
Siti Rahayu Selamat	Universiti Teknikal Malaysia Melaka, Malaysia
Steven Guan	Xi'an Jiaotong-Liverpool University, China
Sung-Bae Cho	Yonsei University, Korea
Swati V. Chande	International School of Informatics and Management, India
Sylvain Piechowiak	Universite de Valenciennes et du Hainaut-Cambresis, France
Takashi Hasuike	Osaka University, Japan
Teresa Ludermir	Federal University of Pernambuco, Brazil
Thomas Hanne	University of Applied Sciences Northwestern Switzerland, Switzerland
Tsu-Yang Wu	Harbin Institute of Technology Shenzhen Graduate School, China
Tzung-Pei Hong	National University of Kaohsiung, Taiwan
Vaclav Snasel	VŠB Technical University of Ostrava, Czech Republic
Valentina Colla	Scuola Superiore Sant'Anna, Italy
Victor Hugo Menendez Dominguez	Universidad Autonoma de Yucatan, Mexico

Contents

Part III: Smart Living Technology

Part IV: Signal Recognition and Image Processing

Part V: Computational Systems and Its Applications

Part VI: Database and Visualization Technologies

Part VII: Computer Networks and Mobile Computing

Part VIII: Multimedia Signal Processing and Classification

Part IX: Intelligent Data Analysis and System Reliability Modeling

Part I

Innovative Computing Technology for Applications

High Availability and High Scalability
to in-Cloud Enterprise Resource Planning System

Bao Rong Chang[1,*], Hsiu-Fen Tsai[2], Ju-Chun Cheng[1], and Yun-Che Tsai[1]

[1] Department of Computer Science and Information Engineering,
National University of Kaohsiung 81148, Taiwan
`{brchang,10985504,m1025512}@nuk.edu.tw`
[2] Department of Marketing Management, Shu Te University, Kaohsiung 82445, Taiwan
`soenfen@stu.edu.tw`

Abstract. In this paper the host system architecture with high availability and high scalability has introduced to an in-cloud Enterprise Resources Planning (in-cloud ERP) deployed in the virtual environment to tackle the crucial problem of unexpected down-time or the failure of system failover that causes data loss and system terminated. Access control authentication has been adopted in the cloud to prevent the service-oriented hosts form external fraud or intrusion. As a result, the experiments have shown that the number of access for in-cloud ERP is 5.2 times as many as in-house ERP. The total cost of in-cloud ERP has decreased significantly to 48.4% of total cost of in-house ERP. In terms of operational speed, the approach proposed in this paper outperforms significantly two well-known benchmark ERP systems, in-house ECC 6.0 and in-cloud ByDesign.

Keywords: in-Cloud Enterprise Resources Planning (in-Cloud ERP), High Availability, High Scalability, Biometrics, Fraud.

1   Introduction

In this paper we introduce high availability and high scalability architecture for an in-cloud services solution to tickle the crucial problem of unexpected down-time or system failure to prevent data loss and system termination, as well as make good use of virtual machine cluster [1][2][3] approach to resolve the failover problem. This paper introduces in-cloud Enterprise Resources Planning (ERP) [4][5] in virtual environment and mobile devices users can easily access the in-cloud services via wired or wireless network with access control authentication [6]. As shown in Fig. 1, a open source ERP, OpenERP [7], has deployed successfully. Additionally, its access control authentication [8] [9] has brought into the virtual machine to achieve identity verification, safe sign-in, and attendance audit, as shown in Figs. 2 and 3. Then, detecting potential BotNet [10] and malicious attacks [11] in the network can efficiently increase the network security.

---

[*] Corresponding author.
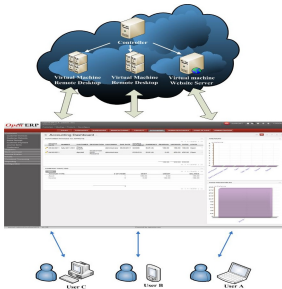
Fig. 1. OpenERP deployment
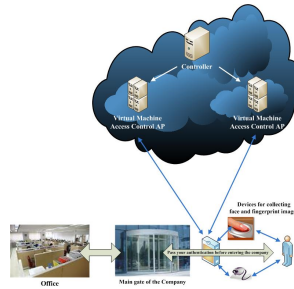
Fig. 2. Access control in a firm
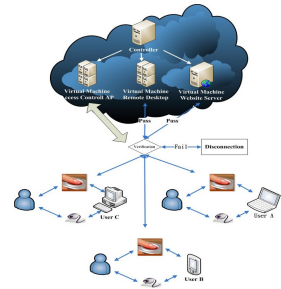
Fig. 3. Access control authentication in cloud

2 In-Cloud ERP and Access Control Authentication

Virtual machine clustering system in cloud is an integration of virtualization, virtual machines, and virtual services so that it can make existing resources be fully applied, such as VMware ESX/ESXi Server [12], Microsoft Hyper-V R2 [12] or Proxmox Virtual Environment [13]. This system can let users run many operating systems in a single physical computer simultaneously which largely decreases the expense of purchasing PCs. Most important of all, it has the following major functions, including virtual machine live migration, virtual storage live migration, distributed resource scheduling, high availability, fault tolerance, backup and disaster recovery, the transfer from physical machines to virtual machines, direct hardware accessing, virtual network switching, and so forth. This study introduces Proxmox Virtual Environment as the cloud computing and service platform with the virtual environment. The kernel-based virtual machine (KVM) acts as the main core of virtual machine, and it has installed the kernel of Linux-based operating system. OpenERP is adopted in this study as ERP application which provides many solutions for open sources software in the future, having it more expandable, making a great progress on cost deduction. The in-cloud ERP is established as follows:

1. Build Proxmox VE virtual machine cluster, and through WebPages manage the virtual machine. The webpage of login and management are shown individually in Fig. 4 and Fig. 5.
2. Create a virtual machine and set up its guest operating system in Proxmox VE virtual machine cluster.
3. Set up OpenERP in virtual machine, inclusive of OpenERP AP, PostgreSQL database, and web interface for end-user.
4. Sign-in at http://localhost:8096 or http://IP:8096 with the browser on virtual machine, pop up a login page of OpenERP as shown in Fig. 6, and then login to administrator to install the necessary modules as a result of an interface of user management as shown in Fig. 7.
5. Set up AP Server for biometric measures security [14]. When users sign-in, it will collect users' biometric features with capturing devices at client side as the evidence of legal or illegal sign-in [15].
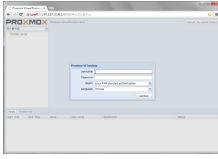
Fig. 4. Login to Proxmox VE Virtual machine cluster server
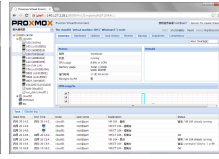
Fig. 5. Management web of virtual machine cluster server

Fig. 6. Remote login to OpenERP system

Fig. 7. User management in OpenERP

6. Collect the traffic flow in network and apply association rule analysis for establishing an identifiable and detectable system to automatically distinguish the network traffic flow either coming from potential BotNet or normal condition [16], as shown in Fig. 8.
7. Employ analysis of Multivariate Normal Model and Markov Chain to detect malicious attacks [17], as shown in Fig. 9.
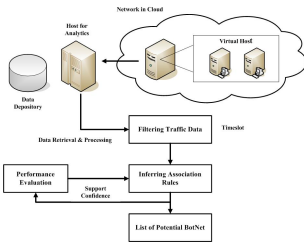
Fig. 8. Potential BotNet detection in cloud
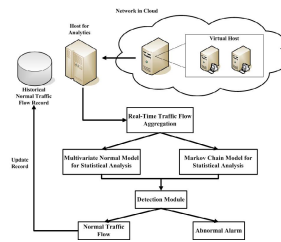
Fig. 9. Malicious attacks detection in cloud

3   High Availability and High Scalability Architecture

3.1   Virtual Machine High Availability

1. **Virtual Machine Live Migration:** when an execution error occurs at a node and causes an interruption, virtual machines at that node can be migrated themselves to the other nodes in which the left tasks of the failure node are also to be continued herein. A prerequisite is to ask for a shared storage as well as two units or more servers, for example, a Proxmox VE system as shown in Fig. 10.
2. **Virtual Storage Live Migration:** the system provides HA in virtual machines and accordingly HA will also support to virtual storage as well. Generally connecting a shared storage (e.g., SAN), the system may achieve the purpose of reaching a low downtime. When an execution error occurs at a node and causes an interruption, virtual storage at that node can be migrated itself to the other nodes to resume the left tasks of the failure node.
3. **Distributed Resource Scheduling:** Virtual machine management system such as Hyper-V [17] imports Non-uniform Memory Access (NUMA) mechanism for the
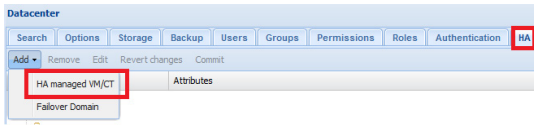
Fig. 10. HA optional setting of VM in Proxmox VE
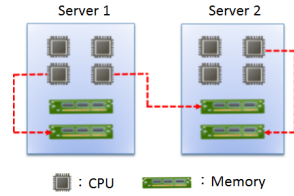
Fig. 11. Hardware resources allocation based on NUMA in Hyper-V R2
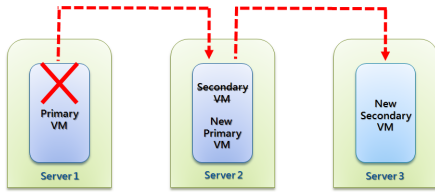
Fig. 12. Fault tolerance mechanism supported by VMware vSphere
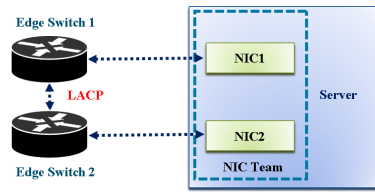
Fig. 13. Realizing the architecture of network high availability

resources allocation, in which computing cores and memory are divided into nodes, and each virtual machine attaches the corresponding node in accordance with the amount of the allocation of resources. That is, the resources of a virtual machine may be allocated from different server hardware resources as shown in Fig. 11.

4. **Fault Tolerance:** The main principle of reaching a zero downtime such as VMware vSphere [16] is that when a primary virtual machine is running, the system automatically generates a redundant virtual machine, totally equal to the primary one, located in other servers to synchronize the task. Once the system detects the primary virtual machine failure, the running task is immediately transferred to the redundant virtual machine, this redundant virtual machine becomes the primary virtual machine at once, and the system will replicate another redundant virtual machine once again as shown in Fig. 12.

3.2   Network High Availability

With Link Aggregation Control Protocol (LACP) [18], network interface cards can utilize Network Bounding techniques that will combine multiple network interface cards together, and in the meantime set the parameters of network interface card related to the HA function. For example, Linux systems can use the software ifenslave to gain fault-tolerant features in the combined network interface cards. That is, as one of network interface cards fails, work load will automatically switch to another one to carry on the successive networking tasks as shown in Fig. 13.

3.3   Storage High Availability

In general, storage device of iSCSI or NAS is able to provide hard drive array (RAID) function. If the system needs to consider both cost and performance, and fault tolerance solution, type of RAID 0+1 disk array is suggested to organize hard drive array. In addition, iSCSI or NAS storage device also probably risks the failure incident and hence the storage device needs to consider HA. At present, the storage device manufacturers have incorporated synchronous backup mechanism, but in contrary the traditional storage devices may not have this feature, which an additional server is required for implementing the synchronization between them as shown in Fig. 14.

According to HA of virtual machine, network, and storage as mentioned above, a diagram of in-cloud platform with high availability is illustrated in Fig. 15. With the minimum facility required for HA structure, the system needs at least two high-performance computing servers, two high-speed network switches, and two high-reliability storages to establish an in-cloud platform with HA.
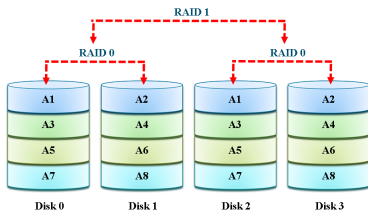
Fig. 14. RAID 0+1 system diagram   **Fig. 15.** an in-cloud platform with high availability
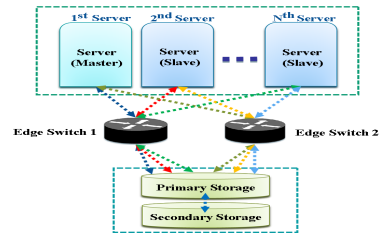
3.4   System High Scalability

In terms of the vertical scalability, virtualization technology can support high availability to ensure hardware maintenance and upgrade/downgrade, and when a single server requires a hardware upgrade/downgrade, virtual machines having the running services in a certain server can live migrate to the others without causing any interruption in these services. Similarly, in horizontal scalability, after virtual machines in a certain server have live migrated to the others, IT manager is able to remove the server from the virtual machine cluster or server cluster, not causing any service interruptions either. The SAN controller of these storage resources to realize so-called storage virtualization [19] could be integrated into different storage pools, which provides the established virtual volumes on demand for virtualized servers. Storage virtualization can more easily integrate heterogeneous storage devices so that changes in the storage devices have a higher elasticity and scalability. As shown in Figs. 16 and 17, they are illustrated to conduct scale-out and scale-in, respectively, in a cluster, and Figs. 18 and 19 have shown the scale-up and scale-down, respectively, of hardware specification in a node.
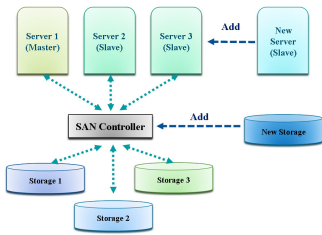
Fig. 16. Scale-out hardware facility
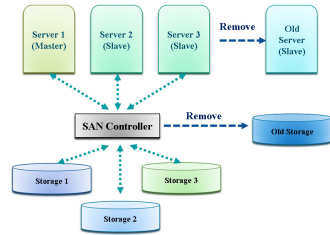
Fig. 17. Scale-in hardware facility
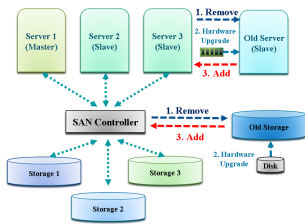
Fig. 18. Scale-up hardware specification
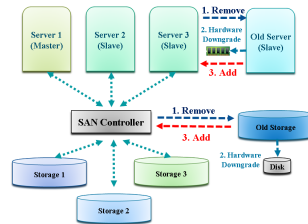
Fig. 19. Scale-down hardware specification

4   Experimental Results and Discussion

4.1   High Availability Testing

First in order to verify the high availability of the network, after the network used the function of Network Bonding, IT manager removed one of the network cables form an edge switch for a few seconds to check whether or not the network satisfies fault tolerance at this situation. After a test of ping command for 50 times, as a result, the connection quality is good because there is no packet loss during the cable removal, achieving the goal of network high availability as shown in Fig. 20. Next in order to verify whether the servers and storage devices achieve high availability, IT manager shut down a server on which a virtual machine was currently running, while the server-mounted storage device will correspondingly fail. Test results show that failover completed successfully because the virtual machine correctly transferred (migrated) to another server as shown in Fig. 21.

4.2   High Scalability Testing

In the virtual environment, the scalability is the most advantage for a cluster with several virtualized servers, and the changes in the scale of hardware devices occur quite often. Therefore the goal of this test in this section will emphasize the non-interrupted services when virtual machine migration is carried out. This test will use two servers denoted as cloud01 and cloud02, IT manager shutdowns server cloud02 normally where the virtual machines running in server cloud02 have migrated to server cloud01 automatically, and removes it from the cluster in order to simulate the

(a) Before VM migration   (b) After VM migration

Fig. 20. Ping command to check the network quality after removing a network cable

Fig. 21. Failover by shutting down a server and initiating a virtual machine migration automatically
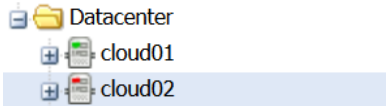
Fig. 22. Shutdown cloud02 in a cluster
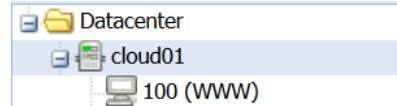
Fig. 23. Remove cloud02 in a cluster
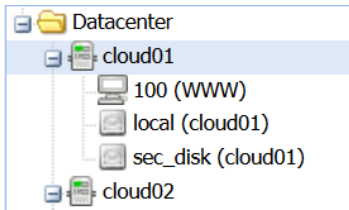
Fig. 24. Add cloud02 to a cluster
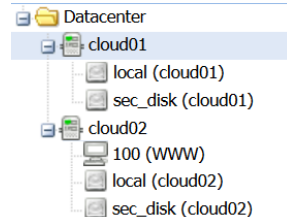
Fig. 25. A VM migrate to cloud02

scale-down of a cluster as shown in Figs. 22 and 23. After that, IT manager re-boot server cloud02 normally and re-join it into the cluster, and IT manager conducts the virtual machines migrated back to server cloud02 from server cloud01 in order to simulate the scale-up of a cluster as shown in Fig. 24 and 25.

4.3   Access Control Authentication and ERP Testing

Users key-in at http://IP:8096 with the browser on Android smart phone to sign-in in-cloud ERP remotely via 3G/WiFi and next based on biometric measures the process of access control authentication is activated to capture human face and fingerprint at mobile device, deliver them to back-end server for identification, and then return the result back to mobile device. It takes about 2 seconds for identity verification as shown in Fig. 26. After that we begin to test ERP routines as shown in Fig. 27. Users
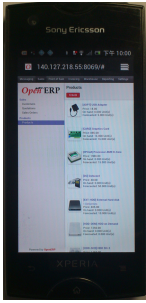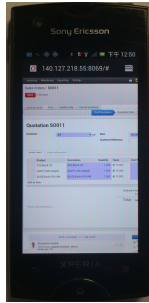
(a) List of products   (b) Sales order
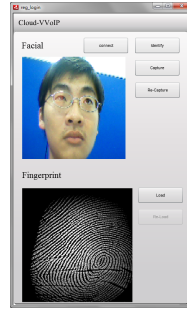
(a) capture the images   (b) identification

Fig. 26. Sign-in in-cloud OpenERP at smart phone

Fig. 27. Access control authentication using face recognition and fingerprint identification at smart phone
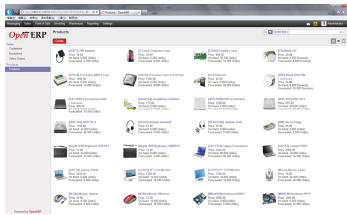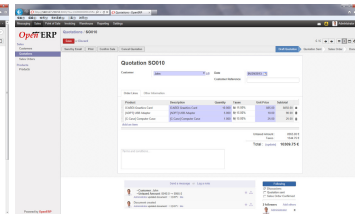
(a) List of products   (b) Sales order

Fig. 28. PC login to in-cloud OpenERP system

key-in at http://IP:8096 with the browser on personal computer to sign-in in-cloud ERP remotely via 3G/WiFi and then go for access control authentication at PC. After that we begin to test ERP routines as shown in Fig. 28.

4.4   Detection of Potential Botnet and Malicious Attacks

In cloud computing platform, Apriori algorithm based on association rules has been executed all the time. When invaders use a BotNet to link to another host or a single host location, the IP pairs between users to a BotNet and a BotNet to an attacked host will coincide more often, and thus the dependence between two IP pairs will be pretty high. As a result, the statistical outcomes derived from Apriori algorithm are fully collected and highlights potential BotNets from multiple period detection. According to the flow of malicious attacks and the relevant features of attack events, introducing the Multivariate Normal Model and Markov Chain is utilized to analyze two above-mentioned situations respectively and then developing a module with the ability of detection and identification is able to combine the two above statistical outcomes. When confronting different types of malicious attacks, it will be more efficient to identify possible causes, even finding the ways for immediate defense.

4.5   Assessment and Discussion

According to the experiments of online testing in the daily use of ERP in enterprise within a week, it was found that the growth rate of the use of in-cloud ERP increased dramatically approximate 5.2 times than the stand-alone ERP. In terms of the hardware cost in Taiwan, it costs the user $1,002.5 on the hardware equipment for a stand alone ERP, i.e. in-house ERP, in which the additional cost will be paid for air conditioning monthly fee of $18.4, space rent of $26.7, and hardware equipment maintenance fee of $16.7. In regard with the amortization expensive per month for a period of two years, the total expenditure costs $2,486.3. In other words, it costs an average monthly usage fee of $103.6. In contrast, renting an in-cloud ERP service in virtual environment only need about $50.1 monthly payment and it saves 1.07 times the cost of in-house ERP, i.e., reducing the total expenditure a lot. As shown in Table 1, a comparison of the number of accesses and the total expenditure for ERP, the proposed in-cloud ERP exclusively superior to in-house ERP. Two well-known benchmark ERP systems, ECC 6.0 [20] and ByDesign [21], are used to compare with the proposed one, according to ERP performance evaluation on system operational speed where the proposed approach outperforms the others as listed in Table 2.

Table 1. ERP assessment

Testing Item	Case A: in-house ERP	Case B: in-cloud ERP	Ratio of Case B to Case A
Number of Access (times/day)	63	328	5.206
Total Cost of Ownership (US dollars/month)	103.6	50.1	0.484

Table 2. Performance comparison of the operational speed of ERP (Unit: min-minutes, sec-seconds)

Function	ECC 6.0 (in-house ERP)	ByDesign (in-cloud ERP)	OpenERP (in-cloud ERP)
Create New Customer Master Data	7:10 min.	4:40 min.	3 min.
Create New Material Master	12:40 min.	10 min.	8:30 min.
Create Sales Order	5:20 min.	2 min.	1:30 min.
Search Function	2:10 min.	5 sec.	2 sec.
Average Access	6.83 min	4.19 min	3.26 min

5   Conclusion

This paper introduces high availability and high scalability architecture for an in-cloud ERP solution to tickle the crucial problem of unexpected down-time or system failure to prevent data loss and system termination, as well as make good use of virtual machine cluster approach to resolve the failover problem. As a result,

according to the experiments the proposed approach in this paper outperforms two well-known benchmark ERP systems, in-house ECC 6.0 and in-cloud ByDesign.

References

1. Beloglazov, A., Buyya, R.: Energy Efficient Allocation of Virtual Machines in Cloud Data Centers. In: 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, pp. 577–578 (2010)
2. Laurikainen, R., Laitinen, J., Lehtovuori, P., Nurminen, J.K.: Improving the Efficiency of Deploying Virtual Machines in a Cloud Environment. In: 2012 International Conference on Cloud and Service Computing, pp. 232–239 (2012)
3. Sotiriadis, S., Bessis, N., Xhafa, F., Antonopoulos, N.: Cloud Virtual Machine Scheduling: Modelling the Cloud Virtual Machine Instantiation. In: 2012 Sixth International Conference on Complex, Intelligent and Software Intensive Systems, pp. 233–240 (2012)
4. Yang, T.-S., Choi, J., Zheng, X., Sun, Y.-H., Ouyang, C.-S., Huang, Y.-X.: Research of Enterprise Resource Planning in a Specific Enterprise. In: 2006 IEEE International Conference on Systems, Man, and Cybernetics, pp. 418–422 (2006)
5. de Carvalho, R.A., Monnerat, R.M., Sun, Y.-H., Ouyang, C.-S., Huang, Y.-X.: Development Support Tools for Enterprise Resource Planning. IT Professional Magazine, 39–45 (2008)
6. Wu, H.-Q., Ding, Y., Winer, C., Yao, L.: Network Security for Virtual Machine in Cloud Computing. In: 2010 5th International Conference on Computer Sciences and Convergence Information Technology, pp. 18–21 (2010)
7. OpenERP, http://v6.openerp.com/
8. Zhao, J.-G., Liu, J.-C., Fan, J.-J., Di, J.-X.: The Security Research of Network Access Control System. In: 2010 First ACIS International Symposium on Cryptography and Network Security, Data Mining and Knowledge Discovery. E-Commerce & Its Applications and Embedded Systems, pp. 283–288 (2010)
9. Metz, C.: AAA protocols: Authentication, Authorization, and Accounting for the Internet. IEEE Internet Computing 3(6), 75–79 (1999)
10. Zhang, L.-F., Persaud, A.G., Johnson, A., Yong, G.: Detection of Stepping Stone Attack Under Delay and Chaff Perturbations. In: 25th Annual International Performance, Computing, and Communications Conference, pp. 256–266 (2006)
11. Yang, H.-Y., Xie, L.-X., Xie, F.: A New Approach to Network Anomaly Attack Detection. Fuzzy Systems and Knowledge Discovery 4, 317–321 (2008)
12. Chang, B.R., Tsai, H.-F., Chen, C.-M.: Evaluation of Virtual Machine Performance and Virtualized Consolidation Ratio in Cloud Computing System. Journal of Information Hiding and Multimedia Signal Processing 4(3), 192–200 (2013)
13. Chang, B.A., Tsai, H.-F., Chen, C.-M.: Empirical Analysis of Server Consolidation and Desktop Virtualization in Cloud Computing. Mathematical Problems in Engineering 2013, Article ID 947234, 11 pages (2013)
14. Wayman, J.L.: Biometrics in Identity Management Systems. IEEE Security & Privacy 6(2), 30–37 (2008)

15. Chang, B.R., Huang, C.-F., Tsai, H.-F., Lin, Z.-Y.: Rapid Access Control on Ubuntu Cloud Computing with Facial Recognition and Fingerprint Identification. Journal of Information Hiding and Multimedia Signal Processing 3(2), 176–190 (2012)
16. Hsiao, H.W., Fan, W.C.: Detecting step-stone with Network Traffic Mining Approach. In: The Fourth International Conference on Innovative Computing, Information and Control, pp. 1176–1179 (2009)
17. Hsiao, H.W., Lin, C.S., Chang, S.Y.: Constructing an ARP Attack Detection System with SNMP traffic data mining. In: The 11th International Conference on Electronic Commerce, pp. 341–345 (2009)
18. Link Aggregation Control Protocol, http://www.ieee802.org/3/ad/public/mar99/seaman_1_0399.pdf
19. Tate, J.: IBM System Storage SAN Volume Controller (IBM Redbooks). Vervante, New York (2006)
20. Doedt, M., Steffen, B.: Requirement-Driven Evaluation of Remote ERP-System Solutions: A Service-oriented Perspective. In: 34th IEEE Software Engineering Workshop, pp. 57–66 (2011)
21. Elragal, A., Kommos, M.E.: In-House versus In-Cloud ERP Systems: A Comparative Study. Journal of Enterprise Resource Planning Studies 2012, Article ID 659957, 13 pages (2012)

Prostate Tumor Identification in Ultrasound Images

Chuan-Yu Chang[1], Meng-Yu Tu[1], and Yuh-Shyan Tsai[2]

[1] Department of Computer Science and Information Engineering,
National Yunlin University of Science and Technology, Yunlin, Taiwan
[2] Department of Urology, National Cheng Kung University Hospital, Tainan, Taiwan
`chuanyu@yuntech.edu.tw`

Abstract. There are various medical imaging instruments used for diagnosing prostatic diseases. Ultrasound imaging is the most widely used tool in clinical diagnosis. Urologist outlines the prostate and diagnoses lesions based on his/her experiences. This diagnostic process is subjective and heuristic. Active contour model (ACM) has been successfully applied to outline the prostate contour. However, application of ACM in outlining the contour needs to give the initial contour points manually. In this paper, an automatic prostate tumor identification system is proposed. The sequential floating forward selection (SFFS) is applied to select significant features. A support vector machine (SVM) with radial basis kernel function is used for prostate tumor identification. Experimental results showed that the proposed method achieved higher accuracy than those of other methods.

Keywords: prostate tumor, feature selection, support vector machine.

1 Introduction

The prostate is a compound tubuloalveolar exocrine gland of the male reproductive system in most mammals. Prostate cancer has been the fifth most common malignancy and the seventh leading cause of cancer deaths in Taiwan. More than 3,000 men are diagnosed annually with prostate cancer, of which approximately 1000 die. Prostate cancer has become a common male urinary tract cancer.

Prostate is located in the pelvis, below the bladder and before the rectum. Most of the early prostate cancers have no symptoms, it's often discovered during routine health examination. Prostate specific antigen (PSA) blood test and digital rectal examination of the prostate are methods for prostate cancer screening. In order to detect early prostate cancer, physicians usually apply the ultrasound imaging to visualize prostate for diagnosis if either elevated PSA or abnormal Digital Rectal Examination (DRE). However, diagnosis by DRE is very subjective and highly depending on the experiences of urologist.

Various imaging modalities, such as magnetic resonance imaging (MRI), computerized tomography (CT), and ultrasound (US) imaging, are widely used in the diagnosis of various diseases with the assistance of medical image analysis techniques. In which, US imaging is inexpensive, non-invasive, and easy to use. Thus,

ultrasound images become the most commonly used in clinical diagnosis, such as inspecting the prostate[4,6,7,8], breast tumor[3] and thyroid nodule[2].

Physician have outlined prostate region in ultrasound image when clinical diagnosis, and then determine whether a tumor within the region, outlined tumor area for tumor diagnosis. However, outlining the tumor area in ultrasound images is highly relied on physicians' experiences, it's inefficient and subjective. Therefore, we proposed a prostate tumor identification system in ultrasound images is proposed in this paper.

This paper is organized as follows: Section 2 presents the proposed prostate tumor identification method. Experimental results are presented in Section 3, where we compared with other method. Conclusions are drawn in Section 4.

2   Proposed Method

In this paper, an automatic SVM-based prostate tumor identification in ultrasound images is proposed. First, a preprocessing is adopted to reduce the influence of speckle noise and therefore enhance the contrast of prostate region. Second, a total of 236 features are extracted from each ROI. Third, the sequential floating forward selection (SFFS) is applied to select significant features. Finally, a support vector machine (SVM) with radial basis kernel function is used for prostate tumor identification. Figure 1 shows the flow diagram of the proposed method. Details of the procedures are described in the following subsections.
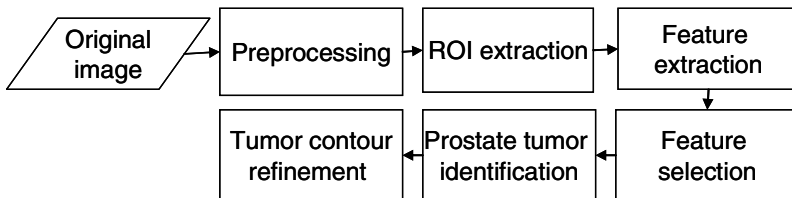
Fig. 1. Flow chart of the proposed method

2.1   Preprocessing

Since speckle noise and poor contrast in ultrasound images will affect the identification accuracy of the prostate tumor. Hence, speckle reduction by means of digital image processing should improve image quality and possibly the diagnostic potential of medical ultrasound.

In this paper, a 5×5 average filter [7] is used to reduce inevitable speckle noise in the prostate ultrasound images. A power-law [7] transformation is then applied to enhance the contrast; parameters c and r of power-law transformation are set as 1 and 0.7, respectively. Figure 2 shows the result of preprocessing. After preprocessing we can clearly see that the contrast between the prostate and background was strengthened.
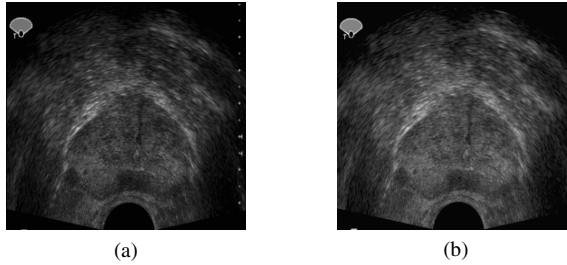
Fig. 2. (a) original image, (b) result after preprocessing

2.2 ROI Extraction

In order to identify the prostate tumor in the ultrasonic images, a lot of ROIs includes prostate tumor and non-tumor were outlined by an experienced urologist and confirmed by biopsy. Based on the sliding-windows method, a block with size of 23×23 is sliding on the outlined ROIs to segment patterns of tumor and non-tumor. The blocks are arrayed adjacent to each other with an $M \times M$ overlap. Figure 3 shows a case of outlined prostate tumor in ultrasound image. The extracted normal and tumor patterns are showed in Fig. 3(b) and (c), respectively.
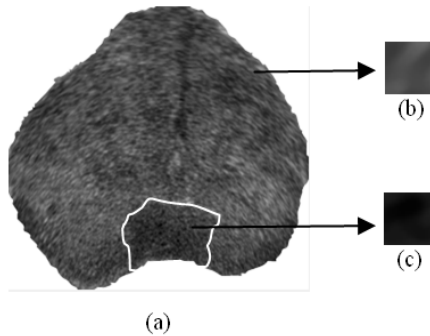
Fig. 3. (a) prostate region, (b) normal prostate pattern, (c) prostate tumor pattern

2.3 Feature Extraction

Textural features contain important information that is used for analysis and explanation of US images. In this paper, 236 features were extracted from the selected ROIs. Features are calculated from 11 different textural matrixes and transformations namely graylevel co-occurrence matrix (GLCM), Statistical Feature Matrix, Gray Level Run-Length Matrix, Laws' Texture Energy Measures, Neighboring Gray Level Dependence Matrix, Haar wavelet, homogeneity, histogram, block difference of inverse probabilities (BDIP) [10], Discrete Cosine Transform, and normalized multi-scale intensity difference (NMSID) [11].

2.4   Feature Selection

In previous stages, 236 features are obtained from each ROI. We can simply use them to identify prostate tumor, but it will be time-consuming. Feature selection is a common way to improve this situation. The sequential floating forward selection (SFFS) is adopted to sift significant features. The SFFS method consists of Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS), which are capable of correcting wrong inclusions and removal decisions. Figure 4 shows the flow diagram of SFFS.
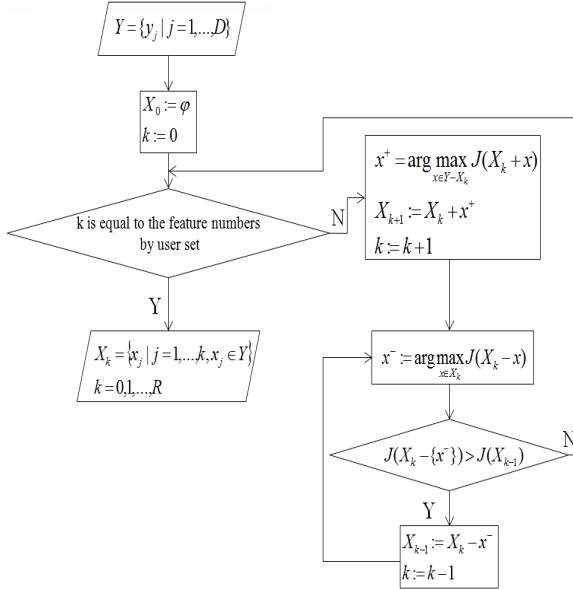
Fig. 4. The flow diagram of SFFS

In which, Y is features for all groups, D is the number of feature item, X is the best feature set selected from Y, R is the numbers of X, $J(x)$ is the cost function that evaluate the classification accuracy of feature set X. The terminative condition of the SFFS algorithm is the number of features k. The SFFS algorithm is summarized as follows:

Step 1: Input a feature group Y and set the terminative condition k.

Step 2: Use the function $J(x)$ to evaluated Y, that we can obtained a best feature x^+, then put the x^+ into the best feature group X.

Step 3: Use the function $J(x)$ to evaluated X to obtain a best feature x^-.

Step 4: Compared the best feature set without x^- ($J(X_k - x^-)$) and the best feature set of previous, if $J(X_k - x^-)$ is bigger than the best feature set of previous, then the best feature set of previous is equal to the best feature set without x^-, and k minus one then return to step 3. If $J(X_k - x^-)$ is not bigger than the best feature set of previous, then return to step 2 until reached the stop condition.

After the SFFS, five representative features were selected. Among them, four features are extracted from GLCM and one is extracted from textural energy matrix. The details of the selected features are described as follows:

The sum average of gray level co-occurrence matrix (GLCM) is defined as:

$$Sum\ Average = \sum_{i=2}^{2L} i p_{x+y}(i) \cdot \tag{1}$$

$$p_{x+y}(k) = \sum_{\substack{i=0 \\ }}^{L} \sum_{\substack{j=0 \\ i+j=k}}^{L} p(i,j), \quad k = 2,3,...,2L \cdot \tag{2}$$

where L is the largest grey value in the image.

The contrast of gray level co-occurrence matrix (GLCM) is defined as:

$$Contrast = \sum_{n=0}^{L} n^2 \left(\sum_{\substack{i=0 \\ }}^{L} \sum_{\substack{j=0 \\ |i-j|=n}}^{L} p(i,j) \right) \cdot \tag{3}$$

$p(i,j)$ is the values from gray level co-occurrence matrix after normalized, it defined as:

$$p(i,j) = \frac{C(i,j)}{\sum_{i,j} C(i,j)} \cdot \tag{4}$$

where $C(i,j)$ is the values of gray level co-occurrence matrix coordinate (i,j).

The difference entropy of gray level co-occurrence matrix(GLCM) is defined as:

$$Difference\ Entropy = -\sum_{i=0}^{L} p_{x-y}(i) \log\left(p_{x-y}(i)\right) \cdot \tag{5}$$

$$p_{x-y}(k) = \sum_{\substack{i=0 \\ }}^{L} \sum_{\substack{j=0 \\ |i-j|=k}}^{L} p(i,j), \quad k = 0,1,...,L \cdot \tag{6}$$

The difference variance of gray level co-occurrence matrix (GLCM) is defined as:

$$Difference Variance = \sum_{i=0}^{L} \left(i - \sum_{j=0}^{L} j p_{x-y}(j) \right)^2 p_{x-y}(i) \cdot \tag{7}$$

In the present study, Law's texture energy [5] computed through the following two masks are used to obtain the texture features:

L5 (Level) = [1 4 6 4 1]
S5 (Spot) = [-1 0 2 0 -1]

The L5 vector gives a center-weighted local average. The S5 vector detects spots. In this study, a 2D convolution masks, namely S5L5 is obtained by computing the

outer products of pairs of 1D masks. By convoluting the image with the 2D mask, a texture energy maps are acquired. The statistics mean of the texture energy maps is then calculated and used as the texture feature.

2.5   Prostate Tumor Identification

Support vector machine (SVM) is a supervised classifier and it has been widely used in regression analysis and statistical classification. The basic idea of SVM is to map the input data into a high-dimensional feature space, and find the hyperplane that maximizes the margin between two classes. Fig. 5 shows the schematic of hyperplane. A SVM is used to identify the prostate tumor. In this paper, the SVM was implemented using the LIBSVM [1], and the radial basis function (RBF) defined as Eq.(8) was selected as the kernel function.

$$K(x^{(s)}, x_i) = \exp\left(\frac{-\left\|x^{(s)} - x_i\right\|^2}{2\sigma^2}\right). \tag{8}$$

where σ^2 is the width of the kernel.
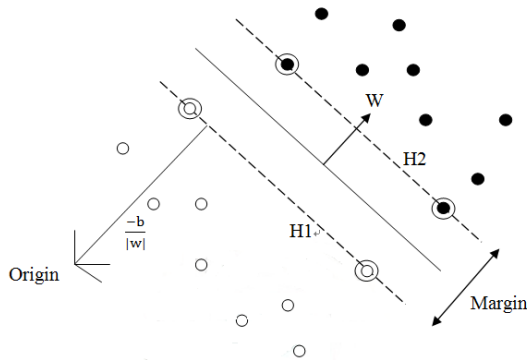
Fig. 5. Schematic concept of hyperplane

2.6   Tumor Contour Refinement

The above procedures can almost segment the prostate tumor completely. However, there still existed some artifact may be classified as part of prostate tumor. A morphological technique is utilized to separate the artifacts from a prostate tumor. Erosion with a 17×17 structuring element is derived to remove the connection between artifacts and prostate tumor. A region filling is then applied to fill holes in tumor region. Finally, dilation with a 17×17 structuring element is performed to recover the original size. The largest region is viewed as the tumor region.

3 Experimental Results

To show the capacity of the proposed method, the prostate ultrasound images for these experiments were taken at the urology department of National Chen Kung University Hospital. Resolution of the image is 3200×2400. A total of 35 cases were used in our experiments. In which, ten cases were chosen as training samples and the remaining cases were chosen as testing samples. The prostate tumor identification system was implemented by Microsoft Visual C# 2008 on a PC with 3.40 GHz Intel Core i7-2600 processor and 4GB RAM.

3.1 Preprocessing

The objective of the preprocessing is to reduce the influence of speckle noise and enhance the contrast of original ultrasound images. Figure 6 shows the results after preprocessing. The left column of Fig. 6 shows original ultrasound images. The results after the preprocessing were showed in the right column of Fig. 6. Obviously, after preprocessing we can clearly see that the contrast between the prostate and background was strengthened.

(a)	(b)	(c)
(d)	(e)	(f)
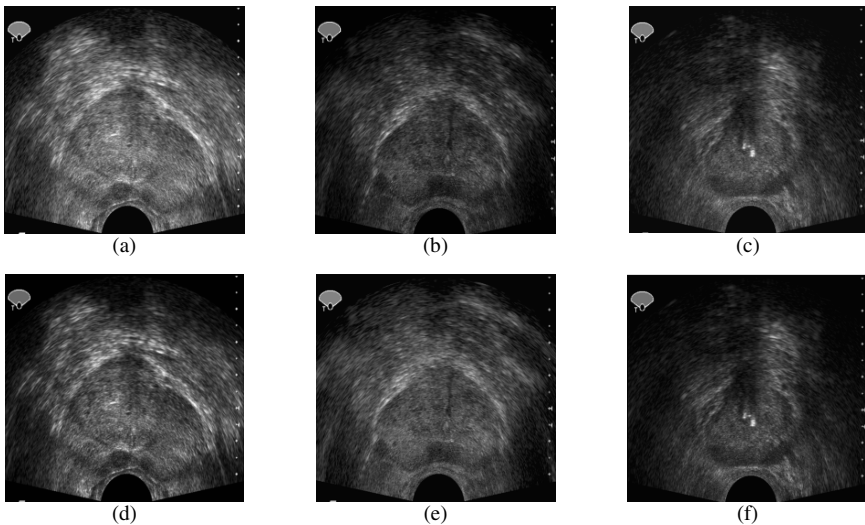
Fig. 6. (a-c) original prostate ultrasound image, (d-f) results after pre-processing

3.2 Prostate Segmentation and Tumor Classification

Before urologists diagnose prostate lesions, the contour of the prostate in US images must be manually outlined. However, manual segmentation is a time-consuming and non-reproducible task. Thus, an automatic prostate segmentation system combines the active contour model (ACM) is adopted [12]. The prostate classifier consists of a validation incremental neural network (VINN) and a radial-basis function neural network (RBFNN). Figure 7 shows the results of outlined the region of prostate in

ultrasound images. Obviously, the prostate region is correctly outlined. Figures 8(a, c, e) show the results of prostate tumor identification. There are many holes existed in prostates. Figures 8 (b, d, f) show the refinement results of Fig. 8 (a, c, e). The holes were removed, and prostate tumors were correctly outlined.

Figure 9 shows the prostate tumor identification results of the proposed method and outlined by urologist manually. Obviously, the outlined tumors of the proposed system are very close to the results outlined by experienced urologist.

(a)   (b)   (c)
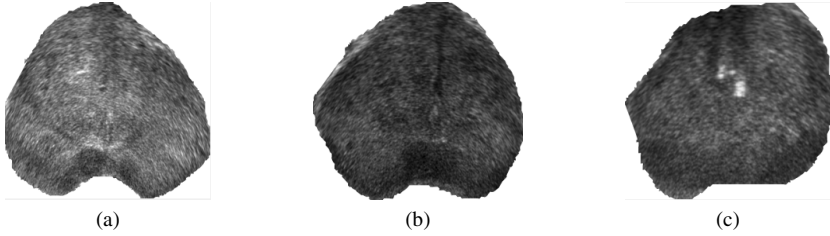
Fig. 7. The outlined prostates in ultrasound images. (a) case1, (b) case2, (c) case3.

(a)   (b)   (c)

(d)   (e)   (f)
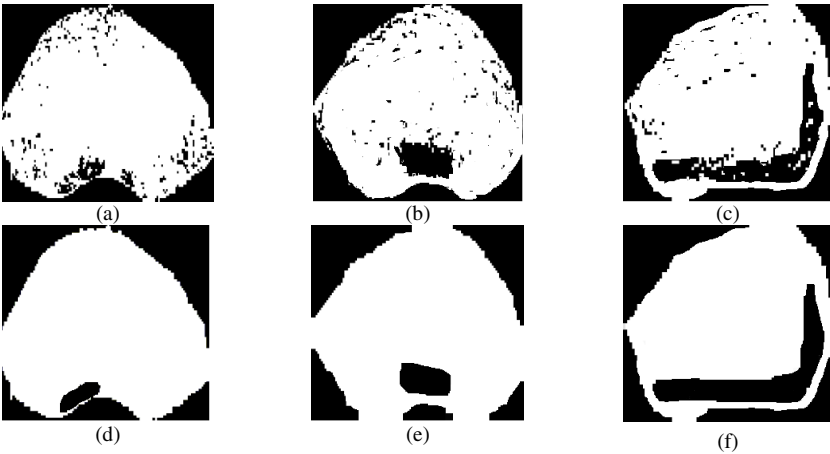
Fig. 8. (a-c) The results after prostate tumor identification. (d-f) The results after prostate tumor refinement.

3.3   Accuracy Evaluation of Prostate Tumor Segmentation

For a fair comparison, the accuracy was used to quantify the performance. Accuracy is defined as follow:

$$Accuracy = (N_{tp} + N_{tn})/(N_p + N_n). \tag{9}$$

where N_p is the total number of pixels of the tumor, N_n is the total number of pixels of non-tumor region, N_{tp} is the number of actual prostate tumor detected by the proposed method, N_{tn} is the number of actual non-tumor pixels detected as
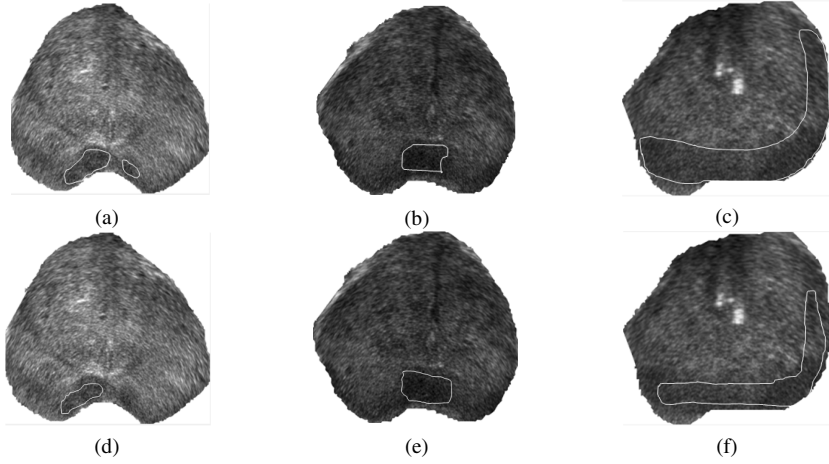
Fig. 9. (a-c)physician outlined tumor, (b-f) outlined tumors of the proposed method

non-tumor pixels, N_{fn} is the number of system classification as the non-tumor region in physician outlined tumor region, N_{fp} is the number of actual non-tumor pixels detected as tumor pixels.

To demonstrate the accuracy of the proposed method is higher than that of Zhang's method [9], Table 1 shows the comparison results. The accuracy of the proposed and Zhang's methods are 91.58% and 90.54%, respectively.

Table 1. Accuracy of Zhang's and proposed method

	The proposed method	Zhang's method
Accuracy	91.58%	90.54%

Table 2. The result for differential parameter

ROI size	Filter size	# of the best feature set	Accuracy
17×17	3×3	6	89.4%
17×17	5×5	5	87.67%
17×17	7×7	6	87.04%
19×19	3×3	7	89.08%
19×19	5×5	6	88.3%
19×19	7×7	6	87.06%
21×21	3×3	4	88.24%
21×21	5×5	5	89.49%
21×21	7×7	6	85.53%
23×23	3×3	4	91.56%
23×23	5×5	5	93.45%
23×23	7×7	6	91.85%
25×25	3×3	4	90.88%
25×25	5×5	5	91.74%
25×25	7×7	5	90.88%

3.4   Determination of ROI Size and Denosing Filter Size

To achieve a high accuracy, the size of ROI, masks, and the number of selected features should be appropriately set in the testing procedure. In order to obtain the appropriate parameters, various ROI sizes and denosing filter sizes were performed. Table 2 shows the accuracies for different parameters. The highest accuracy is achieved when the size of ROI and size of filter is set to 23×23 and 5×5, respectively.

4   Conclusion

In this paper, we proposed a SVM-based method for identification of prostate tumor in ultrasound images. The system can assist urologists on clinical diagnosis. The representative features after feature selection can save lots of time in classification and improve the classification results. According to the experimental results, the average accuracy rate has reach 91.58%. Our experiments have better accuracy than Zhang's method.

References

1. Chang, C., Lin, C.: LIBSVM: a library for support vector machines, 2.91. ed: Citeseer (2001)
2. Maroulis, D.E., Savelonas, M.A., Iakovidis, D.K., Karkanis, S.A., Dimitropoulos, N.: Variable Background Active Contour Model for Computer-Aided Delineation of Nodules in Thyroid Ultrasound Images. IEEE Trans. Infor. Tech. in Biomed. 11, 537–543 (2007)
3. Chen, D.R., Chang, R.F., Wu, W.J., Moon, W.K., Wu, W.L.: 3-D breast ultrasound segmentation using active contour model. Ultrasound in Medicine and Biology 29, 1017–1026 (2003)
4. Sahba, F., Tizhoosh, H.R., Salama, M.M.: Application of Reinforcement Learning for Segmentation of Transrectal Ultrasound Images. BMC Medical Imaging, 1471–2342 (2008)
5. Laws, K.: Textured image segmentation. University of southern california, Los Angeles Image Processing Institute (1980)
6. Betrouni, N., Vermandel, M., Pasquier, D., Maouche, S., Rousseau, J.: Segmentation of Abdominal Ultrasound Images of the Prostate Using a Priori Information and an Adapted Noise Filter. Computerized Medical Imaging and Graphics, 43–51 (2005)
7. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. (1992)
8. Mohamed, S.S., Salama, M.M.: Spectral Clustering for TRUS Images. BioMedical Engineering OnLine (2007)
9. Zhang, Y., Sankar, R., Qian, W.: Boundary delineation in transrectal ultrasound image for prostate cancer. Comp. Biol. Med., 1591–1599 (2007)
10. Chum, Y.D., Seo, S.Y.: Image Retrieval using BDIP and BVLC Moments. IEEE Transactions on Circuits and Systems for Video Tech. 13, 951–957 (2003)
11. Chen, E.L., Chung, P.C., Chen, C.L., Tsai, H.M., Chang, C.I.: An automatic diagnostic system for CT liver image classification. IEEE Trans. on Biomedical Engineering 45, 783–794 (1998)
12. Chang, C.Y., Tsai, Y.S., Wu, I.L.: Integrating Validation Incremental Neural Network and Radial-Basis Function Neural Network for Segmenting Prostate in Ultrasound Images. International Journal of Innovative Computing, Information and Control 7, 3035–3046 (2011)

QoS of Triple Play Services in LTE Networks

Lukas Sevcik, Karel Tomala, Jaroslav Frnda, and Miroslav Voznak

Department of Telecommunications, VSB-Technical University of Ostrava, 17.
listopadu 15, 70833 Czech Republic
`{lukas.sevcik.st1,karel.tomala,`
`jaroslav.frnda,miroslav.voznak}@vsb.cz`

Abstract. This paper deals with studying the effects of performance LTE network throughput for data traffic. Utilisation rate of networks has a significant impact on the quality parameters of triple play services. LTE network was simulated by the software module additionally implemented in a development environment MATLAB. Throughput model has been obtained using this simulation that was used to test the QoS parameters for voice/video. Voice and video data streams using different codecs are transmitted for the obtained throughput. The measured qualitatively QoS parameters determine the resulting quality services from the perspective of end user perception.

Keywords: Quality of service, Long Term Evolution (LTE), E-model, Throughput, packet loss.

1   Introduction

Relative to increasing demands of mobile users for fast transmission of communications, enabling them to efficiently transmit both data and voice, but also multimedia documents, presentations or videos, the issue of QoS (Quality of Service) of multimedia services in modern data networks is topical theme.   With incoming requests from users, directly proportional connected evolution in the development of new technologies for data transmission is through mobile networks from GPRS, through EDGE to contemporary   4G network using LTE technology. Currently, LTE technology is put in test and real traffic, so we decided to take a series of measurements, which showed features of this technology.

2   State of the Art

Network convergence refers to the provision of telephone, video and data communication services within a single network. For VoIP to be a realistic replacement for standard public switched telephone network (PSTN) telephony services, customers need to receive the same quality of voice transmission they receive with basic telephone services—meaning consistently high-quality voice transmissions. Like other real-time applications, VoIP is extremely bandwidth- and delay-sensitive. For VoIP transmissions to be intelligible to the receiver, voice packets should not be dropped,

excessively delayed, or suffer various delay (otherwise known as jitter). Evaluation of the quality of transmitted voice incurred delays and packet loss is addressed in the work [2] a [3].

ITU-T G.107 recommendation [9] defines computation model - E-model, which takes into account all the links between transmission parameters. Its output is a scalar labelled R, which is a function of total expected call quality [1]. To evaluate speech quality, MOS (Mean Opinion Score) scale as defined by the ITU-T recommendation P.800 is applied [8].

The traditional method of determining voice quality is to conduct subjective tests with panels of human listeners.   Extensive guidelines are given in ITU-T recommendations P.800/P.830. The results of these tests are averaged to give mean opinion scores (MOS), but such tests are expensive and impractical for testing in the field. Because of this, the ITU recently standardized a new model, Perceptual evaluation of Speech Quality (PESQ), that automatically predicts the quality scores that would be given in a typical subjective test.   This is done by making an intrusive test and processing the test signals through PESQ [6].

Publication [10] deals with quality video and comparing the uploaded video codecs and their impact on the packet loss that causes errors in the transmission. In the work [7] we deal with focus on Packet loss and overall delay influence to objective QoS parameters of Triple play services. In the mentioned work, missing a comprehensive solution triple play services in real deployment; the impact of the policies used for the network and its quality parameters such as delay, jitter. The aim of our work and measurements in it was to analyze the quality of voice/video in various scenarios LTE network at the physical layer. From each scenario was used measured percentage network throughput. Testing the quality of voice/video was therefore focused on examining the impact of network utilization on its quality.

3   Methodology

3.1   Simulation throughput LTE

Technology LTE (Long Term Evolution) follows on HSPA and UMTS systems. It supports variable width of the frequency band ranging from 1.4 MHz to 20 MHz. The theoretical peak data rate in the downlink is 172.8 Mbit/s, in the uplink 57.6 Mbit/s. LTE system uses the principle of communication IP (MIP - Mobile IP). Communication thus takes place exclusively through the packet. The advantage is a low response in the radio interface (theoretically <10 ms) and high spectral efficiency [13] [14]. LTE system allows symmetrical (paired) or unsymmetrical (unpaired) communication. Symmetric communication using different uplink and downlink frequencies, which are mutually distant for the particular zone by a given value. The highest data rate can be achieved only when using symmetrical communication (FDD - Frequency duplex). When asymmetric communication is using to establish a connection to the constant frequency and communication takes place in alternating moments of time, so it is a time duplex (TDD).

In our case, LTE technology was simulated in an interactive environment MATLAB. We focused on the throughput parameter of this technology in certain scenarios set in the simulation of LTE. While simulating, we adjusted the frequency or time duplex (symmetric and asymmetric communication). Another parameter to measure was the ability to set the SNR, a signal-to-noise ratio. It is a parameter that indicates how many times the signal is fully reduced before we get to the noise level. It is expressed in decibels. Another parameter for testing and verification of the radio transmission, which affects throughput of LTE, were various options of the model channel.

3.2   Quality of Voice

During simulation of LTE, we implemented 4 types of codecs to compare quality of voice on the measured throughput of LTE. One type of codec was with two different bit rates. Capacity utilization of the network ensured IPERF tool. It can make capacity utilization to full at a given throughput LTE. We generated five voice traffic for each codec using program IxChariot with such busy network. Packets had not any priority, when voice was sending. Data were processed by FIFO. Parameters MOS, R-factor, the percentage packet loss and the percentage increase of jitter were evaluated after completion of voice traffic. These parameters evaluate toll IxChariot, which sets the parameters based on the E-model algorithm.

To measure quality of call, mathematical computer model, known as the E-model, was developed. It was designated specifically for the purpose of planning the transmission system, which takes into account many factors (particularly the impact of the delay on packet loss), that affect quality of voice in IP telephony. E-factor model assigns coefficient of deterioration to each partial factor that affects the quality of voice communication. Coefficients of deterioration are indicated in the following figure (figure.1), which presents connection between two participants. The output of the E-model is a value that is closer to the quality of the call in modern networks, either narrowband or broadband. It is called the R-factor and it is in the range of 0 to 100.

The R-factor is determined for the entire transmission chain. It takes into account not only the transmission channel, but also the end device. The higher the R-factor, the higher the quality of telephone service. The minimum acceptable value of R-factor is a ranking value 50. The resulting structure of mathematical relation of E-model is defined according to the recommendations by (1):

$$R = R_0 - I_s - I_d - I_{e-eff} + A \tag{1}$$

The R_0 represents the basic signal-to-noise ratio, which includes all kinds of noise, including noise caused by electrical circuits of equipment and noise caused on electrical wire.

I_S is the deterioration factor reflecting linear distortion of the transmission channel. This deterioration factor is the sum of a few individual factors, which may occur more or less simultaneously in conjunction with the transmission of the call.

Factor I_D represents all the deterioration which is caused by different combinations of delay.

I_{E-EFF} includes deterioration caused by the use of a voice codec, the manifestation of potential packet loss, and its resistance to loss. Finally, parameter A slightly improves the total final quality (for example in the case of satellite phone, the user is more focused to call than by normal use of solid terminal at home).

The final R-factor is a numerical value, usually between 15-94. With a value above 70, it is considered as acceptable. Parameter MOS is used as addition to R-factor. Generally known and most widely used is the five-point scale for quality rate. In this scale, it is possible to display display a value of R-factor (transfer described in recommendation G.107) [1].
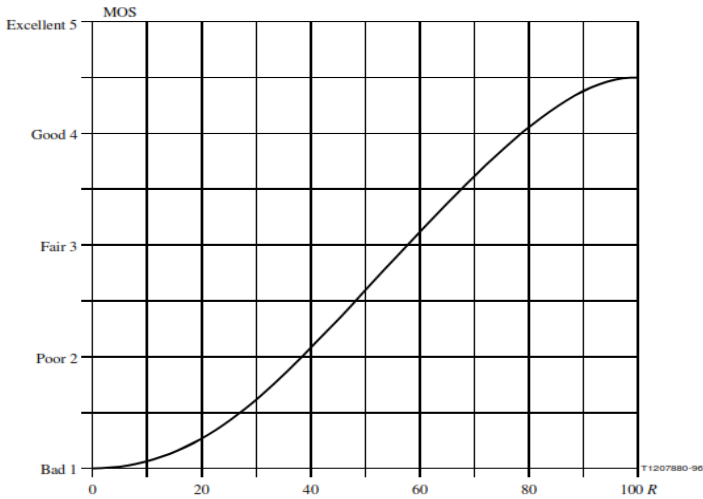
Fig. 1. R-factor as a function of the MOS

We tested four types of codecs in the measurement: G.711 A-law, G.729, G.726 a AMR.

G.711 is the most commonly used voice codec. It isnarrowband audio codec that provides the bit rate of 64 kbit/s. Transmitting audio signals in the range from 300 to 3400 Hz, the sampling rate is 8000 samples per second. A-law encoding takes a 13-bit linear sample as input and converts it into an 8 bit value.

G.729 is the most commonly used voice codec. It is narrowband audio codec that provides the bit rate of 64 kbit/s. Transmitting audio signals in the range from 300 to 3400 Hz, the sampling rate is 8000 samples per second. A-law encoding takes a 13-bit linear sample as input and converts it into an 8 bit value.

G.729 is very effective in comparison with the codec G.711 with relative to the final quality of the call. It uses a modified coding algorithm CELP, called CS-ACELP, bit rate of this codec is 8 kbit/s. Its disadvantage is the higher computational complexity. Total delay in the encoder is about value 25 ms.

G.726 is speech codec that uses the ADPCM coding involving the transmission of voice at rates of 16, 24, 32 and 40 kbit/s. It has size of sample 2, 3, 4 and 5 bits. The most commonly is use transmission rate 32 kbit/s, which doubles the usable capacity of the network by using half speed   of G.711 codec.

3.3   Quality of Video

For measurement was selected two most used video formats for digital broadcasting, specifically MPEG-2 and MPEG-4 Part 10 (H.264). Both formats of video were in HD 720p resolution (1280x720), representing the standard of digital broadcasts at present time. Both formats were processed in a static image "Lenna".

Progam VLC was used to stream of both formats. Linux tool "iperf", simulating a UDP stream, was used for required network load. For the evaluation effects of degradation by network traffic to quality of video were chosen and used two objective methods for assessing the quality of video - PSNR and SSIM. These methods have been implemented by software MSU VQMT.

PSNR method is based on the mean square error (MSE), which is the square deviation between the test and the original sample. PSNR then determines the ratio between the maximum value of the signal to MSE in decibels. Basically, it can be said that while the MSE measures the difference between the two images, PSNR determines how closely the test image is similar to the original reference. Given a noise-free m×n monochrome image I and its noisy approximation K, MSE is defined as (2):

$$MSE = \frac{1}{m \cdot n} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} \left[I\left(i,j\right) - K\left(i,j\right) \right]^2 \tag{2}$$

The PSNR is defined as (3):

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE} \right) \tag{3}$$

$$= 20 \cdot \log_{10}\left(MAX_I \right) - 10 \cdot \log_{10}(MSE)$$

where MAX is the maximum value that pixel can take (e.g. 255 for 8-bit image) [12].

SSIM method takes into account the perception of an image by the naked eye. It evaluate the visual impact of shifts the brightness of the image, change the contrast and other occurrence of errors in the transmitted picture (when compared it with the original image. Reference SSIM values are in the range 0-1, where zero means no similarity with the original and one is two completely identical pictures.

The resulting value of SSIM is a combination of three parameters, when for the original signal x and encoded signal y is valid the following equation (4) [7]:

$$SSIM(x, y) = [l(x, y)]\alpha[c(x, y)]\beta[s(x, y)]\gamma, \tag{4}$$

l(x,y) compares brightness of signal,
c(x,y) compares contrast of signal,
s(x,y) measures the correlation structure,
α > 0, β > 0, γ > 0 controls the weighting of the various parts.

Diagram of the test platform is shown in the following figure:
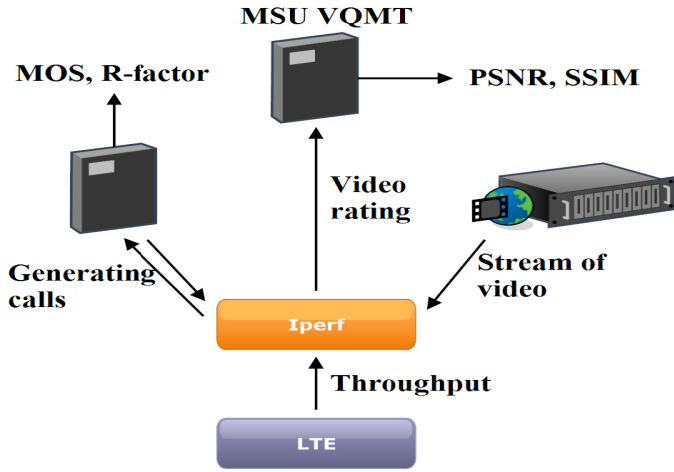
Fig. 2. Scheme of experiment

The simulation model were set through 4 channels:

- EPA (Extended Pedestrian A model): This is a multipath model of propagation of radio waves with pedestrian delay profile.
- EVA (Extended Vehicular A model): This is multipath model of propagation of radio waves with vehicular delay profile.
- ETU (Extended Typical Urban model): multipath model of propagation of radio waves with the town profile delays or wave reflections between buildings.
- HST (High Speed Train condition): It is a model of radio waves with a single component reflects. It represents the Doppler shift, which causes a result of high speed of train moving around the base station.

Some values obtained from tests and measurements are in the following reduced tables. The tables illustrate the relationship between parameters. The graph shows the change in signal-to-noise ratio of throughput at different LTE channel models. The values of signal to noise ratio (SNR) are suitably chosen so that it is possible to measure the throughput of between approximately 15 to 100% throughput.
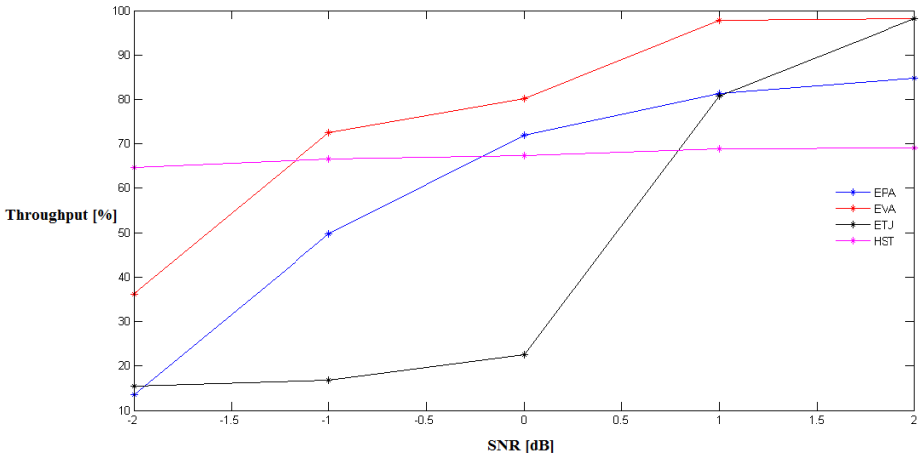
Fig. 3. The dependence of throughput to change SNR - time duplex
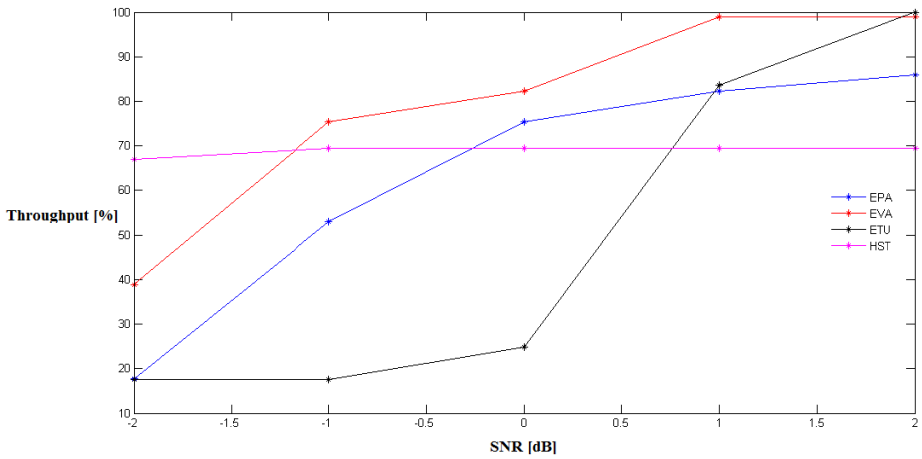
Fig. 4. The dependence of throughput to change SNR –frequencyduplex

Table 1. The quality of voice (codec G.711 A-law) by a throughput LTE with time duplex

Throughput	MOS	R-factor	Packetloss
13.51	3.012	58.616	0.182
22.56	3.362	66.694	1.9598
36.21	3.382	67.416	1.9732
49.73	3.508	69.994	1.4398
69.13	4.24	87.71	1.8534
81.25	4.33	90.22	1.3268
98.23	4.37	91.45	0.1

Table 2. The quality of video by a throughput LTE with time duplex

Throughput [%]	MPEG-2			MPEG-4		
	PSNR [db]	SSIM	PL [%]	PSNR [db]	SSIM	PL [%]
13,51	28,045	0,9396	3,48	17,048	0,7798	8,21
22,56	33,062	0,9533	0,83	19,291	0,8488	2,51
36,21	37,421	0,9557	0,81	21,254	0,8639	1,58
49,73	37,547	0,9562	0,78	21,404	0,8661	1,41
69,13	42,321	0,9658	0,12	24,962	0,9067	0,34
81,25	49,622	0,9827	0,07	38,152	0,9676	0,14
98,23	51,835	0.9897	0	51,2351	0,9848	0

4   Conclusion

Series of measurements were performed to test the characteristics of LTE technology and its impact on the quality of voice and video network. The difference between the time and frequency duplex in LTE had no significant imapact. Size of throughput was reflected in changing SNR and channel model. With increasing signal to noise ratio grew throughput, too. The results of the quality of voice show that at higher network bandwidth, the value of the R-factor increases and thereby the value of the MOS does too. The best results reported codec G.711 A-law at the highest throughput, respectively, it was the worst quality at the lowest throughput. Stable voice quality respectively, the difference between the worst and the best quality at the lowest and highest throughput recorded G.726 and AMR 12.2 kbit/s.

Results for quality of video with the decreasing throughput of the network show a decrease in quality of video and increase the packet loss in the network. Although H.264 provides fold higher compression ratio than the older MPEG-2, there is visible significant drop in quality with the decreasing throughput and packet loss in the network. Practical tests verify the basic hypothesis of lower throughput in the network that automatically decreases the quality of service. Sufficient QoS was achieved with throughput about 50%.

References

1. Voznak, M.: E-model modification for case of cascade codecs arrangement. International Journal of Mathematical Models and Methods in Applied Sciences 5(8), 1439–1447 (2011)

2. Managing Voice Quality with Cisco Voice Manager (CVM) and Telemate, `http://www.cisco.com/en/US/products/sw/voicesw/ps556/products_tech_note09186a00800946f8.shtml`
3. Cole, R.G., Rosenbluth, J.H.: Voice over IP performance monitoring. ACM SIGCOMM Computer Communication (2001)
4. ITU-T P.800.1, Mean Opinion Score (MOS) terminology Geneva, ITU-T Recommendation P.800.1 (July 2006)
5. Karam, J.: A Comprehensive Approach for Speech Related Multimedia Applications. WSEAS Transactions on Signal Processing 6(1) (January 2010)
6. Rix, A.W., Beerends, J.G., Hollier, M.P., Hekstra, A.P.: Perceptual evaluation of Speech Quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs. In: IEEE Signal Processing Society International Conference on Acoustics, Speech, and Signal Processing, ICASSP (May 2001)
7. Frnda, J., Voznak, M., Rozhon, J., Mehic, M.: Prediction Model of QoS for Triple Play Services. In: 21st Telecommunications Forum TELFOR 2013 (2013)
8. ITU-T P.800, Methods for subjective determination of transmission quality, ITU-T Recommendation P.800, Geneva (August 1996)
9. ITU-T G.107, The E-model, a computational model for use in transmission planning, ITU-T Recommendation G.107, ITU-T Geneva (May 2010)
10. Feamster, N., Balakrishnan, H.: Packet Loss Recovery for Streaming Video. In: 12th International Packet Video Workshop, Pittsburgh, PA (2002)
11. Psytechnics Limited: PESQ: An Introduction, `http://www.sageinst.com/downloads/960B/wp_pesq_introduction.pdf`
12. Uhrina, M., Hlubik, J., Vaculik, M.: Correlation Between Objective and Subjective Methods Used for Video Quality Evaluation. Journal Advances in Electrical and Electronic Engineering 11(2) (December 2013) ISSN: 1804-3119
13. Dahlman, E., Parkvall, S., Sköld, J., Beming, P.: HSPA and LTE for Mobile Broadband. Elsevier (2007)
14. Parkvall, S., Dahlman, E., Furuskär, A., Jading, Y., Olsson, M., Wänstedt, S., Zangi, K.: LTE-Advanced – Evolving LTE towards IMT-Advanced. In: Vehicular Technology Conference 2008, pp. 1090–3038 (2008) ISSN: 1090-3038

New Genetic Algorithm
for the p-Median Problem

Pavel Krömer[1,2] and Jan Platoš[2]

[1] Department of Computer and Electrical Engineering,
University of Alberta, Edmonton AB T6G 2V4, Canada
pavel.kromer@ualberta.ca
[2] IT4Innovations & Department of Computer Science
VŠB Technical University of Ostrava
Ostrava, Czech Republic
{pavel.kromer,jan.platos}@vsb.cz

Abstract. The p-median problem is a well-known combinatorial optimization problem with several possible formulations and many practical applications in areas such as operational research and planning. It has been also used as a testbed for heuristic and metaheuristic optimization algorithms. This work proposes a new genetic algorithm for the p-median problem and evaluates it in a series of computational experiments.

Keywords: genetic algorithm, p-median problem, experiments.

1 Introduction

The p-median problem is an NP-hard combinatorial optimization problem with interesting real-world applications in various areas including operational research, planning [4], and clustering [10]. The p-median problem (PMP) can be defined in terms of operational research [9].

Definition 1 (p-median problem). *For a set of m users U, set of n facilities L, and distance matrix $D^{m \times n}$ representing the cost of serving user i from location j find a subset $P \subseteq L$ of exactly p facilities so that the sum of minimal values in the rows of the column-reduced matrix will be minimized:*

$$\min_{P \in \{P \subseteq L \ : \ |P|=p\}} \left\{ \sum_{i \in U} \min_{j \in P} d_{ij} \right\} \tag{1}$$

Alternatively, the PMP can be defined as an integer programming problem or a graph problem. The PMP can be seen as a general combinatorial optimization task of finding an optimal fixed-length subset of p elements out of n, $p < n$. It has a number of real-world applications and it was also used as a testbed for heuristic and metaheuristic optimization algorithms [9].

Intuitively, the problem is challenging due to the large search space of $\binom{n}{k}$ possible solutions. Formally, it is known that the PMP is NP-hard [2,9].

This study uses a novel pure metaheuristic genetic algorithm to solve the PMP and evaluates its performance in a series of computational experiments with PMP instances from the well-known OR-Library[1] of operational research problems.

The rest of this paper is organized in the following way: section 2 briefly introduces genetic algorithms as the metaheuristic optimization method used in this work to find good solutions to the p-median problem. Related work is summarized in section 3. The genetic algorithm proposed in this work is detailed in section 4 and numerical experiments are described in section 5. Section 6 concludes the study and outlines future work.

2   Genetic Algorithms

Genetic algorithms (GA) form a family of well known population-based metaheuristic soft optimization methods [8,1]. GAs solve complex optimization problems by the programmatic evolution of an encoded population of candidate problem solutions. The solutions are ranked using a problem specific fitness function. The artificial evolution is implemented by the iterative application of genetic operators and leads to the discovery of above average solutions. The basic workflow of a simple steady state GA [8] is shown in algorithm 1.

1 Define objective (fitness) function and problem encoding;
2 Encode initial population P of possible solutions as fixed length strings;
3 Evaluate chromosomes in initial population using objective function;
4 while *Termination criteria not satisfied* **do**
5   Apply selection operator to select parent chromosomes for reproduction:
      $sel(P_i) \rightarrow parent1$, $sel(P_i) \rightarrow parent2$;
6   Apply crossover operator on parents with respect to crossover probability to
      produce new chromosomes:
      $cross(pC, parent1, parent2) \rightarrow \{offspring1, offspring2\}$;
7   Apply mutation operator on offspring chromosomes with respect to
      mutation probability: $mut(pM, offspring1) \rightarrow offspring1$,
      $mut(pM, offspring2) \rightarrow offspring2$;
8   Create new population from current population and offspring chromosomes:
      $migrate(offspring1, offsprig2, P_i) \rightarrow P_{i+1}$;
9 end

Algorithm 1. A summary of the genetic algorithm

Problem encoding is an important part of the genetic search. It translates candidate solutions from the problem domain (phenotype) to the encoded search space (genotype) of the algorithm and defines the internal representation of

---

[1] `http://people.brunel.ac.uk/~mastjjb/jeb/orlib/pmedinfo.html`

the problem instances used during the optimization process. The representation specifies chromosome data structure and a decoding function [5]. The data structure defines the actual search space, its size and shape.

The crossover operator is the main operator of genetic algorithms distinguishing it from other population based stochastic search methods [8]. Crossover is primarily a creative force in the evolutionary search process. It is supposed to re-combine parent chromosomes in a stochastic manner and propagate building blocks (low order, low defining-length schemata with above average fitness) from one generation to another and to create new (higher order) building blocks by combining low order building blocks. It is intended to introduce to the population large changes with low disruption of the building blocks [12]. In contrast, mutation is expected to insert new material into the population by random perturbation of the chromosome structure. By doing this, however, new building blocks can be created or old ones disrupted [12].

Genetic algorithms have been successfully used to solve non-trivial multimodal optimization problems including data mining, classification, and prediction problems [13,11]. They inherit the robustness of emulated natural optimization processes and excel in browsing huge, potentially noisy problem domains. This study proposes a new genetic algorithm for the PMP. It defines suitable chromosome encoding, mutation and crossover operators and a fitness function to evaluate candidate PMP solutions.

3   Genetic Algorithms for the p-Median Problem

The p-median problem has been solved by a variety of genetic algorithms in the past [9].

A GA for the PMP is due to Correa et al. [4]. The authors developed a GA for the capacitated PMP in order to find good placement for university facilities with respect to the location and capacity of student residences. In capacitated PMP, each facility can serve only a limited demand. The algorithm used a set of facility indexes to represent selected subset of facilities and modified crossover and mutation operators to evolve the population of candidate solutions. The crossover constructed "exchange vectors" to indicate which alleles can be exchanged between parents. Mutation replaced one facility index by a randomly selected index of another facility that was not part of the solution yet. Additionally, the algorithm used a "heuristic hypermutation" operator that performed local search for good solutions. The proposed algorithm was compared to Tabu Search and the results suggested that the GA with heuristic hypermutation performs better than plain GA and Tabu Search with similar number of fitness function evaluations.

Alp et al. [2] proposed in 2003 another genetic algorithm for the PMP called ADE. The algorithm used specially tailored operators, steady state migration, complex chromosome initialization strategy, and a greedy heuristic to improve solutions found by the artificial evolution. Crossover operator was in ADE based on the union of parents' genes and a greedy method for deletion of extra genes.

No mutation operator was used in the algorithm since the authors reported it had no significant effect on the quality of solutions. The study showed that the proposed algorithm performed well and solved 28 uncapacitated PMP instances from the OR-Library to optimality.

A short paper by Lim and Xu [7] compared the performance of a fixed-length subset GA with and without problem specific heuristic recombination operators on the PMP. As in the previous cases, the GA used set representation of candidate solutions and additional heuristics to improve the solutions. It was concluded that the heuristics improve the quality of solutions found by this type of GA.

Pullan [10] used in 2008 a GA with cut-and-paste crossover operator and hybrid local search to find good PMP solutions. The local search was used to modify candidate solutions found by the evolution to have the required size and to explore the surrounding of evolved solutions. The experiments showed that the algorithm solved all 40 uncapacitated PMP instances from the OR-Library to optimality.

A multiobjective GA for the PMP is due to Arroyo et al. [3]. The algorithm looked for PMP solutions that would minimize both service costs and facility opening costs. Crossover operator of this GA used differences between parents to perform symmetric exchange of genes (whenever possible) and the mutation operator replaced randomly selected facility indexes by different indexes that were not part of the solution yet. The algorithm also used a path relinking strategy to explore search space between best individuals found by the GA in an attempt to improve the evolution.

A more recent work by Landa-Torres et al. [6] solved the capacitated PMP by a new grouping GA and grouping Harmony Search algorithm respectively. Both methods were improved by additional local search steps in order to improve their results. However, solution representation and genetic operators were quite different from those of the previous GAs due to the different nature of grouping algorithms.

This section reviewed recent GAs and GA-based methods for the PMP. A majority of the methods used local search and greedy steps in order to improve solution quality and most of them struggled with the generation of invalid solutions. Moreover, most reviewed works did not discuss computational aspects of chromosome decoding, genetic operator application, and local search steps. In this study we use a new simple GA for the PMP. It uses modified set encoding and only slightly altered crossover and mutation operators. The operations of the algorithm are, in contrast with some previous GAs for PMP, computationally inexpensive. The algorithm does not exploit any form of local search and does not use greedy steps in order to maintain generality. Therefore, it can be used for any other fixed-length subset selection problem when an appropriate fitness function is provided.

4   New Genetic Algorithm for the p-Median Problem

In order to design an efficient genetic algorithm for the PMP, we define the chromosome encoding, genetic operators, and fitness function. The encoding and operators are designed with the aim to obtain a compact representation, to exploit both crossover and mutation, and to avoid the creation of invalid individuals in the course of the evolution.

4.1   Chromosome Encoding

A valid subset of exactly k columns can be defined by the indices of the k columns selected from the full set of n columns in $A^{m \times n}$. No column can appear in the subset more than once in order to achieve the required subset size of k. The natural fixed-length encoding of a column subset is based on the indices of the selected columns, i.e. a chromosome c can be defined by eq. (2):

$$c = (c_1, c_2, \ldots c_k), \quad \forall (i, j) \in \{0, \ldots n\} : c_i \neq c_j \tag{2}$$

However, such an encoding would be prone to the creation of invalid individuals in case traditional genetic operators such as the 1-point crossover or uniform mutation were applied. A few examples of the creation of invalid individuals from valid individuals are given in fig. 1 (conflicting genes are shown in **bold**). To avoid the creation of such invalid individuals, we slightly modify the encoding
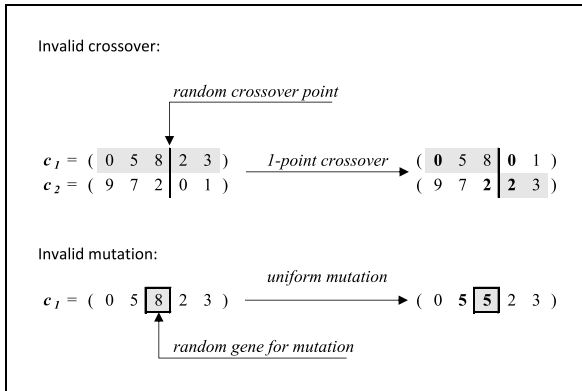
Fig. 1. Examples of the creation of invalid individuals

by sorting the indices within each chromosome, i.e. the encoding of a sample of k columns from n is defined by eq. (3):

$$c = (c_1, c_2, \ldots c_k),$$
$$\forall (i, j) \in \{0, \ldots n\} : i < j \implies c_i < c_j \tag{3}$$

With this modified encoding, the generation of random PMP individuals involves two steps. First, k unique columns out of all n possible columns are selected. Then, the indices in each individual are sorted in an increasing order. However, the ordering of the indices within the chromosome still requires the use of modified mutation and crossover operators that do not generate invalid individuals.

4.2   Genetic Operators

The genetic operators used by the proposed GA are based on the traditional mutation and crossover operators that were modified with respect to chromosome encoding so that they do not create invalid individuals and do not break the ordering of the chromosomes.

Order-Preserving Mutation.   The order-preserving mutation operator replaces ith gene c_i in chromosome $\boldsymbol{c}$ with a value from the interval defined by its left and right neighbour as defined in eq. (4):

$$mut(c_i) = \begin{cases} urand^*(0, c_{i+1}), & \text{if } i = 0 \\ urand(c_{i-1}, c_{i+1}), & \text{if } i \in (0, n-1) \\ urand(c_{i-1}, N), & \text{if } i = N - 1 \end{cases} \tag{4}$$

where $i \in \{0, \ldots, n\}$ and $urand(a, b)$ selects a uniform pseudo-random integer from the interval (a, b) (whereas $urand^*(a, b)$ selects a uniform pseudo-random integer from the interval $[a, b)$). Such mutation operator guarantees that the ordering of the indices within each chromosome remains valid. However, the order-preserving mutation of ith gene has no effect on chromosomes for which it holds that $(c_{i-1} + 1) = c_i = (c_{i+1} - 1)$.

Order-Preserving Crossover.   The order-preserving crossover is based on the traditional one-point crossover operator [1,8]. It selects a random position i in parent chromosomes $\boldsymbol{c1}$ and $\boldsymbol{c2}$ and checks whether it is a suitable crossover point. A position i is a suitable crossover point if eq. (5) is true.

$$c1_i < c2_{i+1} \wedge c2_i < c1_{i+1} \tag{5}$$

If eq. (5) does not hold for i, the remaining positions in the chromosomes are scanned in the search for a suitable crossover point (i.e. first available position for which eq. (5) holds). It should be noted that an order-preserving crossover between 2 chromosomes might not always be possible. Examples of the use of the order-preserving crossover in different situations are shown in fig. 2.

4.3   Fitness Function

The fitness function used in this work is based on the definition of uncapacitated PMP in (1). Fitness of chromosome $\boldsymbol{c}$ is defined by eq. (6):

$$fit(\boldsymbol{c}) = \sum_{i \in U} \min_{j \in \boldsymbol{c}} d_{ij} \tag{6}$$
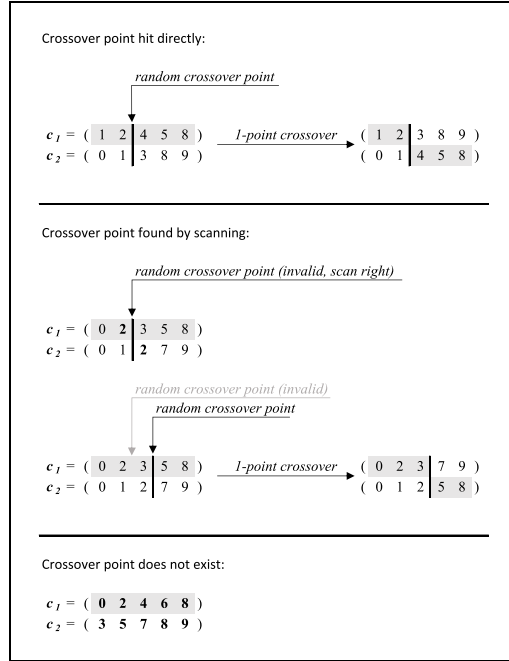
Fig. 2. Examples of order-preserving crossover

The GA for PMP minimizes $fit(\boldsymbol{c})$.

5   Experiments

In order to evaluate the proposed algorithm, a series of computational experiments was performed. The algorithm was implemented in C++ and used to find solutions to 40 uncapacitated PMP instances from the OR-Library. GA parameters used in the experiment were selected on the basis on initial trials and previous experience. The parameters were: steady state migration with generation gap 2 and elitism (only offspring with fitness better than worst member of current population are accepted), population size 100, crossover probability 0.8, mutation probability 0.4, and max. no. of generation 150000. Due to the stochastic nature of GA, all experiments were performed 30 times independently and presented results are averages over the 30 runs.

Overall, the average error of best solutions for all PMP instances found by the GA was 3.231%, the average error of average solutions was 5.46975% and the average error of worst solutions was 8.1845%. Results for different PMP instances are detailed in table 1. Names of the instances that were solved to optimality by at least one GA run are *emphasized*. It can be seen that the GA managed to find optimal solutions for 7 out of 40 instances in at least one run. This result is not as good as the results of previous genetic methods for PMP. However, it must be

Table 1. Service cost and error of solutions found for OR-Library PMP instances

PMP Instance	Service cost found by GA			Solution error [%]		
	best	worst	average (σ)	best	worst	avg.
pmed1	5819	5978	5884.133 (46.985)	0.00	2.73	1.12
pmed2	4105	4358	4202.500 (69.696)	0.29	6.47	2.68
pmed3	4254	4536	4347.500 (73.799)	0.09	6.73	2.29
pmed4	3063	3227	3148.967 (37.464)	0.96	6.36	3.79
pmed5	1392	1539	1468.200 (32.839)	2.73	13.58	8.35
pmed6	7824	7992	7911.233 (46.284)	0.00	2.15	1.11
pmed7	5642	5904	5729.267 (65.984)	0.20	4.85	1.75
pmed8	4565	4881	4687.100 (78.883)	2.70	9.81	5.45
pmed9	2871	3070	2966.100 (58.795)	5.01	12.29	8.49
pmed10	1364	1522	1441.767 (40.390)	8.69	21.27	14.88
pmed11	7702	7941	7783.267 (60.859)	0.08	3.18	1.13
pmed12	6683	7069	6767.667 (79.799)	0.74	6.56	2.01
pmed13	4555	4724	4623.733 (48.046)	4.14	8.00	5.71
pmed14	3151	3350	3248.433 (52.992)	6.17	12.87	9.45
pmed15	1858	2028	1962.800 (38.748)	7.46	17.29	13.52
pmed16	8162	8356	8235.400 (47.539)	0.00	2.38	0.90
pmed17	7037	7331	7169.333 (73.662)	0.54	4.74	2.43
pmed18	4944	5239	5051.800 (67.250)	2.81	8.94	5.05
pmed19	3020	3173	3088.500 (33.516)	6.15	11.53	8.56
pmed20	1982	2148	2061.267 (39.826)	10.79	20.07	15.22
pmed21	9192	9520	9338.767 (67.524)	0.59	4.18	2.20
pmed22	8653	8874	8755.367 (56.862)	0.86	3.44	2.06
pmed23	4797	5022	4903.833 (52.075)	3.85	8.72	6.17
pmed24	3155	3283	3216.367 (28.846)	6.55	10.87	8.62
pmed25	2028	2182	2112.933 (36.535)	10.94	19.37	15.59
pmed26	9917	10175	10045.467 (80.257)	0.00	2.60	1.30
pmed27	8343	8541	8428.100 (54.811)	0.43	2.82	1.46
pmed28	4636	4842	4755.100 (48.889)	3.07	7.65	5.72
pmed29	3253	3369	3322.600 (30.290)	7.25	11.08	9.55
pmed30	2231	2373	2280.200 (35.057)	12.17	19.31	14.64
pmed31	10086	10315	10184.400 (62.649)	0.00	2.27	0.98
pmed32	9394	9790	9488.233 (77.665)	1.04	5.30	2.06
pmed33	4910	5118	4986.800 (46.358)	4.47	8.89	6.10
pmed34	3250	3360	3311.533 (25.425)	7.87	11.52	9.91
pmed35	10400	10663	10527.233 (70.803)	0.00	2.53	1.22
pmed36	9980	10248	10121.700 (65.845)	0.46	3.16	1.89
pmed37	5317	5490	5390.533 (47.432)	5.14	8.56	6.60
pmed38	11060	11375	11171.900 (76.102)	0.00	2.85	1.01
pmed39	9471	9695	9584.933 (53.447)	0.51	2.89	1.72
pmed40	5358	5516	5440.667 (41.527)	4.49	7.57	6.10

noted that, in contrast to the GAs reviewed in section 3, this algorithm used no problem specific heuristic as a local search or greedy exploration step in order to improve the results. Considering it was a pure metaheuristic, the results are encouraging and suggest that the proposed algorithm is a competitive method for the PMP.

6   Conclusions

A new genetic algorithm for the p-median problem was introduced in this work. The algorithm was defined in terms of chromosome encoding, genetic operators, and fitness function. It was designed to be memory efficient and fast. It does not waste time and resources by creating and/or fixing invalid candidate solutions and it also does not need to maintain any hash/lookup tables or bitmaps to check whether a column was already selected or not.

The algorithm was used to find solutions for 40 uncapacitated PMP instances from the well-known OR-Library. The average error of solutions for OR-Library instances was 5.46975% and 7 out of 40 instances were solved to optimality by at least one GA run. The results are encouraging considering the pure metaheuristic nature of the algorithm. Because no problem specific heuristic was used, the algorithm remained general and it can be applied to other tasks that can be expressed by the means of fixed-length subset selection.

Our future work will include further evaluation of the proposed GA, thorough comparison with other metaheuristic and heuristic algorithms for the PMP, and applications of the proposed algorithm to practical problems. The performance of the algorithm in connection with heuristics for PMP will be studied as well.

References

1. Affenzeller, M., Winkler, S., Wagner, S., Beham, A.: Genetic Algorithms and Genetic Programming: Modern Concepts and Practical Applications. Chapman & Hall/CRC (2009)
2. Alp, O., Erkut, E., Drezner, Z.: An efficient genetic algorithm for the p-median problem. Annals of Operations Research 122(1-4), 21–42 (2003)
3. Arroyo, J.E.C., dos Santos, P.M., Soares, M.S., Santos, A.G.: A multi-objective genetic algorithm with path relinking for the p-median problem. In: Kuri-Morales, A., Simari, G.R. (eds.) IBERAMIA 2010. LNCS, vol. 6433, pp. 70–79. Springer, Heidelberg (2010)

4. Correa, E.S., Steiner, M.T.A., Freitas, A.A., Carnieri, C.: A genetic algorithm for the p-median problem. In: Spector, L., Goodman, E.D., Wu, A., Langdon, W., Voigt, H.M., Gen, M., Sen, S., Dorigo, M., Pezeshk, S., Garzon, M.H., Burke, E. (eds.) Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2001), July 7-11, pp. 1268–1275. Morgan Kaufmann, San Francisco (2001)

5. Czarn, A., MacNish, C., Vijayan, K., Turlach, B.: Statistical exploratory analysis of genetic algorithms: The influence of gray codes upon the difficulty of a problem. In: Webb, G.I., Yu, X. (eds.) AI 2004. LNCS (LNAI), vol. 3339, pp. 1246–1252. Springer, Heidelberg (2004)

6. Landa-Torres, I., Del Ser, J., Salcedo-Sanz, S., Gil-Lopez, S., Portilla-Figueras, J., Alonso-Garrido, O.: A comparative study of two hybrid grouping evolutionary techniques for the capacitated p-median problem. Computers and Operations Research 39(9), 2214–2222 (2012)

7. Lim, A., Xu, Z.: A fixed-length subset genetic algorithm for the p-median problem. In: Cantú-Paz, E., et al. (eds.) GECCO 2003. LNCS, vol. 2723, pp. 1596–1597. Springer, Heidelberg (2003)

8. Mitchell, M.: An Introduction to Genetic Algorithms. MIT Press, Cambridge (1996)

9. Mladenović, N., Brimberg, J., Hansen, P., Moreno-Pérez, J.A.: The p-median problem: A survey of metaheuristic approaches. European Journal of Operational Research 179(3), 927–939 (2007)

10. Pullan, W.: A population based hybrid metaheuristic for the p-median problem, pp. 75–82 (2008)

11. Sabeti, M., Boostani, R., Zoughi, T.: Using genetic programming to select the informative eeg-based features to distinguish schizophrenic patients. Neural Network World 22(1), 3–20 (2012)

12. Wu, A.S., Lindsay, R.K., Riolo, R.: Empirical observations on the roles of crossover and mutation. In: Bäck, T. (ed.) Proc. of the Seventh Int. Conf. on Genetic Algorithms, pp. 362–369. Morgan Kaufmann, San Francisco (1997)

13. Yao, J.B., Yao, B.Z., Li, L., Jiang, Y.L.: Hybrid model for displacement prediction of tunnel surrounding rock. Neural Network World 22(3), 263–275 (2012)

Hybrid Bat Algorithm with Artificial Bee Colony

Trong-The Nguyen[1], Jeng-Shyang Pan[2], Thi-Kien Dao[2],
Mu-Yi Kuo[2], and Mong-Fong Horng[2]

[1] Department of Information Technology,
Haiphong Private University, Vietnam
vnthe@hpu.edu.vn
[2] Department of Electronics Engineering,
National Kaohsiung University of Applied Sciences, Taiwan

Abstract. In this paper, a hybrid between Bat algorithm (BA) and Artificial Bee Colony (ABC) with a communication strategy is proposed for solving numerical optimization problems. The several worst individual of Bats in BA will be replaced with the better artificial agents in ABC algorithm after running every Ri iterations, and on the contrary, the poorer agents of ABC will be replacing with the better individual of BA. The proposed communication strategy provides the information flow for the bats to communicate in Bat algorithm with the agents in ABC algorithm. Four benchmark functions are used to test the behavior of convergence, the accuracy, and the speed of the proposed method. The results show that the proposed increases the convergence and accuracy more than original BA is up to 78% and original ABC is at 11% on finding the near best solution improvement.

Keywords: Hybrid Bat Algorithm with Artificial Bee Colony, Bat Algorithm, Artificial Bee Colony Algorithm Optimizations, Swarm Intelligence.

1 Introduction

Computational intelligence algorithms have also been successfully used to solve optimization problems in the engineering, the financial, and the management fields for recently years. For example, genetic algorithms (GA) have been successfully various applications including engineering, the financial, the security [1-3], particle swarm optimization (PSO) techniques have successfully been used to forecast the exchange rates, the optimizing, [4-6], to construct the portfolios of stock, human perception [3, 7, 8], ant colony optimization (ACO) techniques have successfully been used to solve the routing problem of networks, the secure watermarking [9, 10], artificial bee colony (ABC) techniques have successfully been used to solve the lot-streaming flow shop scheduling problem [11], cat swarm optimization (CSO) [12] techniques have successfully been used to discover proper positions for information hiding [13]. Communication between two algorithms is taking advantage of the strength points of each type of algorithms. This idea is based on communication strategies in parallel processing for swarm intelligent algorithms. They only exchange information between populations when the communication strategy is triggered. The existing

methods of these fields are such as such as ant colony system with communication strategies [14], parallel particle swarm optimization algorithm with communication strategies [15], parallel cat swarm optimization [16], Island-model genetic algorithm [17], and parallel genetic algorithm [18]. The parallelized subpopulation of artificial agents increases the accuracy and extends the global search capacity than the original structure. The parallelization strategies simply share the computation load over several processors. The sum of the computation time for all processors can be reduced compared with the single processor works on the same optimum problem. Those algorithms are working only in them self.

In this paper, the concepts of parallel processing and communication strategy are applied to hybrid Bat algorithm with Artificial Bee Colony algorithm is proposed. In the new proposed method, the several poorer individuals in BA algorithm will be replaced with better artificial agents in ABC algorithm after running R_i iterations.

The rest of this paper is organized as follows: a briefly review of ABC and BA is given in session 2; our analysis and designs for the hybrid BA-ABC is presented in session 3; a series of experimental results and the comparison between original BA, original ABC and Hybrid BA-ABC are discussed in session 4; finally, the conclusion is summarized in session 5

2   Related Work

In 2010, Xin-SheYang proposed a new optimization algorithm, namely, Bat Algorithm or original Bat Algorithm (BA), based on swarm intelligence and the inspiration from observing the bats [19]. Original BA simulates parts of the echolocation characteristics of the micro-bat in the simplicity way. It is potentially more powerful than particle swarm optimization and genetic algorithms as well as Harmony search.   The primary reason is that BA uses a good combination of major advantages of these algorithms in some way. Moreover, PSO and Harmony search are the special cases of the Bat Algorithm under appropriate simplifications. Three major characteristics of the micro-bat are employed to construct the basic structure of BA. The used approximate and the idealized rules in Xin-SheYang's method are listed as follows:

All bats utilize the echolocation to detect their prey, but not all species of the bat do the same thing. However, the micro-bat, one of species of the bat is a famous example of extensively using the echolocation.   Hence, the first characteristic is the echolocation behavior. The second characteristic is the frequency that the micro-bat sends a fixed frequency f_{min} with a variable wavelength λ and the loudness A_0 to search for prey.

1. Bats fly randomly with velocity v_i at position x_i. They can adjust the wavelength (or frequency) of their emitted pulses and adjust the rate of pulse emission r ∈ [0, 1], depending on the proximity of their target.
2. There are many ways to adjust the loudness.   For simplicity, the loudness is assumed to be varied from a positive large A_0 to a minimum constant value, which is denoted by Amin.

In Yang's method, the movement of the virtual bat is simulated by equation (1) – equation (3):

$$f_i = f_{min} + (f_{max} - f_{min}) * \beta \tag{1}$$

$$v_i^t = v_i^{t-1} + (x_i^{t-1} - x_{best}) * f_i \tag{2}$$

$$x_i^t = x_i^{t-1} + v_i^t \tag{3}$$

where f is the frequency used by the bat seeking for its prey, f_{min} and f_{max}, represent the minimum   and maximum value, respectively. The x_i denotes the location of the ith bat in the solution space, v_i represents the velocity of the bat, t indicates the current iteration, β is a random vector, which is drawn from a uniform distribution, and $\beta \in$ [0, 1], and x_{best} indicates the global near best solution found so far over the whole population. In addition, the rate of the pulse emission from the bat is also taken to be one of the roles in the process. The micro-bat emits the echo and adjusts the wavelength depending on the proximity of their target. The pulse emission rate is denoted by the symbol r_i, and r_i $\in$[0, 1], where the suffix i indicates the ith bat. In every iteration, a random number is generated and is compared with r_i. If the random number is greater than r_i, a local search strategy, namely, random walk, is detonated. A new solution for the bat is generated by equation (4):

$$x_{new} = x_{old} + \varepsilon A^t \tag{4}$$

where ε is a random number and $\varepsilon \in$[-1, 1], and at represents the average loudness of all bats at the current time step. After updating the positions of the bats, the loudness Ai and the pulse emission rate r_i are also updated only when the global near best solution is updated and the random generated number is smaller than Ai. The update of A_i and r_i are operated by equation (5) and equation (6):

$$A_i^{t+1} = \alpha A_i^t \tag{5}$$

$$r_i^{t+1} = r_i^0[1 - e^{-\gamma t}] \tag{6}$$

where α and γ are constants. In Yang's experiments, $\alpha = \gamma = 0.9$ is used for the simplicity. The process of BA is depicted as follows:

Step1. Initialize the bat population, the pulse rates, the loudness, and define the pulse frequency
Step2. Update the velocities to update the location of the bats, and decide whether detonate the random walk process.
Step3. Rank the bats according to their fitness value, find the current near best solution found so far, and then update the loudness and the emission rate.
Step4. Check the termination condition to decide whether go back to step 2 or end the process and output the result.

In 2005, the Artificial   Bee Colony (ABC) algorithm was proposed Karaboga [20], and in 2008, on the performance of ABC was analyzed [21] whose are based on inspecting the behaviors of real bees on finding nectar and sharing the information of food sources to the bees in the nest. There is three kinds of bee was defined in ABC as being the artificial agent known as the employed bee, the onlooker, and the scout. Every kind of the bees plays different and important roles in the

optimization process. For example: the employed   bee stays   on   a food source, which represents a spot in the solution space, and provides the coordinate for the on-lookers in the hive for reference. The onlooker bee receives the locations of the food sources and selects one of the food sources to gather the nectar. The scout bee moves in the solution space to discover new food sources. The process of ABC optimization is listed as follows:

Step1. Initialization: Spray n_e percentage of the populations into the solution space randomly, and then calculate their fitness values, called the nectar amounts, where n_e represents the ratio of employed bees to the total population. Once these populations are positioned into the solution space, they are called the employed bees. The fitness value of the employed bees is evaluated to take account in their amount of nectar.

$$P_i = \frac{F(\theta_i)}{\sum_{k=1}^{S} F(\theta_k)} \tag{7}$$

Step2. Move the Onlookers: Calculate the probability of selecting a food source by equation (1), where θ_i denotes the position of the ith employed bee, $F(\theta_i)$ denotes   the fitness   function, S represents the   number of employed bees, and P_i is the probabil-ity   of selecting   the ith employed bee. The roulette   wheel   selection method is used to select a food source to move for every onlooker bees and then determine the nectar amounts of them. The onlookers are moved by equation (2), where x_i denotes the position of the ith onlooker bee, t denotes the iteration number, θ is the randomly chosen employed bee, j represents the dimension of the solution and $\Phi(.)$ produces   a series of random variable in the range from -1 to 1.

$$x_{ij}(t+1) = \theta_{ij}(t) + \emptyset(\theta_{ij}(t) - \theta_{kj}(t)) \tag{8}$$

Step3. Update the Best Food Source Found So Far: Memorize the best fitness value and the position, which are found by the bees.
Step4. Move the Scouts: If the fitness values of   the employed   bees do not be im-proved by a continuous predetermined number of iterations, which is   called "*Limit*", those food sources are abandoned, and these employed bees become the scouts. The scouts are moved by equation (3), where r is a random number and $r \in$ range from 0 to 1.

$$\theta_{ij} = \theta_{jmin} + r \times (\theta_{jmax} - \theta_{jmin}) \tag{9}$$

Step5. Termination Checking: Check if the amount of the iterations satisfies the ter-mination condition.   If the termination condition is satisfied, terminate the program and output the results; otherwise go back to Step 2.

3   Hybrid BA with ABC

Hybrid optimization algorithm is structured by communication strategies between two algorithms. This idea is based on replacing the weaker individuals according to fitness evaluation of one algorithm with stronger individuals from other algorithm in parallel processing for swarm intelligent algorithms. Several groups in a parallel structure are

created from dividing the population into subpopulations to construct the parallel processing algorithm. Each of the subpopulations evolves independently in regular iterations. They only exchange information between populations when the communication strategy is triggered. It results in taking advantage of the individual strengths of each type of algorithm, replacing the weaker individuals with the better one from other, the reducing of the population size for each population and the benefit of cooperation is achieved.

The hybrid BA-ABC is designed based on original BA optimization and Artificial Bee Colony optimization algorithm. Each algorithm evolves by optimization independently, i.e. the BA has its own bats and near best solution to replace artificial agents of ABC worst and not near best solution. In contrast, the artificial agents better of ABC are to replace the poorer bats of BA after running Ri iterations. The total iteration contains R times of communication, where $R = \{R_l, 2R_l, 3 R_l,...\}$. The bats in BA don't know the existence of artificial bees of ABC in the solution space.

Let N be the number of population of hybrid BA-ABC, and N_1, N_2 be the number of population of BA and ABC respectively, where $N_1=N_2 = N/2$. If $t \cap R \neq \varphi$, k agents with the top k *fitness* in N_1 will be copied to N_2 to replace the same number of agents with the worst fitness, where t denotes the current iteration count, R_1 and k are the predefined constants. The diagram of the hybrid BA-ABC with communication strategy is shown in figure 1.
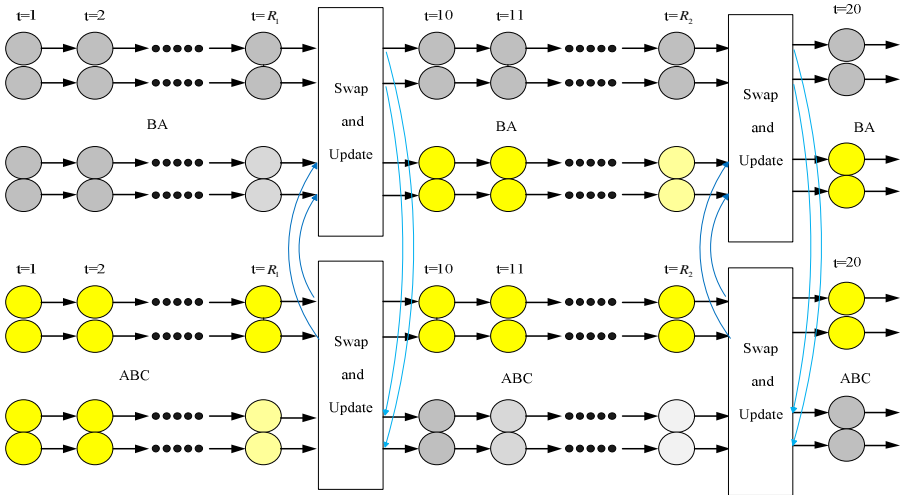
Fig. 1. The diagram of hybrid BA-ABC with a communication strategy

1. **Initialization:** Generate populations for both BA and ABC. Each population is initialized by ABC or by BA independently. Defined the iteration set R for executing the communication strategy. The N_1,N_2 bats and artificial agents S_{ij}^T and X_{ij}^T for populations of BA and ABC respectively, $i = 0, 1, …, N_1- 1, j = 0, 1$, and t is current iteration number. Set $t = 1$.

2. **Evaluation:** Evaluate the value of $f_1(S_{ij}^T)$, $f_2(X_{ij}^T)$ for both BA and ABC in each population. The evolvement of the populations is executed independently by both BA and ABC optimization.

3. **Update:** Update the velocity and the positions of Bat using Eqs. (1), (2) and (3). Update the best Food Source Found So Far: Memorize the best fitness value and the position, which are found by the bees using Eqs. (7),(8) (9).

4. **Communication Strategy:** Migrate the best bats among all the individual of BA's population, copy k individuals with the top k *fitness* in N_1 replace the poorer agents in N_2 of ABC's population and update for each populations in every R_1 iterations. Conversely, migrate the best artificial agents among all the individual of ABC's population, copy k agents with the top k *fitness* in N_2 replace the poorer bats in N_1 and update for each population every R_1 iterations.

5. **Termination:** Repeat step 2 to step 5 until the predefined value of the function is achieved or the maximum number of iterations has been reached. Record the best value of the function $f(S^t)$ and the best bat position among all the bats S^t. Record the best value of the function $f(X^t)$ and the best food Source among all the agents X^t.

4   Experimental Results

This section presents simulation results and compares the hybrid BA-ABC with the original BA, and original ABC, both in terms of solution quality, convergence capability, and the execution time in the number of function evaluations taken. Four benchmark functions are used to test the accuracy and the convergence of hybrid BA-ABC. All the benchmark functions for the experiments are averaged over different random seeds with10 runs.

Let $S =\{s_1, s_2 ,..., s_n\}$, $X =\{x_1, x_2 ,..., x_n\}$ be the m-dimensional real-value vectors for BA and ABC respectively. The benchmark functions are Griewank, Rastrigin, Rosenbrock and Spherical and listed in equation (10) to equation (13). The goal of the optimization is to minimize the outcome for all benchmarks. The population size of Hybrid BA-ABC, original BA and original ABC are set to 20 for all the algorithms in the experiments. The detail of parameter settings of BA can be found in [19] and setting of ABC can be found in [21].

$$f_1(x) = 1 + \sum_{i=1}^{N} \frac{x_i^2}{4000} + \prod_{i=1}^{N} cos \frac{x_i}{\sqrt{i}} \tag{10}$$

$$f_2(x) = \sum_{i=1}^{N}[10 + x_i^2 - 10cos2\pi x_i \tag{11}$$

$$f_3(x) = \sum_{i=1}^{n-1}(100(x_{i-1} - x_i^2)^2 + (1 - x_i)^2 \tag{12}$$

$$f_4(x) = \sum_{i=1}^{N} x_i^2 \tag{13}$$

The initial range and the total iteration number for all test functions are listed in Table 1.

Table 1. The initial range and the total iteration of test standard functions

Function	Initial range $[x_{min}, x_{max}]$	Total iteration
$f_1(x)$	[-100, 100]	400
$f_2(x)$	[5.12, 5.12]	400
$f_3(x)$	[-30,30]	400
$f_4(x)$	[-100, 100]	400

The optimization for all of these test functions is to minimize the outcome. The parameters setting for hybrid BA-ABC with original BA side are the initial loudness $A_i^0 = 0.25$, pulse rate $r_i^0 = 0.5$ the total population size $N_1 = 10$ and the dimension of the solution space $M = 10$, frequency minimum f_{min} = *the lowest of initial range function* and frequency minimum f_{max} = *the highest of initial range function,* and with original ABC side are the initial 'limit'=10 of food source the total population size N_2 = 10 and the dimension of the solution space $M = 10$:. Each benchmark function contains the full iterations of 400 is repeated by different random seeds with 10 runs. The final result is obtained by taking the average of the outcomes from all runs. The results are compared with the original BA and original ABC respectively.

Table 2 compares the quality of performance and time running for numerical problem optimization between Hybrid BC-ABC and original Bat algorithm. It is clearly seen that, almost these cases of testing benchmark functions for Hybrid BA-ABC are better than original BA in terms of convergence. It is special case with test function $f_4(x)$, the Spherical has the mean of value function minimum of total 10 seed runs is 3.18E+07 for Hybrid BA-ABC's performance evaluation, but, for original BA is 2.63E+08, reaches at 78% improvement of convergence. However, all benchmark functions for average time consuming of Hybrid BA-ABC are double time taken in comparison in original BA, for the reasons, the Hybrid algorithm must perform mutation and update operations.

Table 2. The comparison between original BA and hybrid BA-ABC in terms of quality performance evaluation and speed

Function	Performance evaluation		Time running evaluation (seconds)	
	Original BA	*Hybrid BA-ABC*	*Original BA*	*Hybrid BA-ABC*
$f_1(x)$	31.3202	31.0783	0.1657	0.473
$f_2(x)$	4.60E+02	1.39E+03	0.225	0.4318
$f_3(x)$	7.46E-01	3.83E-01	0.1773	0.3897
$f_4(x)$	2.63E+08	3.18E+07	0.1639	0.3491
Average value	**2.63E+08**	**3.18E+07**	**0.7319**	**1.6436**

Figures from 2 to 3 show the experimental results of four benchmark functions in running 10 seeds output with the same iteration of 400 in comparison win original BA.
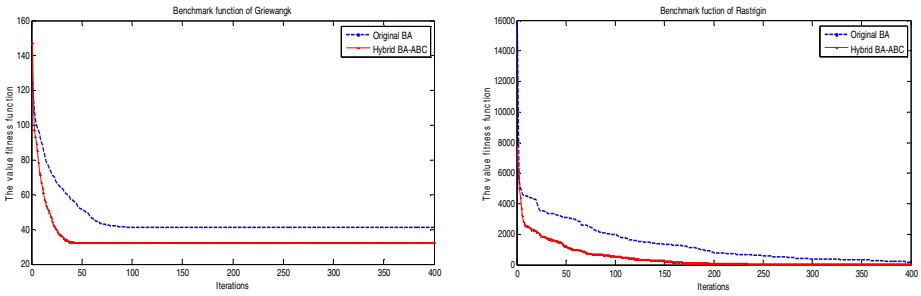
Fig. 2. The mean of function minimum curves in comparing Hybrid BA-ABC and original BA algorithms for function of Griewangk and Rastrigin
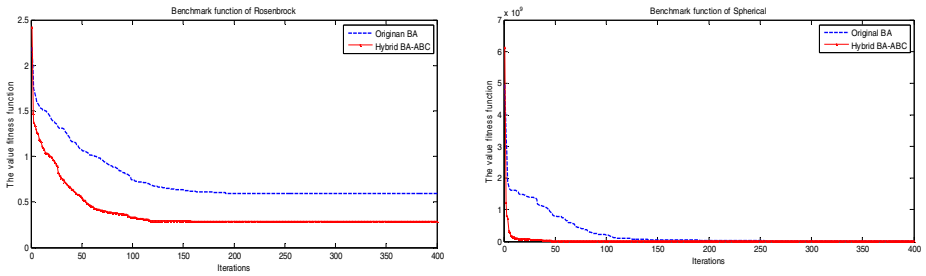
Fig. 3. The mean of function minimum curves in comparing Hybrid BA-ABC and original BA algorithms for function of Rosenbrock and Spherical

Table 3 compares the quality of performance and time running for numerical problem optimization between Hybrid ABC-BA and original ABC. It is clearly seen that, almost these cases of testing benchmark functions for Hybrid ABC-BA are more convergence than original ABC. Average value of all benchmark functions for Hybrid ABC- BA is 8.02E+06 in performance evaluation, but this figure is 9.93E+06 for original ABC, reaches at 11% improvement of convergence. However, average times consuming of all benchmark functions for Hybrid ABC-BA are longer taken than that for original ABC. For this result, the reason is the Hybrid algorithm must perform mutation and update operations.

Table 3. The comparison between hybrid ABC-BA and origianl ABC in terms of quality performance evaluation and speed

Function	Performance evaluation		Time running evaluation (seconds)	
	Original ABC	*Hybrid ABC-BA*	*Original ABC*	*Hybrid ABC-BA*
$f_1(x)$	0.2134	0.078	0.3281	0.4503
$f_2(x)$	2.55E+02	2.04E+02	0.2249	0.4218
$f_3(x)$	9.93E+06	8.02E+06	0.2333	0.3997
$f_4(x)$	8.49E+02	5.94E+02	0.1759	0.3591
Average value	**9.93E+06**	**8.02E+06**	**0.9622**	**1.6309**

Figures from 4 to 5 show the experimental results of four benchmark functions in running 10 seeds output with the same iteration of 400 in comparison win original ABC.
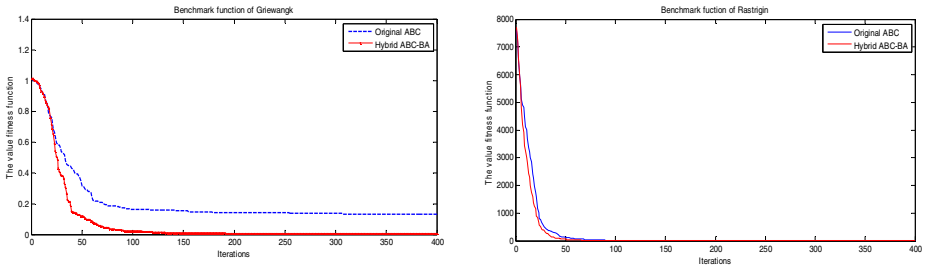
Fig. 4. The mean of function minimum curves in comparing Hybrid ABC-BA and original ABC algorithms for function of Griewangk and Rastrigin
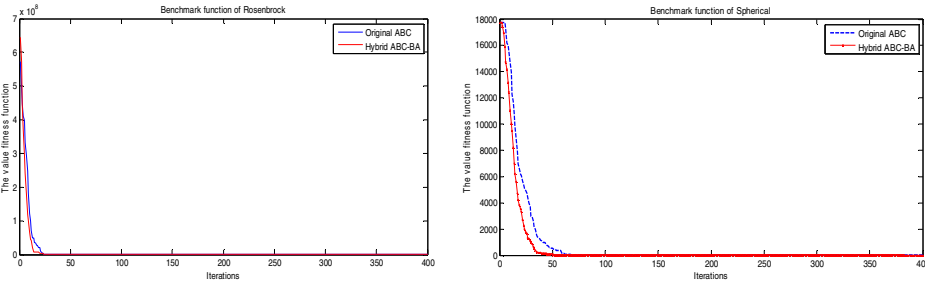
Fig. 5. The mean of function minimum curves in comparing Hybrid ABC-BA and original ABC algorithms for function of Rosenbrock and Spherival

5   Conclusion

This paper, a novel proposed optimization algorithm is presented, namely Hybrid BA-ABC (hybrid Bat algorithm with Artificial Bee Colony). The implementation of hybrid for optimization algorithms could have important significance for taking advantage of the power of each algorithm and achieving cooperation of optimization algorithms. In new proposed algorithm, the several worse individual of Bats in BA are replaced with the better artificial agents in ABC algorithm after running every Ri iterations, and on the contrary, the poorer agents of ABC are replacing with the better individual of BA. The proposed communication strategy provides the information flow for the bats to communicate in Bat algorithm with the agents in ABC algorithm. The performance of Hybrid BA-ABC algorithm is better than both original BA and ABC in terms of convergence and accuracy. The results of proposed algorithm on a set of various test problems show that Hybrid BA-ABC increases the convergence and accuracy more than original BA is up to 78% and original ABC is at 11% on finding the near best solution improvement.

References

1. Srinivas, M., Patnaik, L.M.: Genetic Algorithms: A Survey. Computer 27, 17–26 (1994)
2. Wang, S., Yang, B., Niu, X.: A Secure Steganography Method based on Genetic Algorithm. Journal of Information Hiding and Multimedia Signal Processing 1, 8 (2010)
3. Ruiz-Torrubiano, R., Suarez, A.: Hybrid Approaches and Dimensionality Reduction for Portfolio Selection with Cardinality Constraints. IEEE Computational Intelligence Magazine 5(2), 92–107 (2010)
4. Chen, S.-M., Chien, C.-Y.: Solving the traveling salesman problem based on the genetic simulated annealing ant colony system with particle swarm optimization techniques. Expert Systems with Applications 38(12), 14439–14450 (2011)
5. Hsu, C.-H., Shyr, W.-J., Kuo, K.-H.: Optimizing Multiple Interference Cancellations of Linear Phase Array Based on Particle Swarm Optimization. Journal of Information Hiding and Multimedia Signal Processing (4), 292–300 (2010)
6. Chen, S.-M., Kao, P.-Y.: TAIEX forecasting based on fuzzy time series, particle swarm optimization techniques and support vector machines. Information Sciences 247, 62–71 (2013)
7. Jui-Fang, C., Shu-Wei, H.: The Construction of Stock's Portfolios by Using Particle Swarm Optimization, p. 390 (2007)
8. Parag Puranik, P.B., Abraham, A., Palsodkar, P., Deshmukh, A.: Human Perception-based Color Image Segmentation Using Comprehensive Learning Particle Swarm Optimization. Journal of Information Hiding and Multimedia Signal Processing 2(3), 227–235 (2011)
9. Pinto, P.C., Nagele, A., Dejori, M., Runkler, T.A., Sousa, J.M.C.: Using a Local Discovery Ant Algorithm for Bayesian Network Structure Learning. IEEE Transactions on Evolutionary Computation 13(4), 767–779 (2009)

10. Khaled Loukhaoukha, J.-Y.C., Taieb, M.H.: Optimal Image Watermarking Algorithm Based on LWT-SVD via Multi-objective Ant Colony Optimization. Journal of Information Hiding and Multimedia Signal Processing 2(4), 303–319 (2011)
11. Pan, Q.-K., Tasgetiren, M.F., Suganthan, P.N., Chua, T.J.: A discrete artificial bee colony algorithm for the lot-streaming flow shop scheduling problem. Inf. Sci. 181(12), 2455–2468 (2011)
12. Chu, S.-C., Tsai, P.-W.: Computational Intelligence Based on the Behavior of Cats. International Journal of Innovative Computing, Information and Control 3(1), 8 (2006)
13. Wang, Z.-H., Chang, C.-C., Li, M.-C.: Optimizing least-significant-bit substitution using cat swarm optimization strategy. Inf. Sci. 192, 98–108 (2012)
14. Chu, S.-C., Roddick, J.F., Pan, J.-S.: Ant colony system with communication strategies. Information Sciences 167(1-4), 63–76 (2004)
15. Chang, J.F., Chu, S.C., Roddick, J.F., Pan, J.S.: A parallel particle swarm optimization algorithm with communication strategies. Journal of Information Science and Engineering 21(4), 9 (2005)
16. Pei-Wei, T., Jeng-Shyang, P., Shyi-Ming, C., Bin-Yih, L., Szu-Ping, H.: Parallel Cat Swarm Optimization, pp. 3328–3333 (2008)
17. Whitley, D., Rana, S., Heckendorn, R.B.: The Island Model Genetic Algorithm: On Separability, Population Size and Convergence. Journal of Computing and Information Technology 1305/1997, 6 (1998)
18. Abramson, D., Abela, J.: A Parallel Genetic Algorithm for Solving the School Timetabling Problem. Division of Information Technology, pp. 1–11 (1991)
19. Yang, X.-S.: A New Metaheuristic Bat-Inspired Algorithm. In: González, J.R., Pelta, D.A., Cruz, C., Terrazas, G., Krasnogor, N. (eds.) NICSO 2010. SCI, vol. 284, pp. 65–74. Springer, Heidelberg (2010)
20. Karaboga, D.: An Idea based on Honey Bee Swarm for Numerical Optimization. Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department (2005)
21. Karaboga, D., Basturk, B.: On the Performance of Artificial Bee Colony (ABC) Algorithm. Applied Soft Computing 1, 687–697 (2008)

Compact Bat Algorithm

Thi-Kien Dao[1], Jeng-Shyang Pan[1], Trong-The Nguyen[1],
Shu-Chuan Chu[2], and Chin-Shiuh Shieh[1]

[1] Department of Electronics Engineering,
National Kaohsiung University of Applied Sciences, Taiwan
jvnkien@gmail.com
[2] School of Computer Science, Engineering and Mathematics,
Flinders University, Australia

Abstract. Addressing to the computational requirements of the hardware devices with limited resources such as memory size or low price is critical issues. This paper, a novel algorithm, namely compact Bat Algorithm (cBA), for solving the numerical optimization problems is proposed based on the framework of the original Bat algorithm (oBA). A probabilistic representation random of the Bat's behavior is inspired to employ for this proposed algorithm, in which the replaced population with the probability vector updated based on single competition. These lead to the entire algorithm functioning applying a modest memory usage. The simulations compare both algorithms in terms of solution quality, speed and saving memory. The results show that cBA can solve the optimization despite a modest memory usage as good performance as oBA displays with its complex population-based algorithm. It is used the same as what is needed for storing space with six solutions.

Keywords: Bat algorithm, compact Bat algorithm, Optimizations, Swarm intelligence.

1   Introduction

Computational intelligence algorithms have also been successfully used to solve optimization problems in the engineering, the financial, and the management fields for recently years. For example, genetic algorithms (GA) have been successfully various applications including engineering, the financial, the security [1-3], particle swarm optimization (PSO) techniques have successfully been used to construct the portfolios of stock, human perception [3-5], ant colony optimization (ACO) techniques have successfully been used to solve the routing problem of networks, the secure watermarking [6, 7], artificial bee colony (ABC) techniques have successfully been used to solve the lot-streaming flow shop scheduling problem [8], cat swarm optimization (CSO) [9] techniques have successfully been used to discover proper positions for information hiding [10]. Some applications require the solution of a complex optimization problem event though in limited hardware conditions. These conditions are to use a computational device due to cost and space limitations. For example, wireless sensor networks (WSN) are networks of small, battery-powered, memory-constraint

devices named sensor nodes, which have capability of wireless communication over a restricted area [11]. Due to memory and power constraints, they need to be well arranged to build a fully functional network. The other applications require a very fast solution of the optimization problem due to the communication time between a control/actuator devices and an external computer, real-time necessities within the control/actuator devices. For instance, in telecommunications[12] or in industrial plants for energy production[13]. Special applications require fault-tolerance in a high priority and/or to avoid rebooting of the device. For example, in the space shuttle control [14], or in communication underwater [15]. The mentioned problem is not enough memory of computational devices to store a population composed of numerous candidate solutions of those computational intelligence algorithms.

Compact algorithms are a promise answer for this problem. An efficient compromise is used in compact algorithms to present some advantages of population-based algorithms but the memory is not required for storing an actual population of solutions. Compact algorithms simulate the behavior of population-based algorithms by employing, instead of a population of solutions, its probabilistic representation. In this way, a much smaller number of parameters must be stored in the memory. Thus, a run of these algorithms requires much less capacious memory devices compared to their correspondent population-based structures.

The very first implementation of compact algorithms has been the compact Genetic Algorithm (cGA) [16]. The cGA simulates the behavior of a standard binary encoded Genetic Algorithm (GA). It can be seen that cGA has a performance almost as good as that of GA and that cGA requires a much less capacious memory. The compact Differential Evolution (cDE) algorithm has been introduced in [17]. The success of cDE implementation is the combination of two factors. The first is that a DE scheme seems to benefit from the introduction of a certain degree of randomization due to the probabilistic model. The second is that the one-to-one spawning survivor selection typical of DE (the offspring replaces the parent) can be naturally encoded into a compact logic. The compact Particle Swarm Optimization (cPSO) has been defined in [18]. The implementation of cPSO algorithm benefits from the same natural encoding of the selection scheme employed by DE and another ''ingredient'' of compact optimization, i.e. a special treatment for the best solution ever detected and reinterpreted as an evolutionary algorithms in order to propose a compact encoding of PSO.

In this paper, the behavior and the characteristic of the Bat are reviewed to improve the Bat algorithms [19, 20] and to present the compact Bat Algorithm (cBA) based on the framework of the original BA (oBA). According to the experimental results, our proposed cBA presents same result in finding original Bat algorithm.

The rest of this paper is organized as follows: a briefly review of BA is given in session 2; our analysis and designs for the cBA is presented in session 3; a series of experimental results and the compare between oBA and cBA are discussed in session 4; finally, the conclusion is summarized in session 5.

2   Related Works

A random walk is a mathematical formalization of a path that consists of a succession of random steps. This work is primarily inspired by the random walk model in [21,

22]. This model focused on building block for representing individual in warms. Compact algorithms is represented the population as probability distribution based on random steps over the set of solutions. By discretizing its probability representation, the proposed algorithm reduces the oBA's memory requirements. A set of frequencies of Bats is a building block as a whole given high contribution to the fitness of an individual. There are no interactions among building blocks, so they could be solved independently. The behavior of building blocks for solving to optimality could be simulated by the dynamics of the random walk model[22].

In 2010, Xin-SheYang proposed a new optimization algorithm, namely, Bat Algorithm or original Bat Algorithm (oBA), based on swarm intelligence and the inspiration form observing the bats [19] . oBA simulates parts of the echolocation characteristics of the micro-bat in the simplicity way. Three major characteristics of the micro-bat are employed to construct the basic structure of BA. All bats utilize the echolocation to detect their prey, but not all species of the bat do the same thing. However, the micro-bat, one of species of the bat is a famous example of extensively using the echolocation. Hence, the first characteristic is the echolocation behavior. The second characteristic is the frequency that the micro-bat sends a fixed frequency f_{min} with a variable wavelength λ and the loudness A_0 to search for prey.

1. Bats fly randomly with velocity v_i at position x_i. They can adjust the wavelength (or frequency) of their emitted pulses and adjust the rate of pulse emission $r \in [0, 1]$, depending on the proximity of their target;
2. There are many ways to adjust the loudness. For simplicity, the loudness is assumed to be varied from a positive large A_0 to a minimum constant value, which is denoted by Amin.

In Yang's method, the movement of the virtual bat is simulated by equation (1) – equation (3):

$$f_i = f_{min} + (f_{max} - f_{min}) * \beta \tag{1}$$

$$v_i^t = v_i^{t-1} + (x_i^{t-1} - x_{best}) * f_i \tag{2}$$

$$x_i^t = x_i^{t-1} + v_i^t \tag{3}$$

where f is the frequency used by the bat seeking for its prey, f_{min} and f_{max}, represent the minimum and maximum value, respectively. x_i denotes the location of the ith bat in the solution space, v_i represents the velocity of the bat, t indicates the current iteration, β is a random vector, which is drawn from a uniform distribution, and $\beta \in [0, 1]$, and x_{best} indicates the global near best solution found so far over the whole population. In addition, the rate of the pulse emission from the bat is also taken to be one of the roles in the process. The micro-bat emits the echo and adjusts the wavelength depending on the proximity of their target. The pulse emission rate is denoted by the symbol r_i, and $r_i \in [0, 1]$, where the suffix i indicates the ith bat. In every iteration, a random number is generated and is compared with r_i. If the random number is

greater than r_i, a local search strategy, namely, random walk, is detonated. A new solution for the bat is generated by equation (4):

$$x_{new} = x_{old} + \varepsilon A^t \tag{4}$$

where ε is a random number and $\varepsilon \in [-1, 1]$, and at represents the average loudness of all bats at the current time step. After updating the positions of the bats, the loudness A_i and the pulse emission rate r_i are also updated only when the global near best solution is updated and the random generated number is smaller than A_i. The update of A_i and r_i are operated by equation (5) and equation (6):

$$A_i^{t+1} = \alpha A_i^t \tag{5}$$

$$r_i^{t+1} = r_i^0 [1 - e^{-\gamma t}] \tag{6}$$

where α and γ are constants. In Yang's experiments, $\alpha = \gamma = 0.9$ is used for the simplicity. The process of oBA is depicted as follows:

Step 1. Initialize the bat population, the pulse rates, the loudness, and define the pulse frequency
Step 2. Update the velocities to update the location of the bats, and decide whether detonate the random walk process.
Step 3. Rank the bats according to their fitness value, find the current near best solution found so far, and then update the loudness and the emission rate.
Step 4. Check the termination condition to decide whether go back to step 2 or end the process and output the result.

3   Compact Bat Algorithm

As mentioned above that compact algorithms process an actual population of solution as a virtual population. This virtual population is encoded within a data structure, namely Perturbation Vector (*PV*) as probabilistic model of a population of solutions. The distribution of the individual in the hypothetical swarms must be described by a probability density function (PDF) [23] defined on the normalized interval is from -1 to +1. The distribution of the each Bat of swarms could be assumed as Gaussian PDF with mean μ and standard deviation σ [16]. A minimization problem is considered in an *m*-dimensional hyper-rectangle in Normalization of two truncated Gaussian curves (*m* is the number of parameters). Without loss of generality, the parameters assume to be normalized so that each search interval is [-1,+1] . Therefore PV is a vector of $m \times 2$ matrix specifying the two parameters of the PDF of each design variable being defined as:

$$PV^t = [\mu^t, \sigma^t] \tag{7}$$

where μ and σ are mean and standard deviation values a Gaussian (PDF) truncated within the interval [-1, +1], respectively. The amplitude of the PDF is normalized in order to keep its area equal to 1. The apex t is time steps. The initialization of the virtual population is generated for each design variable i, $\mu_i^1 = 0$ and $\delta_i^1 = k$ where k is set as a large positive constant (e.g. $k = 10$). The PDF height normalization is obtained approximately sufficient in well the uniform distribution with a wide shape. The generating for a candidate solution x_i is produced from $PV(\mu_i, \delta_i)$. The value of

mean μ and standard deviation δ in *PV* are associated equation of a truncated Gaussian PDF is described as following:

$$PDF\big(trucNormal(x)\big) = \frac{e^{\frac{(x-\mu_i)^2}{2\delta_i^2}}}{\delta_i(\text{erf}\big(\frac{\mu_i+1}{\sqrt{2}\delta_i}\big)-\text{erf}\big(\frac{\mu_i+1}{\sqrt{2}\delta_i}\big))} \tag{8}$$

The PDF in formula (8) is then used to compute the corresponding Cumulative Distribution Function (CDF). The CDF is constructed by means of Chebyshev polynomials by following the procedure described in [24], the codomain of CDF is arrange from 0 to 1. The distribution function or cumulative distribution function (*CDF*) describes the probability that a real-valued random variable X with a given probability distribution will be found at a value less than or equal to x_i CDFs are also used to specify the distribution of multivariate random variables.

$$CDF = \int_0^1 PDF * dx \tag{9}$$

The sampling of the design variable x_i from *PV* is performed by generating a random number rand (0, 1) from a uniform distribution and then computing the inverse function of CDF in rand (0, 1). The newly calculated value is x_i.

$$x_i = inverse(CDF) \tag{10}$$

When the comparison between two design variables for individuals of the swarm (or better two individuals sampled from PV) is performed the winner solution biases the PV. Let us indicate with winner the vector that scores a better fitness value and with loser the individual losing the (fitness based) comparison. Regarding the mean values l, the update rule for each of its elements is $\mu_i^t, \delta_i^t => \mu_i^{t+1}, \delta_i^{t+1}$.

$$\mu_i^{t+1} = \mu_i^t + \frac{1}{N_p}(winner_i - loser_i) \tag{11}$$

where Np is virtual population size. Regarding δ values, the update rule of each element is given by:

$$\delta_i^{t+1} = \sqrt{(\delta_i^t)^2 + (\mu_i^t)^2 - (\mu_i^{t+1})^2 + \frac{1}{N_p}(winner_i^2 - loser_i^2)} \tag{12}$$

$$[winner, loser] = complete(x_{best}, x^{t+1}) \tag{13}$$

The construction of formulas (11) and (12) are persistent and non-persistent structures with tested results given in [25]. Similar to the binary cGA case, it was impossible to assess whether one or another elitist strategy was preferable. As reported in [17], it is fundamental to remember that the virtual population size N_p is parameter typical of compact algorithms and does not strictly correspond to the population size in a population-based algorithm. The virtual population size, in real-valued compact optimization, is a parameter which biases the convergence speed of

the algorithm. In [25] has been mentioned that, for a given problem (with its dimensionality), this parameter should be set several times bigger than the population size of a corresponding population-based algorithm. In elitist compact schemes, at each moment of the optimization process, the solution displaying the best performance is retained in a separate memory slot. If a new candidate solution is computed, the fitness based comparison between it and the elite is carried out. If elite is a winner solution, it biases the *PV* as shown in formulas (11) and (12).

---

1) *Initialization probability vector* $(PV(\mu, \delta))$
   for *i=1:n* **do** μ_i^t *=0;* δ_i^t*=k= 10;*
2) *Initialization parameters: pulse rate* r_i*,* the loudness A_i*,* β *= random, and* $v_i^1 = 0$*;*
   Generate global best solution x_{best} *from PV; Define pulse frequency* f_{min}*,* f_{max} *as search range;*
3) Evaluate new solutions
   while *termination is not satisfied* **do**
   x^I = generateFrom(*PV*)
   Update velocities and locations
   Use equations (1),(2), and (3)
   if (β > ri)
   Select a solution among the best solution,
   Generate a local solution around selected best solution
   endif
    [*winner, loser*]=**compete**(x^{t+1}, x_{best})
4) *Update PV*
   Use equations (11), and (12)
5) *Global best update*
   x_{best} *=winner;* $\mu_i^{t+1} = \mu_i^t$
6) *Accept new solutions,*
    Fnew =fitness(x);
        Rank the bats and find the current best
        Update if the solution improves, or not too loud
       if (Fnew<=fmin) & (β <A)
         best=xt;   fmin=Fnew;
       endif
    endwhile

Fig. 1. The pseudo code of compact Bat algorithm

The fitness value of the position x^{t+1} is calculated and compared with x_{best} to determine a winner and a loser. Equation (11) and equation (12) are then applied to update the probability vector *PV*. If f(x^{t+1}) <= *f(min)* and β <A, the value of the global best is then updated: $f_{min}=F_{new}$ and the process is repeated over for the subsequent steps. Current values for x^{t+1} and v^{t+1} are retained for subsequent algorithm steps. Figure 1 shows the pseudo code of algorithm working principles of cBA.

4   Experimental Results

This section presents simulation results and compares the cBA with the oBA, both in terms of solution quality and in the number of function evaluations taken. To evaluate the accuracy and the computational speed of the proposed cBA, four benchmark functions are chosen to use in the experiments. All experiments are averaged over different random seeds with 25 runs. All benchmark functions are listed in equation (14) - equation (19).

$$f_1(x) = \sum_{i=1}^{n-1}(100(x_{i-1} - x_i^2)^2 + (1 - x_i)^2 \qquad (14)$$

$$f_2(x) = \sum_{i=1}^{n}(\sum_{k=1}^{i} x_i) \qquad (15)$$

$$f_3(x) = 1 + \sum_{i=1}^{N}\frac{x_i^2}{4000} + \prod_{i=1}^{N} cos\frac{x_i}{\sqrt{i}} \qquad (16)$$

$$f_4(x) = \sum_{i=1}^{N}[10 + x_i^2 - 10cos2\pi x_i \ ] \qquad (17)$$

The initial range and the total iteration number for all test functions are listed in Table 1.

Table 1. The initial range and the total iteration of test standard functions

Function	Initial range $[x_{min}, x_{max}]$	Total iteration
$f_1(x)$	[-100,100]	5000
$f_2(x)$	[-100,100]	5000
$f_3(x)$	[-100,100]	5000
$f_4(x)$	[-5.12,5.12]	5000

The optimization goal for all of these test functions is to minimize the outcome. The parameters setting for both cBA and oBA: are the initial loudness $A_i^0 = 0.25$ (range 0.1 to 0.9), pulse rate $r_i^0 = 0.5$ (range 0.5 to 0.9) the total population size $n = 20$ and the dimension of the solution space $M = 30$, frequency minimum f_{min} = the lowest of initial range function and frequency minimum f_{max} = the highest of initial range function.   Each function contains the full iterations of 5000 is repeated by different random seeds with 25 runs. The final results are obtained by taking the average of the outcomes from all runs. The results are compared with the original BA (oBC).

4.1   Comparison Optimizing Performance Algorithms

Table 2 compares the quality of performance and time running for numerical problem optimization between cBA and oBA. It is clearly seen that, almost cases of benchmark functions for optimizing in compact Bat algorithm are faster convergence. It is special case with test function $f_2(x)$ quadric has the mean of value function

minimum of total 25 runs is 44865 with average time running equal 3.1764 seconds for oBA evaluation. However, for cBA this value of function minimum of total 25 runs is 36321 with time running equal 0.9936 seconds in same executing computer. The mean of four test functions evaluation of minimum function 25 runs is 9.40E+08 with average time consuming 7.0093 for oBA but, 8.87E+08 with average time consuming 3.5843 for cBA respectively. It improved the accuracy and the convergence is up to 4%.

Table 2. The comparison between oBA and cBA in terms of quality performance evaluation and speed

Function	Performance evaluation		Time running evaluation	
	oBA	*cBA*	*oBA*	*cBA*
$f_1(x)$	9.4E+08	8.87E+08	1.2659	1.0141
$f_2(x)$	44865	36321	3.1764	0.9936
$f_3(x)$	6.5331	6.1153	1.3823	0.8845
$f_4(x)$	209.356	256.2685	1.1847	0.6921
Average value	**9.40E+08**	**8.87E+08**	**7.0093**	**3.5843**

Figure 2 shows the average of function minimum of four test functions in 25 runs output in the same iteration of 5000. It can be clearly seen that the curves of compact BA (red and star lines) are faster in convergence.
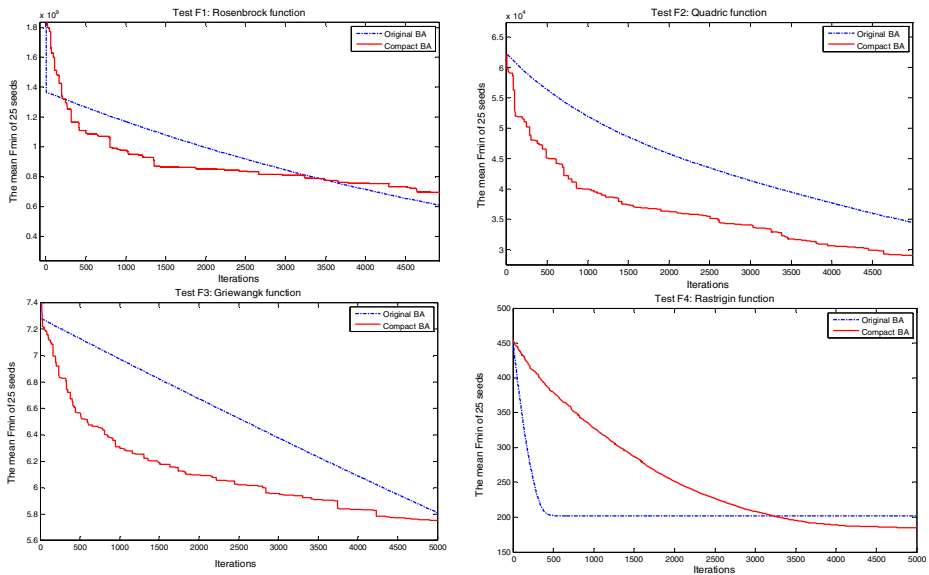
Fig. 2. The mean of function minimum curves in comparing cBA and oBA algorithms for four benchmark functions

4.2   Comparison Saving-Memory and Time-Complexity Algorithms

Table 3 compares the saving-memory computations of two algorithms cBA and oBA. It can be clearly seen that the number memory variables of cBA is smaller than that of oBA in the same condition of computation such as iterations. The real number of population or population size of oBA is N, but that size for cBA is only one. Even though, the number equations used for optimizing computation in cBA is six such as equations (1), (2), (3), (12), (13) and (14), and the number equations used for optimizing computation in oBA is only three of them such as equations (1), (2), and (3), the computing complexity of cBA is $6{\times}T{\times}iteration$ and it for oBA $is\ 3{\times}T{\times}N{\times}iteration$. Thus, the rate of saving-memory equals the computing complexity of cBA per the computing complexity of oBA as given: rate = 2/N.

Table 3. The saving-memory comparison between cBA and oBA

Algo-rithms	Population size	#Memory variable	# Equ-ations	Computing complexity
oBA	N	$3{\times}N$	(1),(2),(3)	$3{\times}T{\times}N$ $\times iteration$
cBA	1	6	(1),(2),(3),(12), (13),(14)	$6{\times}T$ $\times iteration$

The considered computational times, for both the algorithms cBA and oBA, have been calculated by means of a PC Intel Core 2 Duo 2.4 GHz with 4 GB RAM employing in Windows7-OS, with Matlab (R2011b), version 7.13.0.564 32bits. Figure 4 illustrates the comparison of executing time between cBA and oBA in 25 seed runs with iteration 5000 for four benchmark functions. It is clearly seen that the most cases of test functions time executing in the proposed cBA (red colored bar) are smaller than that executing in oBA (blue colored bar). The bar of test function number 2 is longest distance different oBA and cBA more than double time executed.
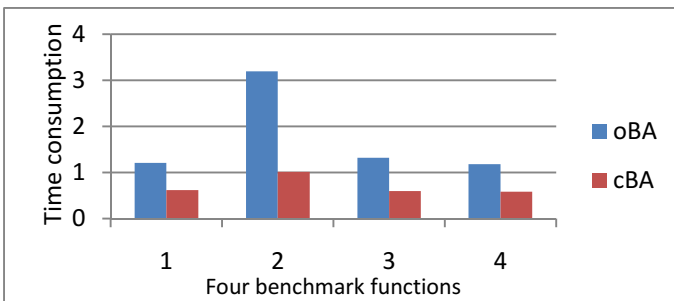
Fig. 3. Comparison two algorithms in term of time running for 6 chosen benchmark functions

For the computational times, several dimensionality levels such as 5, 10, 20, 30, and 40 dimensions should be tested in standard test functions. Figure 4 compares two algorithms in term of different dimensions for test functions. The most of cases test functions for convergence of cBA is smaller than that for oBA.
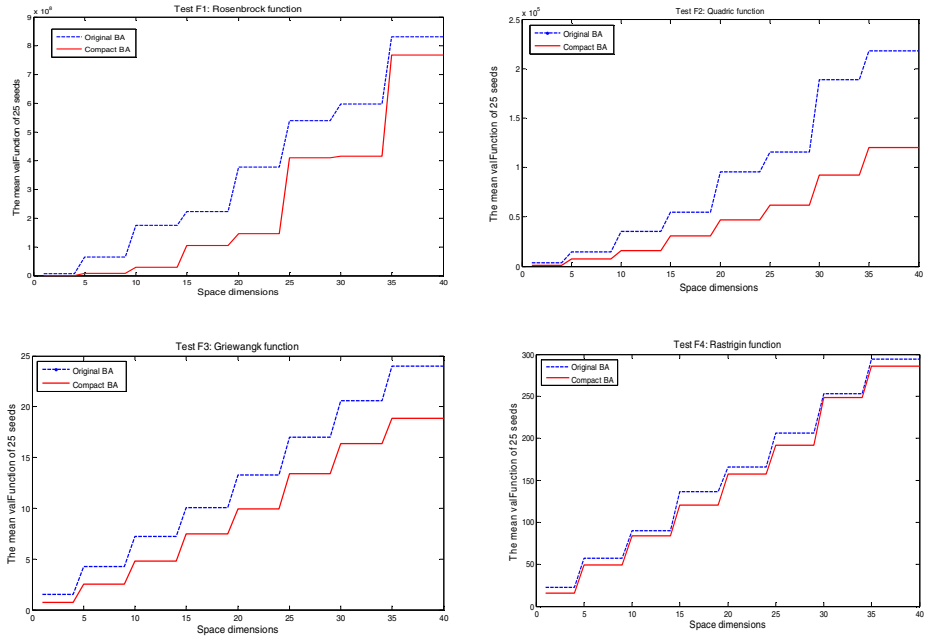
Fig. 4. Comparison two algorithms in term of different dimension for test functions

For variety population size of swarms *N*, the most cases of test functions employing in cBA is not effected much in comparison with oBA because of population size in cBA is virtual population, so the mean of value functions test for cBA are more stable.

5   Conclusion

This paper, a novel proposed optimization algorithm is presented, namely compact Bat algorithm (cBA). The implementation of compact for optimization algorithms could have important significance for the development of embedded devices with small size, low price and being suitable for trend of ubiquitous computing today. In new proposed algorithm, the actual design variable of solutions search space of Bat algorithm is replaced with a probabilistic representation of the population. This feature is important for application problems characterized by a limited memory since it allows the embedded implementation in small and cheap devices. The performance of cBA algorithm is as good as the other previous works in literature with respect to compact algorithms. The results of proposed algorithm on a set of various test

problems show that cBA seems to be a valid alternative for optimization problems plagued by a limited memory. Experimental results on this real-world application also show the applicability of the proposed approach and highlight the good cBA performance within the category of memory-saving algorithms.

References

1. Srinivas, M., Patnaik, L.M.: Genetic algorithms: a survey. Computer 27(6), 17–26 (1994)
2. Wang, S., Yang, B., Niu, X.: A Secure Steganography Method based on Genetic Algorithm. Journal of Information Hiding and Multimedia Signal Processing 1(1), 8 (2010)
3. Ruiz-Torrubiano, R., Suarez, A.: Hybrid Approaches and Dimensionality Reduction for Portfolio Selection with Cardinality Constraints. IEEE Computational Intelligence Magazine 5(2), 92–107 (2010)
4. Jui-Fang, C., Shu-Wei, H.: The Construction of Stock's Portfolios by Using Particle Swarm Optimization, pp. 390–390
5. Bajaj, P., Puranik, P., Abraham, A., Palsodkar, P., Deshmukh, A.: Human Perception-based Color Image Segmentation Using Comprehensive Learning Particle Swarm Optimization. Journal of Information Hiding and Multimedia Signal Processing 2(3), 227–235 (2011)
6. Pinto, P.C., Nagele, A., Dejori, M., Runkler, T.A., Sousa, J.M.C.: Using a Local Discovery Ant Algorithm for Bayesian Network Structure Learning. IEEE Transactions on Evolutionary Computation 13(4), 767–779 (2009)
7. Chouinard, J.-Y., Loukhaoukha, K., Taieb, M.H.: Optimal Image Watermarking Algorithm Based on LWT-SVD via Multi-objective Ant Colony Optimization. Journal of Information Hiding and Multimedia Signal Processing 2(4), 303–319 (2011)
8. Pan, Q.-K., Tasgetiren, M.F., Suganthan, P.N., Chua, T.J.: A discrete artificial bee colony algorithm for the lot-streaming flow shop scheduling problem. Inf. Sci. 181(12), 2455–2468 (2011)
9. Chu, S.-C., Tsai, P.W.: Computational Intelligence Based on the Behavior of Cats. International Journal of Innovative Computing, Information and Control 3(1), 8 (2006)
10. Wang, Z.-H., Chang, C.-C., Li, M.-C.: Optimizing least-significant-bit substitution using cat swarm optimization strategy. Inf. Sci. 192, 98–108 (2012)
11. Akyildiz, I.F., Weilian, S., Sankarasubramaniam, Y., Cayirci, E.: A survey on sensor networks. IEEE Communications Magazine 40(8), 102–114 (2002)
12. Gene, C., Wai-tian, T., Yoshimura, T.: Real-time video transport optimization using streaming agent over 3G wireless networks. IEEE Transactions on Multimedia 7(4), 777–785 (2005)
13. Pourmousavi, S.A., Nehrir, M.H., Colson, C.M., Caisheng, W.: Real-Time Energy Management of a Stand-Alone Hybrid Wind-Microturbine Energy System Using Particle Swarm Optimization. IEEE Transactions on Sustainable Energy 1(3), 193–201 (2010)
14. Norman, P.G.: The new AP101S general-purpose computer (GPC) for the space shuttle. Proceedings of the IEEE 75(3), 308–319 (1987)

15. Simpson, J.A., Hughes, B.L., Muth, J.F.: Smart Transmitters and Receivers for Underwater Free-Space Optical Communication. IEEE Journal on Selected Areas in Communications 30(5), 964–974 (2012)

16. Harik, G.R., Lobo, F.G., Goldberg, D.E.: The compact genetic algorithm. IEEE Transactions on Evolutionary Computation 3(4), 287–297 (1999)

17. Mininno, E., Neri, F., Cupertino, F., Naso, D.: Compact Differential Evolution. IEEE Transactions on Evolutionary Computation 15(1), 32–54 (2011)

18. Neri, F., Mininno, E., Iacca, G.: Compact Particle Swarm Optimization. Information Sciences 239, 96–121 (2013)

19. Yang, X.-S.: A New Metaheuristic Bat-Inspired Algorithm. In: González, J.R., Pelta, D.A., Cruz, C., Terrazas, G., Krasnogor, N. (eds.) NICSO 2010. SCI, vol. 284, pp. 65–74. Springer, Heidelberg (2010)

20. Tsai, P.W., Pan, J.S., Liao, B.Y., Tsai, M.J., Istanda, V.: Bat Algorithm Inspired Algorithm for Solving Numerical Optimization Problems. Applied Mechanics and Materials 148-149, 134–137 (2012)

21. Pearson, K.: The Problem of the Random Walk. Nature, 72 (1905)

22. Pemantle, R.: A survey of random processes with reinforcement. Probability Surveys 4(2007), 9 (2007)

23. Billingsley, P.: Probability and Measure. John Wiley and Sons (1979)

24. Cody, W.J.: Rational Chebyshev approximations for the error function. Mathematics of Computation 23(107), 631–637 (1969)

25. Mininno, E., Cupertino, F., Naso, D.: Real-Valued Compact Genetic Algorithms for Embedded Microcontroller Optimization. IEEE Transactions on Evolutionary Computation 12(2), 203–219 (2008)

The New Procurement System Based on MRP Algorithm

Lei Meng and Chuansheng Zhou

Software College, Shenyang Normal University, Shen Yang, 110142, China
netmenglei@126.com, 252752602@qq.com

Abstract. To a good number of medium and small enterprises, it is a serious problem how to decline production cost and improve the response to the market by normative management and sharing the information. As an advanced management method and an IT tool, ERP can solve all these problems well. And how to combine ERP with other advanced management methods (such as SCM) and make ERP fit to the mid-small enterprises becomes a hot question recently. Purchase stands on the upstream of SC and takes a strategic effect on cost and quality control. Advanced and efficient purchase management not only can supply continuous material to keep enterprise run perfectly but also can make enterprise more competitive in controlling the whole SC. MRP algorithm designed in this paper can be a good application of procurement management in the ERP system, and achieved good economic.

Keywords: MRP, SCM, DWT, production, Algorithm.

1 Introduction

Human society has entered the information age, with social and economic development and scientific and technological progress, especially the rapid development of information technology, information resources and their management have become an increasing concern all sectors of society in general, human and efficient acquisition, processing, transmission and the ability to greatly enhance the use of information, making information and knowledge in human life and economic activities more and more prominent role. Information resources is a very important strategic resources, effective use of information resources to promote social and economic development has become an important part. For the vast majority of enterprises in China, give them the information age has brought unprecedented opportunities for development, so that they face new challenges. In the market economy, increasing competition among enterprises, the traditional way by hand more and more management business model can not meet the needs of enterprise survival and development, can not adapt to fast-paced changes in the information age, it can not meet the market rapid response to changing requirements. In this case, only the full use of information technology enterprises, the use of modern management methods to improve the overall quality of employees, the use of modern means of purchase, sales, inventory and other operations management, advanced information technology and advanced management science integration in order to truly improve the operational efficiency of enterprises to maximize existing equipment, resources, people, technology, role in promoting the development of enterprises, enhance

their competitiveness and create more economic benefits, which is to vigorously promote enterprise information. Enterprise information is the use of information and information technology to improve production and efficiency of management decision-making process, thereby enhancing competitiveness and economic efficiency, and promote information technology is to strengthen the popularization and application of information technology to promote the development and use of information resources. Advance information of all the content and process are to develop information resources as a starting point on, he use of information resources for the ultimate home. Enterprise information resources are very rich, such as market information products and design information, equipment, materials, technology and equipment and various types of fixed information appliances, quality information, stock information, capital and financial information, sales information, supplier and customer information, statistical information and decision-making information. Production and operation management of the process is essentially a variety of information processing and delivery of data generation and use of the process, the product is the ultimate physical embodiment [1].

2   The Main Principles of ERP Modules

2.1   MPS Works

Manufacturing production plans (Production Plans), a factory manufacturing activities guiding framework, it focuses on the required business plan and production facilities. The scheduled time is usually longer (2-5 years), the number and types of products in the enterprise too much, mostly by product category rather than individual products according to plan, plan to work properly simplified. When used in production planning cell (Time Bucket) is often the month.

Production and marketing schedule (Master Production Scheduling, referred to as MPS), also known as master production scheduling, detailed production plan is the expression usually used by the MPS for the week or day when the grid, covering the time is often three to twelve months to individual products for the object, that is, the products specified in MPS must be specific, individual BOM described [2]. In principle, it is the content of the production plan to do further subdivided and received.

2.2   BOM Functionality

Using computer-aided production management, first get the computer to read out the composition of products manufactured by enterprises and all the materials to be involved, in order to facilitate computer recognition, must be expressed with the icon into the product structure of a data format, such the data format to describe the product structure of the document is a list of materials, that is, BOM (Bill-Of-Material). It is the technical definition of product structure file, so it is called the product structure and product structure tree table. In some industries, may be called the "recipe", "elements of the table" or other name [3].

In the MRP Ⅱ and ERP systems, the material has a broad meaning of the word, it is all products, semi-finished products, work in process, raw materials, parts, collaboration parts, consumables, etc. and production of related materials collectively.

In the usual MRP Ⅱ and ERP systems is the BOM components and sub-components by the parents formed tree. BOM can be decomposed in the form of top-down or bottom-up track to the form of information.

In the ERP system, MRP Ⅱ and BOM is a relationship between the data organization, hierarchical relationship between the use of these data can be used as the basis for many features modular design, some form of these data is a summary of all of us are familiar with the report.

BOM is the MRP Ⅱ / ERP information system is the most important basic data, reasonable or not their organizational format is designed to directly affect the system's processing performance, therefore, according to the actual use of the environment, flexible and reasonable and effective BOM design is very important.

BOM MRP Ⅱ system is not only important input data and cost accounting of the financial sector, manufacturing and production in an important basis, therefore, BOM's impact was greatest, it also had the highest accuracy requirements. Proper use and maintenance of BOM management system is very important work during the operation.

2.3   MRP Decomposition

Manufacturing operations in a large number of materials used in the demand is certain to produce material goods containing these materials determined. In fact these materials, not to sustain a uniform rate is used, and they are composed of material goods is not required prior to production.

MRP system in Fig 1 is a schematic diagram [4]:
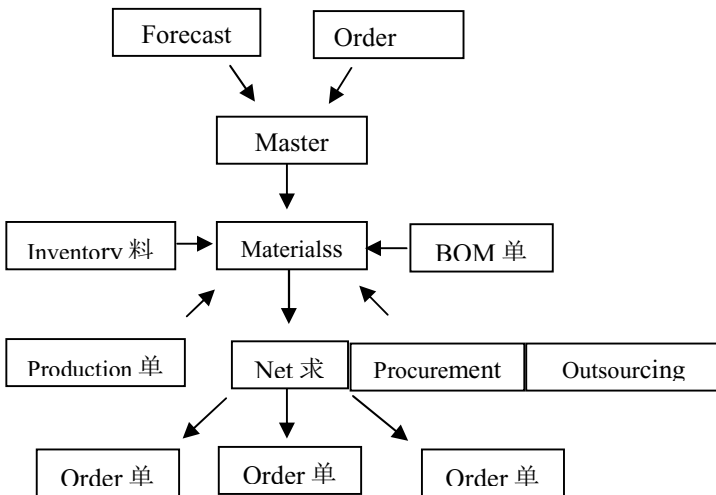
Fig. 1. MRP Principle chart

3   Algorithm Design

3.1   MRP Algorithm Flowchart
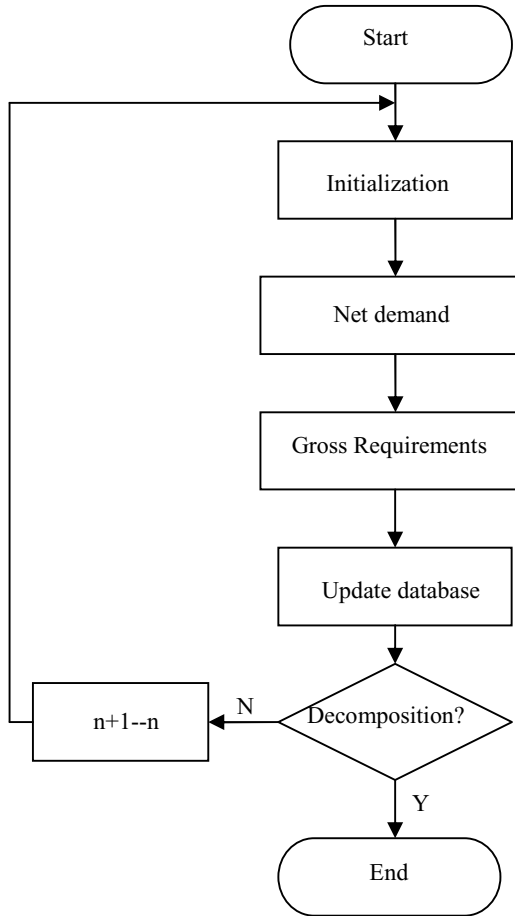
Fig. 2. Long-lost manufacturing MRP in the mode calculation flow chart

3.2   Algorithm Steps

Remove the master production schedule based on the final demand for the product number.

First of all levels of the table followed by BOM calculated the net demand for each component, the net demand in the process of seeking attention compared to the lower parts of the code is equal to the current level of N. If the code is not equal to the current level of level N, the time being, does not deal with the material requirements, and then continue down the decomposition.

Find the components of net requirements make the appropriate data into the temporary table to the database and find the corresponding gross demand.

Update the database.

BOM table to determine whether the complete decomposition of all levels, if not complete, repeat steps 2-4.

4   Conclusion

Certain enterprise resource planning (ERP) has begun the next chapter; the title is "ERP II". Companies are beginning to own from a focus on optimization of the internal functions of the vertical integration organization, into a more flexible core competency-based entities, to enable enterprises in the supply chain and value network to find the best Ding Wei. The main features of this positioning in the B2B and B2C e-only transactions, but also participate in collaborative commerce (c-commerce) process. MRP algorithm written in this article, it is consistent with the development trend of ERP and achieved good results.

References

1. Gan, L.: Construction and management of enterprise information. University Press, Beijing (2002)
2. Wang, H.: Enterprise Application MRPII / ERP Theory and Case Study. University Press, Beijing (2000)
3. Zeng, F.H., Wu, G.Y.: Machinery & Electronics, pp. 535–538 (2001)
4. Su, Y.Z.: Limited Partnership, pp. 20–23 (2001)

Part II

Context Awareness Services and Intelligent Computing Applications

Wireless Sensor and Mobile Application of an Agriculture Security System

Chun-Chieh Fan[*], Rong-Hou Wu, Liang-Lin Jau, and Yu-Ming Li

Department of Computer and Communication Engineering,
St. John's University, 499, Sec.4, Tam King Road,
Tamsui Dist., New Taipei City 25135, Taiwan, R.O.C.
`chun@mail.sju.edu.tw`

Abstract. This study uses wireless sensor and mobile APP technologies to plan and build an agriculture security system to prevent farmers' losses. This security system consists of a variety of different sensor functions. The data collected will be analyzed and judged through the Arduino microprocessor and the information is sent to the control end by the ZigBee wireless transmission module. The control end will activate all buzzers and the searchlight if an invasion has been confirmed. Also, the installation of the CCD camera will activate when suspicious activity occurs and will thus take clear and definite photos as evidence and send an email to notify the user. Using the mobile APP program developed by ourselves, when the smart phone APP detects new photos being stored in the cloud database, it sends an alert message to the user to remind him/her to check the new photo.

Keywords: wireless sensor, mobile APP, microprocessors, ZigBee wireless transmission module, smart phone.

1 Introduction

In recent years, the weather globally has been going through drastic changes. As a result, food supply shortage is also becoming more apparent. In addition, hardworking farmers often suffer heavy losses as their crops are often stolen by thieves. Police reports that in recent years, agriculture related thefts are rising so there is a high demand for agricultural security systems. This can create a large market.

Wireless sensor network techniques (WSN)[1,2,3] are becoming popular and a variety of different sensors are developed, such as temperature/humidity, pressure, and photo-frequency sensors, making WSN applications more extensive. ZigBee is a wireless network protocol defined by ZigBee Alliance[4]. It is currently the most common technique used. It has characteristics of low complexity, low cost, and energy saving. According to the Gartner report, 210 million Smartphone were sold in the first season in 2013. Therefore, various smart phone APP are being developed

---

[*] Corresponding author.

continuously. To improve farmers' life, we develop an agriculture security system to prevent the farmers' losses.

2   Introduction of Related Techniques

2.1   Introduction of Related Sensors

A variety of different sensors incorporated in this agricultural security system are introduced below.

2.1.1   PIR Body Sensor

Pyro-electric Infrared Detector (PIR) body sensor is shown in Fig.1. A multilayer coating on the sensor cover can block most of the Infrared wavelength and only let wavelength of Infrared temperature near 36.5 degree to pass. Therefore, this is suitable for human body motion detection. A detection angle in the front is around 120 degree and the detection distance is around 6 meter.
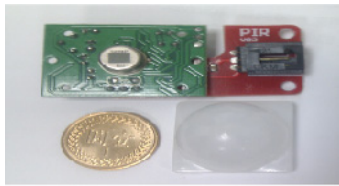
Fig. 1. PIR body sensor

2.1.2   FSR Pressure Sensor

Pressure sensor is a pressure sensitive resistor as shown in Fig.2. Reaction distribution area is 0.25*24 inch. Reaction pressure is detected by the change of the resistor value. The resistor value of FSR408 is greater than $1M\Omega$ when there is no pressure and the ideal reaction pressure value is between 100g and 10Kg.
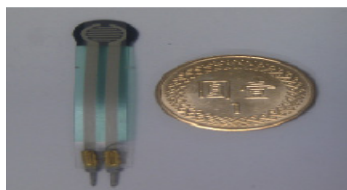
Fig. 2. FSR pressure sensor

2.1.3   SW-180 Vibration Sensor

A root of metallic stick is held between two electrodes of the mini vibration sensor, as shown in Fig.3. When there is no vibration or inclination, the resistor of the sensor is stable. However, when vibration is detected, the conduct resistor value increases

rapidly. The signal amplitude increases in direct proportion to the amount of vibration, thus triggering the circuit.
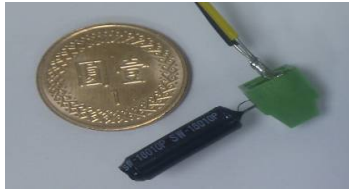
Fig. 3. SW-180 vibration sensor

2.1.4   Bending Rate Detection Sensor

Bending rate detection sensor is a simple Flex transducer as shown in Fig.4. It uses different of bending angle to change the resistor value.
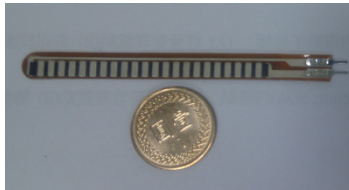
Fig. 4. Bending rate detection sensor

2.1.5   X-band Motion Detector

X-band motion detector, as shown in Fig.5, uses 10.525GHz of X-band frequency to detect the presence of absence of relative motion of an object. It also produces output of high/low waveform to show the motion speed. X-band motion detector has a general application in security systems and automatic door opening detection. It can detect motions indoor, a general area, and even motion detection on the other side of a partition wall.

Fig. 5. X-band motion detector

2.2   Single-Chip System Arduino Develop Board

This study adopts the single-chip system Arduino develop board [5]. The develop board has wide range of function and uses low cost of Atmel Mega single-chip as core. This development board can supply digital I/O, analog I/O, UART, ISP, source code, and circuit board.

2.3   Introduction of XBee Wireless Transmission Module

XBee is designed by Digi International Company for network standard module and is manufactured by Maxstream Company. XBee passed the IEEE 802.15.4 protocol of 2.4GHz RF wireless transmission module and it also holds all advantages of WSN, such as small volume of hardware, low energy consumption, high stability, and low cost. The transmission range of indoors and outdoors are 30 and 150 meters, respectively, due to the low energy consumption.

2.4   Cloud End Database

The Dropbox [6] is used in this study. Dropbox is a free of charge and popular cloud storage space. It can operate on Windows, Mac, and Linux operation system. Beside Dropbox, there are other clouds such as Google cloud hard disk、SkyDrive, and SugarSync.

3   System Structure and System Operating

3.1   System Structure

Figure 6 shows the system structure of an agriculture security system builds in four parts: sensor, alarm, central control, and mobile device application. The main sensor is composed of 5 different sensor components (PIR body infrared sensor, FSR pressure sensor, SW-18010P vibration sensor, bending rate detector, and X-band motion detector) and Arduino ATMEGA328 single-chip microprocessor and XBee wireless transmission module. If an invasion has been confirmed by the sensor, Arduino single-chip will generate a control package signal and broadcast to the alert and host control end via XBee wireless transmission module. The alert end is composed of LED, buzzer, CCD camera, server, Arduino single-chip, Blue-tooth speech system, and XBee wireless transmission module. When XBee wireless transmission module of alert end receives the control signal from sensor end, the Arduino single-chip will start the relative alert device. Photos taken by CCD camera at alert end will be transmitted and stored into cloud database. The host end use XBee wireless transmission module to receive control signals from the sensor and display these signals on C# interface. Afterwards, these data are stored in host database and also into cloud database for convenient access anytime and from any device. If needed, the host end can transmit an E-mail to Google G-mail so the use can receive the real-time situation at the farm, granted that the user's mobile device and G-mail are synchronized. Besides receiving E-mail, mobile devices (smart phones) can also use the mobile APP program that we have developed to receive notifications when new photos are being stored in the cloud database. This will remind the user to check the photos.
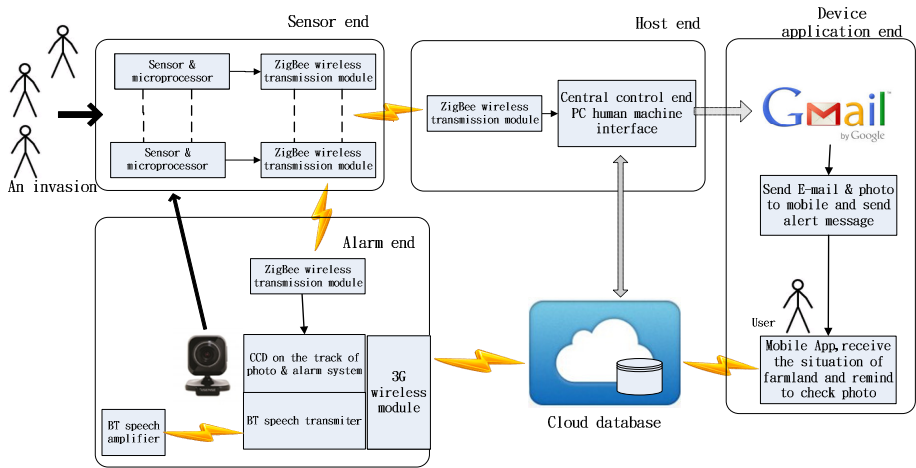
Fig. 6. The system structure of an agriculture security system

3.2   System Operating Procedures

The operating procedures of the agriculture security system is explained and illustrated below.

3.2.1   Arrangement Procedures of the Sensor End of Agriculture Security System

Figure 7 shows arrangement procedures of agriculture security system of sensor end.
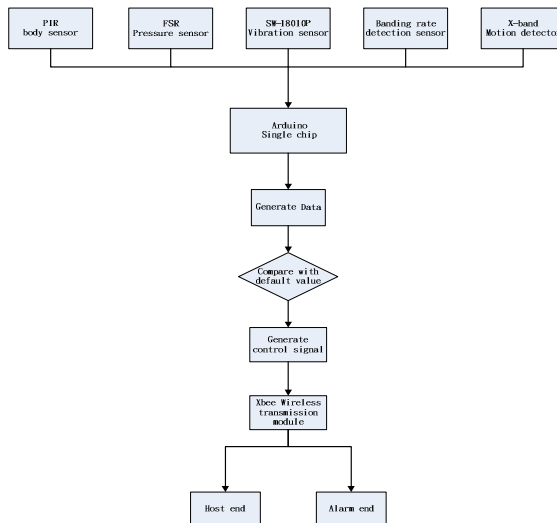
Fig. 7. Arrangement procedures of sensor end

3.2.2   Arrangement Procedures of Alarm End of Agriculture Security System

Figure 8 shows arrangement procedures of agriculture security system of alarm end.
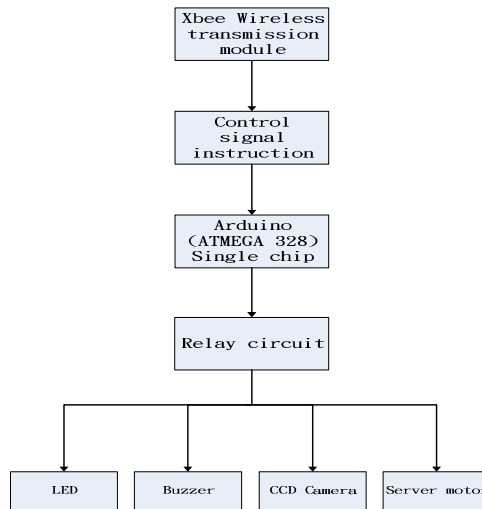
Fig. 8. Arrangement procedures of alarm end

3.2.3   Host End of Agricultural Security System

The main function of host end is to transmit signals from sensor end at the farmland to the host computer through XBee wireless transmission module. These signals are carefully viewed and documented by means of C# of the human machine interface (HMI). Then they are saved at the host end Access database to allow the user to conveniently know the exact time each sensor has been triggered.

4   Results from System Implementation

Hardware practical of agriculture security system along with software design and program development is introduced in this section.

4.1   Hardware Used in the Agriculture Security System

A sketch of whole installation of agriculture security system is shown in Fig. 9. Fig.10 and 11 are separate parts physical hardware.

Fig. 9. A sketch of whole installation of agriculture security system
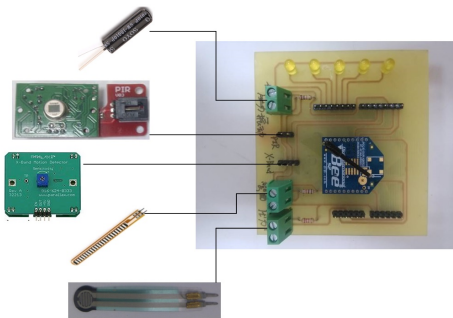
Fig. 10. A physical hardware of sensor end   Fig. 11. A physical hardware of alarm end
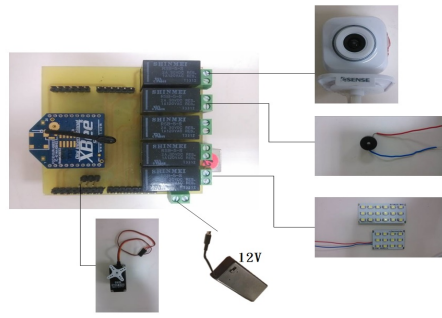
4.2   Human Machine Interface of Agriculture Security System

When host end initiates C# interface software, XBee connects to a computer through the TTL connection board. Fig.12 shows the detection result after COMPORT is chosen and Fig. 13 shows the sensor state of access database.
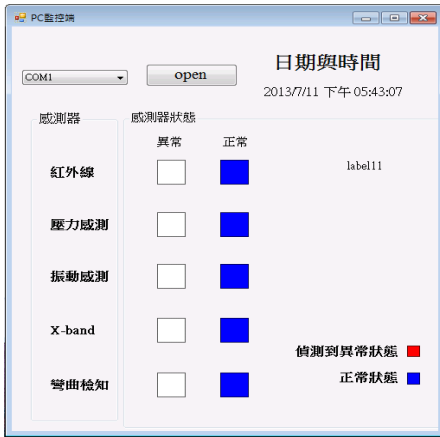
Fig. 12. Interface of the agriculture security system

Fig. 13. The sensor state of access database

4.3   Mobile Device End of Agriculture Security System

When there's detection of human invasion into the farmland, the central control of the computer will automatically send an email to notify the user. If the user's mobile device and G-mail are synchronized, a notification can be received along with description of the message in the email, as illustrated in figure 14.

Fig. 14. The contents of the E-mail transmitted from the agriculture security system

In addition to sending via E-mail, photos taken by CCD will be sent and saved in a cloud database. Using the mobile APP program developed by ourselves, when the smart phone APP detects new photos being stored in the cloud database, it sends an alert message to the user to remind him/her to check the new photo, as illustrated in figure 15.
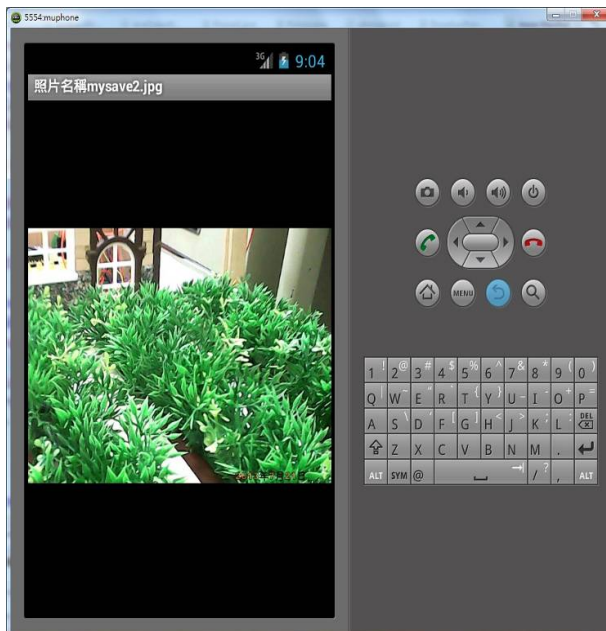
Fig. 15. Display of CCD photo on mobile device

5   Conclusion

This study simulates an orchard and a farmland. We have successfully built numerous sensors and ZigBee wireless transmission module to detect and send information to the central control end. If it has been verified that a thief has entered the premise, an alarm or a notification via APP will be sent to the user. This system has the function of guarding and scaring off any intruders. In addition, the system built in this study for sensor detection and verification can be used as a reference for other related studies.

References

1. Robert, F.: Building Wireless Sensor Network. O'Relly (2011)
2. Gutierrez, J.A.: On the use of IEEE 802.15.4 to enable wireless sensor network in building automation. In: Proceeding of IEEE PIMRC 2004, vol. 3, pp. 1865–1869 (2004)
3. Yau, C.L., Chung, W.Y.: IEEE 802.15.4 Wireless Mobile Application for Healthcare System. In: Proceeding of Convergence Information Technology 2007, pp. 1433–1438 (2007)
4. ZigBee Alliance: ZigBee Specification ZigBee Allience, Ver.1.0 (June 2005)
5. Massimo, B.: Getting Started with Arduino. O'Relly (2008)
6. Dropbox, http://www.dropbox.com

Knowledge Integration for Diabetes Drugs Ontology

Rung-Ching Chen, Yu-Wen Lo, Bo-Ying Liao, and Cho-Tscan Bau

Department of Information Management, Chaoyang University of Technology,
168, Jifeng E. Rd., Wufeng District, Taichung City 41349, Taiwan ROC
`crching@cyut.edu.tw`

Abstract. The rising and developing of information technologies has made information overflow. Even in the same topics, a lot of different aspects of knowledge were setup. So, the knowledge integration is one of the important research topics. In this paper, the diabetes drug prescriptions are used as examples to do the knowledge integration which according to American Diabetes Association (ADA), American Journal of Clinical Endocrinology Society (AACE), the Republic of China Diabetes Association and the British National Health Service Bureau (NHS). The system will integrate the four medication diabetes associations to establish knowledge ontologies. The system includes three parts. First, the ontologies pre-processing will calculate the similarity between the ontologies and then find out the correlation between ontologies, Next, the system transfers the ontologies format into Joseki, and finally the user through a graphical user interface to obtain information.

Keywords: Ontology, Ontology Integration, Diabetes Drugs, OWL, SPARQL.

1 Introduction

The knowledge may come from different sources in a particular area such as medical, business and so on. Therefore, the knowledge presentation and structure will be different from various people. So it is important to solve communication barriers from knowledge integration. Then it can achieve to share the knowledge. Through information technology, knowledge can be presented by ontology. At present there are many specific areas of the ontology of knowledge are presented. The main benefits are to find the relationships of knowledge.

The term of ontology is from the philosophy field which refers to "a systematic approach to explain the existence of things in the world" [1]. Ontology can also be used in different fields such as artificial intelligence, semantic web, systems engineering, software engineering and so on. In this paper, we use ontology as a knowledge expression tool. The advantage of ontologies, compared to other data structuring technologies, is the ability to reason upon data. By adding rules and logic it is possible to merge data from different sources, extract and combine data in new ways to produce new information. In general, the elements of ontology include class, attribute and instance [2,3]. Class defines the general things, such as diabetes, diabetes drugs, drugs testing and so on. Attribute: It describes the property of class or the relationship between the concepts. In addition, the relationship between the superclass and the

subclass is also the property. Instance is the entity of concept or class in the ontology. It would inherit all the properties or relation of their class. Nicola [4] proposed the definition of ontology can be based on different conditions to give the different structure and application.

The diabetes drugs are divided into six categories: (1) biguanides, (2) sulfonylurea, (3) DDP4 (DPP4 inhibitor), (4) thiazolidnedione, (5) alpha-glucosidase, (6) meglitinide in Taiwan. Therefore, the six categories of drugs are only considered in this paper[5,6]. In addition, the diabetes drugs are divided into two types: oral hypoglycemic agent (OHA) and insulin. In this paper, only oral hypoglycemic agent to be considered because OHA is convenient to use. Although insulin can effectively control blood glucose, it must be long-term injection. In addition, patients are fear.

Ontology integrations are divided into three kinds: (1) ontology mapping, (2) ontology merging, and (3) ontology alignment. Ontology Mapping is also divided into two parts: One-way mapping and Two-way mapping. One-way mapping refers to ontology A replace ontology B; Two-way mapping refers to ontology A and ontology B can interact with each other. OWL is proposed by W3C and it combines the Darpa Agent Markup Language (DAML) and OIL. The main spirit of OWL is describing the concept of logic[7]. OWL can be divided into three sub-languages: OWL Lite, OWL DL and OWL Full. In this paper, we will use OWL DL to establish diabetes associations and protégé to setup the ontologies of domain knowledge.

In this paper, we will use the domain knowledge of diabetes as an example to do knowledge integration. The main reason of selecting the diabetes domain knowledge is that this disease is one of leading cause of death in the world which is ranked at the top five every year in Taiwan. In addition, diabetes medication knowledge is extensive and it hard to effectively interact and share. This paper will base on four associations included ADA, AACE, Diabetes Association of the Republic of China and NHS to do knowledge integration. We integrated the four medication diabetes associations to establish knowledge ontologies to recommend drugs to the doctors for better treatment for patients.

2   Methodology

Figure 1 shows the flow chart of the drugs recommendation system which divided into five steps:

(1) collect data, (2) create four diabetes associations ontologies, (3) calculate the similarity between the diabetes associations of ontologies, (4) evaluate the relationship between the diabetes associations.

Protégé is used to construct ontologies database in this paper. Protégé is an ontology editor developed by Standford University and the knowledge acquisition software. Protégé use JAVA programming language to develop ontology editor. It provides an interface to users for more flexible applications and access their database and supports RDF (Resource Description Framework). RDF is uesd to describe the relations between web pages and other resources. Protégé mainly has the advantages: easy to learn, free software, supporting a variety of ontologies, supporting multiple store formats, providing complete interface, and providing plug-in program 1. Protégé
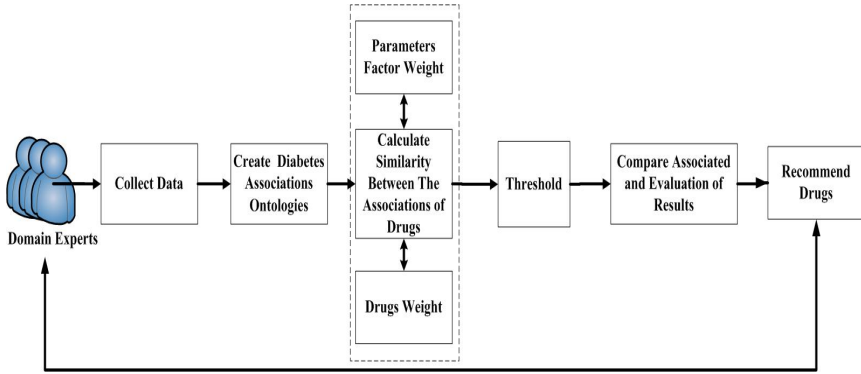
Fig. 1. The flow chart of the drugs recommend system

is an excellent design which allowed many plug-ins. Hence, it has become one of the most widely used ontology editors. In addition, the others more popular ontology editors are such as OntoEdit, OILED, WebODE and so on. In this paper, ontology composition is divided into four parts: (1) class, (2) instance, (3) property, (4) relationship.

2.1 Diabetes Drugs Knowledge Integration

Protégé is used to construct patient ontology and diabetes association ontology. And then we will use Jambalaya to show the various class relationships graph. This paper will use four diabetes association included ADA, AACE, Diabetes Association of the Republic of China and NHS to do knowledge integration. The system can help doctors to diagnose patients in the same situation can give the treatment of diabetes-related knowledge from different countries. It can give more accurate drugs or provide doctors as assistant in the diagnostic process. In this study, four diabetes associations are the main difference between them that they identified the range for glycated hemoglobin (or HbA1c) are different. In this paper, drugs administration divided into two parts: single drugs and combo drugs. Single drug have MET, DPP4, SU, Glinide, TZD and AGI; combo drugs have MET+DPP4, MET+TZD, MET+SU, MET+Glinide, MET+AGI, SU+DPP4, SU+TZD and TZD+DPP4.

The AACE, ADA, Diabetes Association of the Republic of China and NHS ontologies are sequential represented to O_1, O_2, O_3, and O_4.

In addition, the six categories of drugs for diabetes MET, DPP4, SU, Glinide, TZD and AGI are sequential defined as X_{m1}, X_{m2}, X_{m3}, X_{m4}, X_{m5}, and X_{m6}. And then we will define the weight of drugs as W_{m1}, W_{m2}, W_{m3}, W_{m4}, W_{m5}, and W_{m6}. Table 1 shows the symbol of drugs and weights. Each diabetes association of the drugs weight are sum up to one. For example: the sum weight of drugs in O_1 are $W_{m1}+ W_{m2}+ W_{m3},+W_{m4},+W_{m5}+W_{m6} =1$.

Table 1. The definition of drugs and weights

Drug	MET	DPP4	SU	Glinide	TZD	AGI
Symbol	X_{m1}	X_{m2}	X_{m3}	X_{m4}	X_{m5}	X_{m6}
Weight	W_{m1}	W_{m2}	W_{m3}	W_{m4}	W_{m5}	W_{m6}

First, the formula (1) will be used calculate the diabetes association all parameters between the single drug factor values. It's not pure single drug to give to patients if the patient's condition is more serious, the patients will be given the combo drugs. Formula (2) calculates the diabetes association for all parameters between the combo drugs factor values. In this paper, we are focused on single drug and combo drugs.

The parameters have fourteen: blood glucose, liver fat, BMI, waistline, triglycerides (TG), cholesterol (or HDL), hypoglycemia, gastrointestinal dysfunction, renal function indices, liver cirrhosis, liver function, heart failure, fracture risk, and glycated hemoglobin (or HbA1c). The parameter factors are defined as Y_1, Y_2, Y_3, …, Y_j.

$$sum(X_{mi}) = w_{mi} \sum_{j=1}^{n} X_{mi} w_j Y_j \qquad (1)$$

$$w_{m1} + w_{m2} + w_{m3} + \cdots + w_{mi} = 1$$

W_{mi} is drug weight, W_j is parameter weight, X_{mi} is drug value, and Y_j is parameter value.

$$sum(X_{mi}) = w_{mi_1} \sum_{j=1}^{n} X_{mi_1} w_j Y_j + w_{mi_2} \sum_{j=1}^{n} X_{mi_2} w_j Y_j \qquad (2)$$

X_{mi1} and X_{mi2} are two of the six categories of drugs, W_{mi1} and W_{mi2} are the weights for the two types of drugs.

After individual diabetes association was evaluated, the system will calculate similarity measurement of four diabetes associations of ontology with the conditions of combination 11 kinds: $sim(O_1,O_2)$, $sim(O_1,O_3)$, $sim(O_1,O_4)$, $sim(O_2,O_3)$, $sim(O_2,O_4)$, $sim(O_3,O_4)$, $sim(O_1,O_2,O_3)$, $sim(O_1,O_2,O_4)$, $sim(O_1,O_3,O_4)$, $sim(O_2,O_3,O_4)$ and $sim(O_1,O_2,O_3,O_4)$. Then, the system calculates the single drug similarity measurement of the two ontologies by the formula (3), and the formula (4) will be used to measure the similarity of combo drugs from two ontologies. In addition, the formula (5) will be used to calculate the single drug and combo drugs similarity measurement form three ontologies, and the formula (6) will be used to calculate the single drug and combo drugs similarity measurement from four ontologies.

$$sim(O_{as}, O_{bs}) = \frac{O_{as}\left(w_{mi} \sum_{j=1}^{n} X_{mi} w_j Y_j\right)}{O_{bs}\left(w_{mi} \sum_{j=1}^{n} X_{mi} w_j Y_j\right)} \qquad (3)$$

$sim(O_{as},O_{bs})$ is selected the similarity of single drug from two ontologies.

$$sim(O_{ac}, O_{bc}) = \frac{O_{ac}(w_{mi\,1}\sum_{j=1}^{n} X_{mi_1} w_j Y_j + w_{mi_2}\sum_{j=1}^{n} X_{mi_2} w_j Y_j)}{O_{bc}(w_{mi\,1}\sum_{j=1}^{n} X_{mi_1} w_j Y_j + w_{mi_2}\sum_{j=1}^{n} X_{mi_2} w_j Y_j)} \tag{4}$$

$sim(O_{ac}, O_{bc})$ is selected the similarity of combo drugs from two ontologies .

$$sim(O_a, O_b, O_c) = \frac{sim(O_a, O_b) + sim(O_a, O_c) + sim(O_b, O_c)}{3} \tag{5}$$

$sim(O_a, O_b, O_c)$ is selected the similarity of single drug and combo drugs from three ontologies.

$$sim(O_a, O_b, O_c, O_d) = \frac{sim(O_a, O_b, O_c) + sim(O_a, O_b, O_d) + sim(O_a, O_c, O_d) + sim(O_b, O_c, O_d)}{4} \tag{6}$$

$sim(O_a, O_b, O_c, O_d)$ is selected the similarity of single drug and combo drugs form four ontologies.

2.2   Ontology Mapping Algorithm

In this paper, we will use two algorithms to calculate the similarity of single drug and the combo drugs respectively. They show the results of the ontology mapping. And then they will recommend the availability of drugs. The ontology mapping between the single drug algorithm will input O_1, O_2, O_3, and O_4 parameter factors of drugs; output the single drug

---

Input: O_1, O_2, O_3, and O_4 parameter factors of drugs
Output: The single drug of ontology mapping results and recommendation drug **Step 1:** extract the ontologies;
  Step 2: calculate individual the single drug values for all parameter factors by the formula (1) and $w_{m1} + w_{m2} + w_{m3} + \ldots + w_{mi} = 1$;
Step 3: repeat Step 1 and Step 2, and calculate all the possibilities by the formula $C_2^4 + C_3^4 + C_4^4$, it can find eleven similar combinations;
Step 4: use formula (3) to compare the two of ontologies, formula (5) to compare the three of ontologies and formula (6) to compare the four of ontologies, then $O_{as}(w_{mi}\sum_{j=1}^{n} X_{mi} w_j Y_j)$

and $O_{bs}(w_{mi}\sum_{j=1}^{n} X_{mi} w_j Y_j)$ are random selected ontologies;
Step 5: repeat Step 4, until the completion of the eleven combinations of the calculated;
Step 6: substitute the single drug threshold (T);
Step 7: overtake the threshold value representation the correlation between each other and it's recommended the single drug as the first choice, then it is marked as (**), otherwise, lower than threshold value is recommended as the second choice and it is marked as (*). If the drugs between the ontologies do not be used, this drug is marked as (-);
Step 8: return the single drug of ontology mapping result and recommendation drugs.

Fig. 2. The ontology mapping between the single drug algorithm

of ontology mapping results and recommendation drug. Figure 2 shows the ontology mapping between the single drug algorithms. The ontology mapping between the combo drugs algorithm will input O_1, O_2, O_3, and O_4 parameter factors of drugs; output the combo drugs of ontology mapping results and recommendation drug.

It includes eight steps: (1) extract the ontologies, (2) calculate individual the single drug values for all parameter factors, (3) calculate all the possibilities, (4) use formula (3),(5),(6) to compare all ontologies combination, (5) repeat Step 4, until the completion of the eleven combinations of the calculated, (6) substitute the single drug threshold (T), (7) overtake the threshold value representation the correlation between each other and it's recommended the single drug as the first choice, then it is marked as (**), otherwise, lower than threshold value is recommended as the second choice and it is marked as (*). If the drugs between the ontologies do not be used, this drug is marked as (-), and (8) return the single drug of ontology mapping result and recommendation drugs.

3   The Experiments

The experiments uses Intel(R) Core(TM) Duo CPU P7450 2.13GHz and 4.00 GB RAM. The operation system is Windows 7. Protégé was the platform for constructing ontology. Joseki was used to build the query service. Microsoft Visual Studio 2008 was utilized to develop the user interface system.

This paper will provide recommendation drugs from four diabetes association for doctors. They were based on fourteen different parameter factors to give the drugs in different circumstances. The user interface included three query parts: category query, drugs query and the drugs recommendation. Category query is the overall introduction of the system query and the users can understand the system operation; drugs query is the single drug and combo drugs recommendation results in any parameter factors; the drugs recommendation is shown 11 kinds comparison results.

For drugs query, there are forty-three options. They are included BMI-N (Single Drug), BMI-N (Combo Drugs), Cirrhosis (Single Drug), Cirrhosis (Combo Drugs) and so on. Figure 3 shows the drugs query options. In this paper, three examples are used to explanation for the drugs query. First, it selected the "BMI-N (Single Drug)" to execute. The "BMI-N (Single Drug)" SPARQL query syntax content included three parts: "prefix default: <http://www.owl-ontologies.com/Ontology1305864796.owl#>" is defined to use the prefix, "select ?Parameters ?Drugs" representation wants to query the information, and "where{ {?Parameters a default:BMI} {?Parameters default:single_drug ?Drugs } FILTER regex(str(?Parameters),"BMI-N")}" represents to find BMI parameter, and then find the "default:single_drug" associated with the information between parameters and drugs. In addition, it filters BMI to find subclass BMI-N. Figure 4 shows the "BMI-N (Single Drug)" query result. It displays recommendation drugs of the four diabetes association such as the single drug of BMI-N can use AGI, DPP4 and MET in AACE.
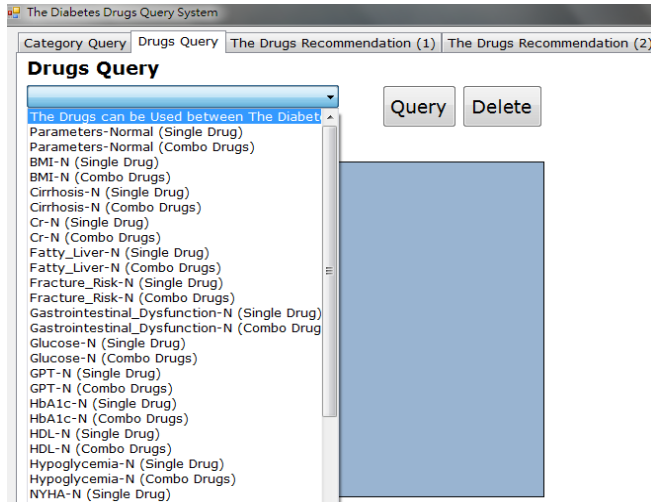
Fig. 3. The drugs query options
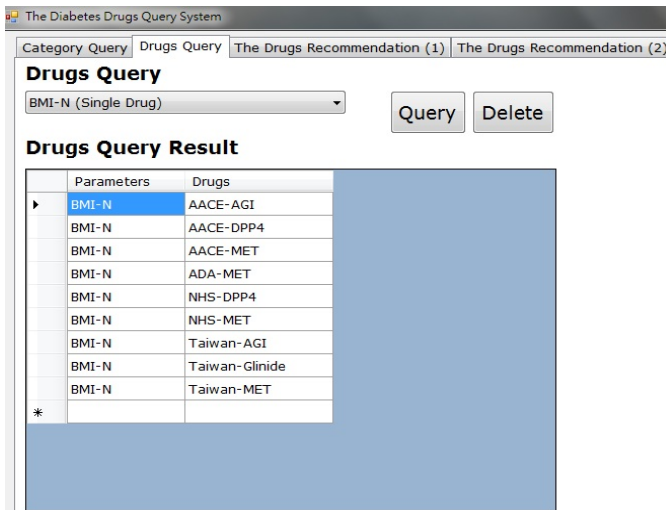
Fig. 4. The "BMI-N (Single Drug)" query result

4   Conclusions and Future Works

Due to the current in the same topics, there are many relevant domains knowledge which come from different countries, domain experts, systems and so on. Therefore, the various related domains knowledge integration is an important work. Ontology can

explain more clearly between the concepts in the particular domain. Hence, we use Protégé to construct domain ontology. In this thesis, the main purpose is to complete the four diabetes drugs knowledge integration and compare similarity between them. In addition, we implement an user interface to help doctors more easier to query diabetes medication knowledge. In this way, our method can help physicians to get more medication knowledge and improve the diagnosis occasion would be more accurate in the drugs administration. In the future works, we will research the triple drugs recommendation and integration. And then we will try to add more medication knowledge among diabetes associations.

References

1. Wu, Y.J., Lang, J.S., Shang, F.H.: A Similarity-Based Approach for Ontology Mapping. In: International Conference on Computer Science & Education, pp. 165–169 (2009)
2. Thorsen, K.A.H., Eftestøl, T., Tøssebro, E., Rong, C., Steen, P.A.: Using Ontologies to Integrate and Share Resuscitation Data From Diverse Medical Devices. Resuscitation 80(5), 511–516 (2009)
3. Nicola, G.: Formal Ontology and Information Systems. In: Proceedings of FOIS 1998, Trento, Italy, pp. 3–15. IOS Press (1998)
4. Mario, C., Pietro, H.G., Pierangelo, V.: Using Ontologies for Querying and Analysing Protein-Protein Interaction Data. Procedia Computer Science 1(1), 991–998 (2010)
5. http://www.wretch.cc/blog/f92420f2/10044728
6. http://www.kmu.edu.tw/~kmcj/data/9306/16.htm
7. Corsar, D., Sleeman, D.: Reusing JessTab Rules in Protégé. Knowledge-Based Systems 19(5), 291–297 (2006)

Increasing Customer Loyalty in Internet Marketing

Long-Sheng Chen[*] and Tzung-Yu Kevin Yang

Department of Information Management,
Chaoyang University of Technology, Taichung, Taiwan
{lschen,s10114616}@cyut.edu.tw

Abstract. In recent years, with rapid development of social networking websites, more and more travelers make their travel and accommodation decisions by referring to online comments (electronic word of mouth). It's especially true for the customers who live in bed and breakfast. But, due to the limited marketing budget, a bed and breakfast (B&B) enterprise needs an effective and cheap to promote their products and services through internet marketing. Social media marketing could one of cheap and powerful internet advertising channels. However, most of bed and breakfast enterprises lack sufficient human resource and time to interact to the online users of social networking websites. Moreover, there are lots of social media marketing techniques, but we don't know which one is crucial for a bed and breakfast enterprise. Therefore, this study aims to define the key factors of social media marketing, and then use decision tree to identify the important factors for increasing customers' loyalty. A survey of social media marketing in Facebook will be provided to demonstrate the effectiveness of our utilized methods.

Keywords: Internet Marketing, Electronic Word of Mouth, Decision Trees, Feature Selection.

1 Introduction

With rapid development of social networking websites such as Facebook, Trip Advisor, and Twitter, more and more travelers make their travel and accommodation decisions by referring to online comments (electronic word of mouth) [27]. It's especially true for the customers who live in bed and breakfast which has been of major travel trends in Taiwan according to the 2012 survey of Tourism Bureau, Taiwan. But, due to the limited marketing budget, a bed and breakfast (B&B) enterprise needs an effective and cheap to promote their products and services through internet marketing.

In addition, social networking websites has become an important source of information. Travelers will share their travel experiences to others and respond to related comments. This many-to-many interaction spread through the internet can be called electronic word of mouth (e-WOM) [6, 7]. This kind of trusted information has been produced by the trusting relationship between/among community members. In Taiwan, 75% of consumers trust the opinions of Internet users. Before making their purchase decisions, about 80% of consumers will refer to other comments on the Internet and as a basis for

---

[*] Corresponding author.

future purchase reference [12, 16, 18]. Many companies also interact to consumers and understand their needs by social media. So, we can say that the rapidly growing popularity of social networking sites has changed the traditional marketing content [24].

According to the report of 2013 social media marketing industries [25], the most frequently used social media is Facebook (92%), then are Twitter(80%), Linkedln (70%), Blog (58%), and Youtube(56%). The current status of the use of social media marketing, 79% of marketing companies integrate social media into traditional marketing. The benefits about social media marketing, 89% respondents said the efforts they made on social media can generate more exposure for businesses. 75% think there is a positive effect of increased customer traffic, again providing market information (69%), the establishment of fan loyalty (65%), resulting in potential customers (61%), etc. However, sales are driven by long-time relationships. Among those who have more than three years of experience in the social media industry, only 50% said social media could assist them to increase sales. And the performance of social media marketing is proportional to the numbers of hours which weekly input by companies. Related researches have shown that travelers will choose a tourist destination by searching others' experience via Facebook to compare features of different Bed and Breakfast. Thus, to interact with consumers through social media and on the field of tourism including B & B industry has become increasingly important [14].

However, limited by firm size, B & B industry generally lacks sufficient human resources and time for communicating to consumers, cultivating potential customers, and conduct social media marketing business. Therefore, understanding the important techniques of social media marketing practices and critical service quality can help B & B companies to make precise social media marketing decisions and obtain the voice of the customers.

Consequently, this study aims to define the factors (techniques) of social media marketing in B & B industry, and then utilize decision trees (DT) to recognize the crucial factors for building loyalty, respectively. Once we know the important factors of social media marketing, B & B enterprises can use their limited resource to do effective marketing and improve service quality. A survey of social media marketing in Facebook will be provided to demonstrate the effectiveness of our utilized methods.

2   Related Works

2.1   Internet Marketing

Social media is to achieve the purposes of dissemination by interaction so that people can share information and give comments, establish circle of friends, maintain relationships, and communicating to others [11, 21, 22, 40]. Following the growth of web 2.0 websites such as Blog, Twitter (information exchanging and sharing); Youtube (vedio and photo sharing); Wikipedia (knowledge sharing), this trend has continuously increases [11, 37]. Yu et al. [37] compared 4 types of media (blogs, forum, Twitter, conventional news) and they found social media outperforms conventional media.

Eduardo et al. [11] found that tourists using virtual communities to exchange views and experiences already exist for more than 10 years. Tourists will obtain information, learn the experiences of others, and compare tourism-related services before travelling or during the trip. Tan [32] mentioned that there are three motivating

factors of searching travel-related information, the exchange of information, enter-tainment, and social networking. Because travelers use the Internet to communicate with each other and share travel-related information, they can do detailed travel plans by understanding tourism destination through others shares [8, 22, 35]. Laroche et al. [20] developed a model to explain how social media community to strengthen the relationship between the brand and the customers. Boley et al. [3] compared the dif-ference of visitors' behaviors between the websites which did post pictures and those which didn't post pictures, when they buy souvenirs. The discovered information allows website owners to develop strategies to stimulate the effectiveness of their marketing. In the work of Tomas & Elena [29], they utilized modified balance score card to evaluate the hotel websites. They confirmed that the hotel website can build long-term customer relationships through the hotel's policy. Jacques et al. [15] uti-lized "hotel stars" and "hotel facilities (restaurant, business center, gym, free wireless internet)" to predict roads, destination types and etc.

To sum up, social media marketing must build a persuasive rhetoric, through the exchange of experience and quality, to influence customer loyalty and maintain long-term customer relationship [2]. However, these available literatures didn't focus on B & B industry. In addition, they cannot tell the enterprises how to use social media marketing and which social marketing techniques are important when they don't have enough time, budget and human resource. Therefore, it is necessary to discover the important factors which influence consumers' behaviors for tourism enterprises.

2.2   Feature Selection

You et al. [36] mentioned that there are three types of feature selection approaches. They are filters, wrappers, and embedded approaches. The filter based approaches are simple and quick. The wrappers based methods are dependent on selection functions in the feedback and learning of classifiers. Embedded approaches utilize machine learning methods to evaluate the selected feature subset.

Considering the purpose of being easily used and computational cost, this study uses decision trees. Decision tree is one of most popular supervised machine learning methods [17]. It belongs to embedded feature selection approach. When using decision tree method to be feature selection tool, the nodes in the constructed tree will be considered as impor-tant. Other features will be removed. DT is also widely applied to travel areas. For exam-ples, Tyrvainen et al. [30] studied the traveler's preferences and select tourist destination. They used DT to find the factors which influence travelers' selection to develop overall planning of resorts, to enhance eco-efficiency and sustainable development. Kim & Upne-ja [19] used decision tree (C4.5) in the restaurant which is in financial difficulties to accu-rately predict financial distress. Duchessi & Lauria [10] also utilized the decision tree model (CART) to predict the internet marketing performance of ski resorts to make adver-tising strategies, to enhance brand awareness, and to attract customers. These studies have shown that the decision tree method can be successfully used for feature selection.

3   Approach

This section describes the research methods and procedure. The study flow chart shown in Figure 1 can be divided into six main steps. The implementation produce will be illustrated by following instructions.

Step 1: Define the Factors of Social Media Marketing

From the available literatures, we try to identify and define the potential social media marketing elements. Then, we modified their original definitions for B & B industry. They constitute factors investigated in this work.

Step 2: Design and Pretest the Questionnaire

This study used a questionnaire as a measurement tool. Based on defined factors defined in step 1, we design the questionnaire and this questionnaire will be employed to survey respondents' views regarding the importance of every factors. Before issuing a formal questionnaire, in order to avoid being misunderstood the meaning of the respondents to affect the validity of the questionnaire, a pilot questionnaire will be pretested. We use one by one to have interview with respondents who should have the experiences of living in B&B and participating in social media.

```
┌─────────────────────────────────────┐
│ 1. Define the factors of social media mar- │
│         keting factors               │
└─────────────────────────────────────┘
                  ⇩
┌─────────────────────────────────────┐
│ 2. Design and pretest the questionnaire │
└─────────────────────────────────────┘
                  ⇩
┌─────────────────────────────────────┐
│ 3. Collect data                      │
└─────────────────────────────────────┘
                  ⇩
┌─────────────────────────────────────┐
│ 4. Implement decision tree           │
└─────────────────────────────────────┘
                  ⇩
┌─────────────────────────────────────┐
│ 5. Draw conclusions                  │
└─────────────────────────────────────┘
```
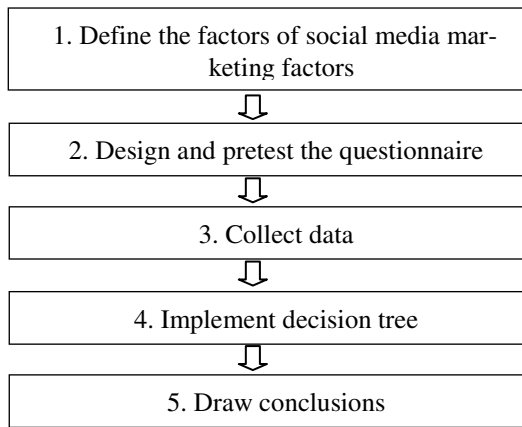
Fig. 1. The implemental procedure of employed approaches

Step 3: Collect Data

The completed questionnaire will be distributed to target respondents who have lived in B&B and the B&B enterprisers. In order to select valid samples, the internet and hardcopy questionnaires sent simultaneously to the respondents who lived in different areas such as northern, central and eastern Taiwan, and different age groups.

Step 4: Implement Feature Selection

In this step, we attempt to select important factors from the collected data. We use the data of the importance items (defined social media marketing factors) as our input data (attributes) and the labels "the probability of revisiting if B&B fulfills the important factors which respondents answered in questionnaire" as our output data. Then, we can construct decision trees for selecting the important factors. The detailed implemental process could be found as following sub-steps.

Step 5.1 Data preparation

Use 10-fold cross validation experiment and construct a decision tree for each fold of data. In other words, the collected data set will be divided into 10 equal sized sets and each set is then in turn used as the test set. Beside test set, we use other 9 sets as our training set to build decision trees. Therefore, we will have 10 trees.

Step 5.2 Determine the input and output variables
Step 5.3 Construct decision trees following C5.0 algorithm for each fold data set
Step 5.3.1 Create an initial rule tree
Step 5.3.2 Prune this tree
Step 5.3.3 Process the pruned tree to improve its understandability
Step 5.4 Pick a tree whose performance is the best among all constructed trees

Step 5: Draw Conclusions

Finally, from the results of step 4, we find the important factors of social media marketing in B&B industry. And then we can draw conclusions based on them.

4 Results

Table 1 shows the 16 defined social media marketing factors. Totally there are 1200 questionnaires have been issued. 210 examples are returned for further analysis. Among these collected samples, 17 examples are from B&B enterprises and others (193 examples) responded by customers. In all samples, males accounted for 45%, female 55%; samples were mainly aged 18 to 30 years old (51%), and then 31~40 years old (21%) and above 41 years old (25%); the majority (50%) samples' monthly income is from 20,000 to 50,000 NTD; the source of B & B information comes from shared information from social media (49%); the main reason (38%) of living in B&B is "unique characteristics"; 54% respondents spent less than 3 hours to browse social networking sites. We separate the customers and B&B enterprises to show their background in table 2. From this table, we can find some major differences between customers and B&B enterprisers. For examples, B&B enterprises think "unique characteristics" is only one motivation for attracting customers. But, except this motivation, customers think "pricing" is another crucial motivation. Besides, compared to customers, B&B enterprises obviously spent too little time to interact to members of social networking websites.

Table 1. The defined social media marketing factors in B&B industry

No.	Factors	Supports	No.	Factors	Supports
Q1	Fan page	[13]	Q9	Aesthetics and visual quality	[9, 39]
Q2	Group	[5, 13]	Q10	Interaction quality	[34]
Q3	Event	[13]	Q11	Functional needs	[2, 32]
Q4	Advertise-ment	[13, 33]	Q12	Psychological and hedonic	[32]
Q5	Market-place	[13]	Q13	Altruism	[11, 26]
Q6	Beacons and Polls	[4, 9]	Q14	Socialization and learning	[23]
Q7	Applica-tions	[28]	Q15	Relaxation	[23, 38]
Q8	Selection of features	[1, 21]	Q16	Novelty	[31, 38]

Table 2. The background of collected data

Variable		Customers	B&B enterprises
Gender	Male	47%	24%
	Female	53%	76%
Age	<18	2%	6%
	18~30	53%	24%
	31~40	20%	35%
	>40	24%	35%
Income	<20,000 NTD	34%	24%
	20,000~50,000 NTD	50%	53%
	50,000~100,000 NTD	14%	24%
	>100,000NTD	3%	0%
Source of B & B information	Keyword advertisements	9%	18%
	Newspapers & magazines	34%	24%
	Social media	48%	53%
	Others	9%	6%
Motivation for living B&B	Pricing	35%	18%
	Unique characteristics	36%	65%
	Word of mouth	16%	12%
	Natural view and landscapes	10%	6%
	Others	3%	0%
Spent time in social networking websites	<3hrs	51%	88%
	3~6hrs	39%	12%
	6~9hrs	7%	0%
	>9hrs	3%	0%

Table 3. Performances of 10-fold cross validation experiment

Fold	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10
Errors	33.3%	28.6%	52.4%	38.1%	42.9%	52.4%	33.3%	38.1%	33.3%	28.6%
Rule number	14	7	14	14	11	14	16	5	8	16

Table 4. The extracted knowledge rules for the tree with the best performance (fold #2)

No.	Rules
1	IF Q8 <= 1 AND Q9 > 2 AND Q15 > 2 THEN Class=HIGH [0.750]
2	IF Q4 <= 1 AND Q8 <= 1 AND Q13 > 2 THEN Class=HIGH [0.750]
3	IF Q8 <= 1 AND Q13 <= 2 THEN Class=HIGH [0.508]
4	IF Q6 > 1 AND Q6 <= 2 AND Q8 > 2 THEN Class=MEDIUM [0.857]
5	IF Q8 > 1 AND Q8 <= 2 THEN Class= MEDIUM [0.747]
6	IF Q15 > 2 THEN Class= MEDIUM [0.710]
7	IF Q6 > 2 AND Q8 > 2 AND Q10 <= 2 AND Q14 > 1 THEN Class=LOW [0.889]

Note:
(1) The value of attributes (Q1~Q16) "1,2,3,4,5" means "very important, important, neutral, unimportant, not very important".
(2) Class = "the probability of revisiting if B&B fulfills the important factors which respondents answered in questionnaire"

Next, we implement 10 fold cross validation experiments. Therefore, there are 10 trees have been built. And they are listed in table 3. From this table, we can find fold #2 that has the best performance (the lowest error rate). Consequently, this tree has been picked to select features. Table 4 summarizes all extracted knowledge rules by this decision tree. There are seven rules and the attributes left in the nodes will be considered as important. We also divided the all samples into "customers" and "B&B enterprisers" to find the differences. In table 5, we can find 8 important factors from 16 candidate attributes. They are "Q4(Advertisement)", "Q6(Beacons and Polls)", "Q8(Selection of features)", "Q9(Aesthetics and visual quality)", "Q10(Interaction quality)", "Q13(Altruism)", "Q14(Socialization and learning)", and "Q15(Relaxation)".

Table 5. Summary of selected important factors of social media marketing

	All data	Customers	B&B enterprisers
Selected impor-tant factors	Q4, Q6, Q8, Q9, Q10 Q13, Q14, Q15	Q1, Q2, Q3, Q6, Q8 Q9, Q10, Q12, Q14 Q15	Q15
Best performance (Over accuracy)	71.4%	73.7%	100%

5   Conclusions

This study aims to define potential social media marketing factors for B & B industry. From the defined 16 candidate factors, we use feature selection approaches to find the crucial factors. Customers think "Q4(Advertisement)", "Q6(Beacons and Polls)", "Q8(Selection of features)", "Q9(Aesthetics and visual quality)", "Q10(Interaction quali-ty)", "Q13(Altruism)", "Q14(Socialization and learning)", and "Q15(Relaxation)" are the key social media marketing factor for increasing customer royalty.

In the future research directions, readers can find aother new feature selection to analyze sample data, and then compare the difference between consumers and B&B enterprisers. To discover the true consumers' thinking about social media marketing can bring financial benefits. In addition, the investigation can be conducted for differ-ent industry types and other social media.

References

1. Albee, A.: eMarketing Strategies for the Complex Sale. McGraw-Hill (2010)
2. Andersen, P.H.: Relationship marketing and brand involvement of professionals through web-enhanced brand communities: The case of Coloplast. Industrial Marketing Manage-ment 34(1), 39–51 (2005)
3. Boley, B.B., Magnini, V.P., Tuten, T.L.: Social media picture posting and souvenir pur-chasing behavior: Some initial findings. Tourism Management 37, 27–30 (2013)

4. Boyd, D., Ellison, N.: Social network sites: definition, history, and scholarship. Journal of Computer-Mediated Communication 13(1), 210–230 (2007)
5. Bressler, S.E., Grantham, C.: Communities of commerce: Building internet business communities to accelerate growth, minimize risk, and increase customer loyalty. McGraw-Hill, Inc., New York (2000)
6. Chen, Y.C., Shang, R.A., Li, M.J.: The effects of perceived relevance of travel blogs' content on the behavioral intention to visit a tourist destination. Computers in Human Behavior 30, 787–799 (2014)
7. Cantallops, A.S., Salvi, F.: New consumer behavior: A review of research on eWOM and hotels. International Journal of Hospitality Management 36, 41–51 (2014)
8. Chung, J.Y., Buhalis, D.: A study of online travel community and Web2.0: Factors affecting participation and attitude. In: Proceedings of ENTER 2008, Innsbruck, pp. 267–278. Springer, Wien (2008)
9. Cross, R.L., Parker, A.: The hidden power of social networks: Understanding how work really gets done in organizations, p. 304. Harvard Business Review Press (2004)
10. Duchessi, P., Lauria, E.J.M.: Decision tree models for profiling ski resorts' promotional and advertising strategies and the impact on sales. Expert Systems with Applications 40(15), 5822–5829 (2013)
11. Eduardo, P.L., Jacques, B.G., Desiderio, G.T., Ricardo, D.A.: Intentions to use social media in organizing and taking vacation trips. Computers in Human Behavior 27(2), 640–654 (2011)
12. Hu, N., Bose, I., Koh, N.S., Liu, L.: Manipulation of online reviews: An analysis of ratings, readability, and sentiments. Decision Support Systems 52(3), 674–684 (2012)
13. Holzner, S.: Facebook marketing: Leverage social media to grow your business, p. 288. QUE (2009)
14. Hughes, D.J., Rowe, M., Batey, M., Lee, A.: A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. Computers in Human Behavior 28(2), 561–569 (2012)
15. Jacques, B.G., Santiago, M.G., Beatriz, G.L.V.: A social media analysis of the contribution of destinations to client satisfaction with hotels. International Journal of Hospitality Management 35, 44–47 (2013)
16. Kerstetter, D., Cho, M.H.: Prior knowledge, credibility and information search. Annals of Tourism Research 31(4), 961–985 (2004)
17. Kirkos, E., Spathis, C., Manolopoulos, Y.: Audit-Firm group appointment: An artificial intelligence approach. Intelligent Systems in Accounting, Finance and Management, 1–17 (2010)
18. Koetsier, J.: Nielsen's state of social 2012 report: more social, more mobile, more minutes, more TV, and more ads, Nielsens' 2012 state of the social union report (2012)
19. Kim, S.Y., Upneja, A.: Predicting restaurant financial distress using decision tree and AdaBoosted decision tree models. Economic Modelling 36, 354–362 (2014)
20. Laroche, M., Habibi, M.R., Richard, M.O.: To be or not to be in social media: How brand loyalty is affected by social media? International Journal of Information Management 33(1), 76–82 (2013)
21. O'Connor, P., Hopken, W., Gretzel, U.: Information and communication technologies in Tourism 2008. In: Proceedings of the International Conference in Innsbruck, p. 591 (2008)
22. Pan, B., MacLaurin, T., Crotts, J.C.: Travel blogs and the implications for destination marketing. Journal of Travel Research 46(1), 35–45 (2007)
23. Park, D.B., Yoon, Y.S.: Segmentation by motivation in rural tourism: A Korean case study. Tourism Management 30(1), 99–108 (2009)

24. Pallis, G., Zeinalipour-Yazti, D., Dikaiakos, M.D.: Online social networks: Status and trends. In: Vakali, A., Jain, L.C. (eds.) New Directions in Web Data Management 1. SCI, vol. 331, pp. 213–234. Springer, Heidelberg (2011)
25. Stelzner, M.A.: How marketers are using social media to grow their businesses, Social Media Marketing Industry Report, pp. 2–43 (2013)
26. Sigala, M., Marinidis, D.: Exploring the transformation of tourism firms' operations and business models through the use of web map services. In: European and Mediterranean Conference on Information Systems 2009 (EMCIS 2009), Brunel University, Izmir, pp. 1–13 (2009)
27. Safko, L., Brake, D.K.: The social media bible: Tactics, Tools, and Strategies for business success, p. 640. Wiley (2009)
28. Treadaway, C., Smith, M.: Facebook Marketing: An Hour a Day, Sybex (2010)
29. Tomas, E.R., Elena, C.T.: An evaluation of Spanish hotel websites: Informational vs. relational strategies. International Journal of Hospitality Management 33, 228–239 (2013)
30. Tyrvainen, L., Uusitalo, M., Silvennoinen, H., Hasu, E.: Towards sustainable growth in nature-based tourism destinations: Clients' views of land use options in Finnish Lapland. Landscape and Urban Planning 122, 1–15 (2014)
31. Terri, L.R., Anny, R.: Bed and breakfasts, small inns, and the internet: The impact of technology on the globalization of small businesses. Journal of International Marketing 8(2), 86–97 (2000)
32. Tan, W.K., Chang, Y.C.: Credibility assessment model of travel information sources: An exploratory study on travel blogs. In: Proceedings of ENTER, pp. 457–469. Springer, Vienna (2011)
33. Wang, Y., Yu, Q., Fesenmaier, D.R.: Defining the virtual tourist community: implications for tourism marketing. Tourism Management 23(4), 407–417 (2002)
34. Wilson, R.F.: The six simple principles of viral marketing. E-commerce consultant (2005)
35. Xiang, Z., Gretzel, U.: Role of social media in online travel information search. Tourism Management 31(2), 179–188 (2010)
36. You, W., Yang, Z., Ji, G.: Feature selection for high-dimensional multi-category data using PLS-based local recursive feature elimination. Expert Systems with Applications 41(4), 1463–1475 (2014)
37. Yu, Y., Duan, W., Cao, Q.: The impact of social and conventional media on firm equity value: A sentiment analysis approach. Decision Support Systems 55(4), 919–926 (2013)
38. Yuksel, A., Akgul, O.: Postcards as affective image makers: An idle agent in destination marketing. Tourism Management 28(3), 714–725 (2007)
39. Zarella, D.: The social media marketing book, p. 232. O'Reilly Media (2010)
40. Zhou, L., Wang, T.: Social media: A new vehicle for city marketing in China. Cities 37, 27–32 (2014)

A Novel Approach for Sustainable Supplier Selection Using Differential Evolution: A Case on Pulp and Paper Industry

Sunil Kumar Jauhar[1], Millie Pant[1], and Ajith Abraham[2]

[1] Indian Institute of Technology, Roorkee, India
[2] Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence, Auburn 98071, WA, USA
{suniljauhar.iitr,millidma}@gamil.com, ajith.abraham@ieee.org

Abstract. Diverse sustainable supplier selection (SSS) methodologies have been suggested by the practitioners in earlier, to find a solution to the SSS problem. A SSS problem fundamentally is a multi-criteria practice. It is a judgment of tactical significance to enterprises. The nature of this decision usually is difficult and unstructured. Optimization practices might be useful tools for these types of decision-making difficulties. During last few years, Differential Evolution has arisen as a dominating tool used for solving a variety of problems arising in numerous fields. In the current study, we present an approach to find a solution to the SSS problem using Differential Evolution in pulp and paper industry. Hence this paper presents a novel approach is to practice Differential Evolution to select the efficient sustainable suppliers providing the maximum fulfillment for the sustainable criteria determined. Finally, an illustrative example on pulp and paper industry validates the application of the present approach.

Keywords: Sustainable Supplier Selection, Sustainable Supply Chain Management, Differential Evolution, DEA.

1 Introduction

Now days in view of the growing awareness concerning sustainability in business firm, the sustainable supplier selection would be the vital element in the process of managing a sustainable supply chain. Developing an efficient and robust sustainable supply chain (also known as sustainable SCM) is a crucial task for the success of a business firm. One of the most significant factors that help in building a strong sustainable supply chain is the SSS process. It is natural that for a particular product, huge amounts of suppliers/ vendors are available in the market. Now, it is the duty of the purchasing managers to recognize the most suitable clusters of sustainable suppliers for their product. Evaluation plus selection of sustainable suppliers is a complex process and depend on a large number of qualitative as well as quantitative factors to ensure a cost effective model without compromising with the quality.

Differential evolution (DE) algorithm proposed by Stron and Price in 1995 [1], It is a population set based evolutionary algorithm that has been applied successfully to a wide range of problems [2,3,4,5,6]. The motive of this paper is to bring about to SSS approach. In the current research article, DE is used for solving the SSS problem modeled with the help of data envelopment analysis (DEA).

This paper is organized in eight sections. Subsequent to the introduction in Section 1, the SCM in pulp and paper industry, sustainable supplier selection (SSS) in addition to Problem statement and methodology are briefed in Sections 2, 3 and 4 respectively. Section 5 describes the Mathematical Formulation of the Problem with DEA used in this article. Section 6 describes the DE algorithm for SSS. Finally, a discussion and conclusion with summary drawn from the current research article are given in Section 7 and 8.

2   SCM in Pulp and Paper Industry

The pulp and paper industry produces a great amount of paper as well as cellulose based fibre products. Bulletin papers, copy papers, different kinds of tissue, bottle sticky label, cigarette papers plus coffee filter are just a small number of patterns of the products frequently used in our daily life. There is a huge amount of activity involved in the chain behind these products; such a system of actions is acknowledged as supply chain in management as well as operation research works [7].   The manufacturing of paper can be separated into six main process areas, which are [8]: (1) Wood preparation (2) Pulping (3) Bleaching (4) Chemical recovery (5) Pulp drying (non-integrated mills only) and (6) Papermaking. Fig. 1 presents a flow diagram of the paper manufacturing process in pulp and paper industry.
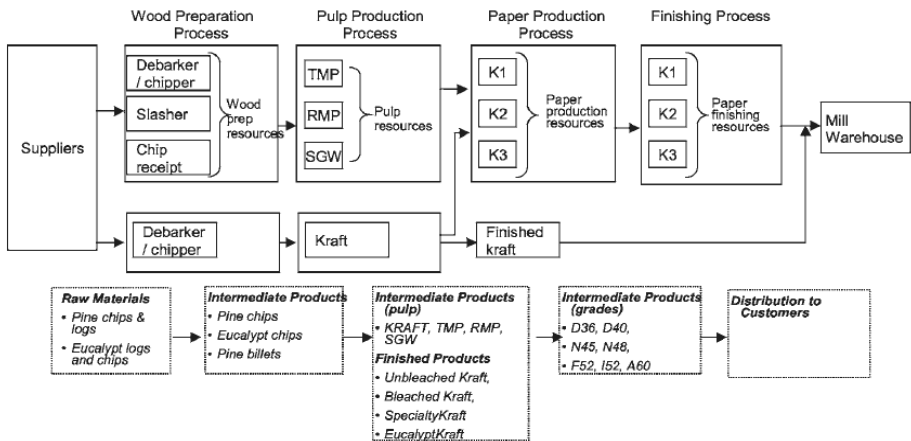
Fig. 1. Paper manufacturing process in pulp and paper industry (Source: [9])

Business organizations are gradually identifying that the effective management of sustainable supply chains is a primary driver of value creation as well as environmental performance. While lots of organizations in the automotive, consumer goods, and electronics industries have exploited the environment concern of SCM, industries in

the pulp and paper area are just start to identify the immense scope of the potential future prospects that exist. These industries are categorized by a huge and extremely incorporated supply chain. The entire supply chain starts in the procurement network, carry on through production network, distribution network and finishes by sales network. The fig. 2 presents an easy illustration of the pulp and paper industry supply chain organization with corresponding main SCM level.
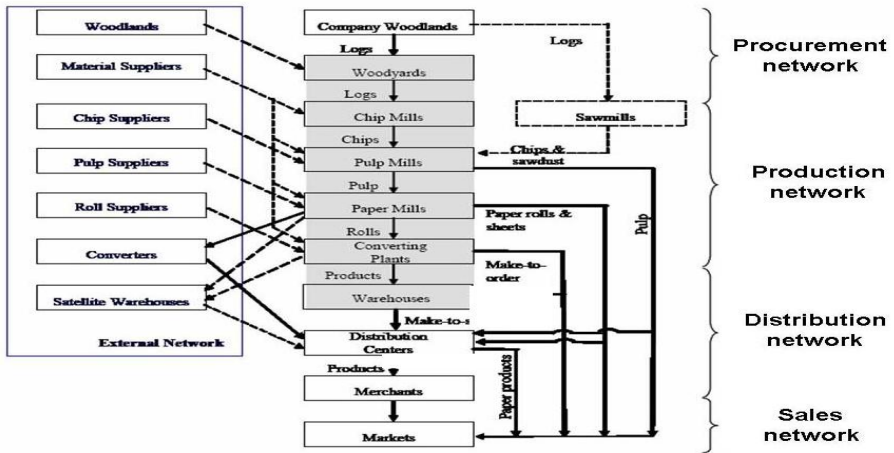
Fig. 2. The pulp and paper supply chain organization. (Source: [10])

2.1 Pulp and Paper CO_2 Emission Sources

The paper products manufacturing companies has a significant and complicated role in the worldwide carbon cycle. Pulp and paper are massive consumers of energy. In fact, the world's fifth- leading consumers of energy is the pulp and paper industry. The World Resources Institute, a body of experts, placed the industry's CO_2 emissions at around 500m tons worldwide in 2005. Rather, European companies are reasonably green. The Confederation of European Paper Industries (CEPI), a trade association, states his associates' emissions were 46m tons in 2011 [9].

Greenhouse gas (GHG) emissions from the pulp and paper source group are most of the part of CO_2 with lesser quantity of CH_4 as well as N_2O. The GHG emissions allied with the pulp and paper industry's processes can be caused by [10]: (1) The process of burning on-site energies; (2) Non-energy-related emission causes, for example by-product CO_2 emissions from the lime kiln chemical reactions and CH_4 emissions from wastewater treatment and (3) indirect emissions of GHG are related through the off-site production of steam in addition to electricity that are procured by or pass on to the industry.

Environmental concern over the pulp and paper industry has until recently been concentrated on emissions to air and water from pulping, bleaching, and paper making practices. A blend of strategy tools (usually emission ceilings) has reduced emissions to air as well as water from these industries significantly [11]. In this days and

age, concern is usually for the sectors effect on external network of pulp and paper supply chain organization (see Fig. 2) and its emissions of CO_2.

3   Sustainable Supplier Selection (SSS)

Supplier selection is a significant portion of supply chain management (SCM). Supplier selection comprises numerous criteria including price, lead time, quality, speed, delivery performance, reliability, etc and often encompasses the choice of one meanwhile give up the other. The supplier selection practices are comprehensively studied in the literature with multi-criteria decision making models. These models comprise such practices, as the AHP, ANP, CBR, DEA, fuzzy set theory, GA, mathematical programming, SMART, and their hybrid variants. Different researchers have studied the works in the past relating the supplier evaluation and selection problem [12,13,14,15,16,17,18,19,20,21,22,23,24].

At the present time, sustainable development has turn out to be a buzzword that acknowledged a lot of attentions in numerous fields like manufacturing [25], business development [26], tourism [27] and agriculture [28]. Also, in SCM both academics and general practitioner contemplate the sustainable concerns in their workings. Sustainable SCM is the managing of resources, informations and capital flows, as well as collaboration between firms alongside the supply chain, meanwhile taking into account the objectives from all three dimensions, such as financial, ecological and societal, of sustainable growth derived from client and investor wants [29]. To select the potential suppliers, two focuses comprising significance one is the degree of the selection criteria, and second one is the suppliers' sustainable performance, these two focuses need to be verified with the appropriate decision makers. Therefore, to manage this problem and cope with the imprecision that is be found in the SSS problem, use of Differential evolution (DE) algorithm is explored in this article.

4   Problem Statement and Methodology

The case study presented in this paper stands a hypothetical paper manufacturer in India (X Company). After verifying a group of criteria in a view point of sustainable merits, some criteria including lead time, quality, price, service quality and CO_2 emissions of the delivered products are derived for SSS problem.

In current study we split the criteria in two manners: the input and output criteria. The input criteria are the traditional supplier selection criteria, such as lead time, price and quality of the delivered goods. The output criteria are the service quality and CO_2 emission of the product and services. We assume that the service quality and CO_2 emission are the output of the examined model. The data is shown in Table I with the supplier's database covering input as well as output criteria of an item provided in the shipment of pulp and paper industry.

In this study, a supplier is considered efficient if the efficiency score is 1 otherwise it is considered as inefficient. Data for Service quality of items is taken from concept of Service quality dimension based on 12 questionnaires including 27 questions given

by [30] and for CO_2 Emissions, *LOCOG* Guidelines on Carbon Emissions of Products and Services –Version 1 [31] is considered.

4.1   Methodology

To measure and analyze the relative efficiency of 18 suppliers, we follow a four step methodology:

- ➢ Design a criteria containing input and output criteria
- ➢ Select a problem
- ➢ Formulate the mathematical model of the SSS problem with the help of DEA.
- ➢ Apply DE on mathematical model.

The present model can be carry out for any quantity of suppliers and there is no limitation, by using this model, the company can obtain a recommended combination of efficient suppliers.

Table 1. Data for numerical example

Criteria	Inputs			Output	
	Lead time (L) (Day)	Quality(Q) %	Price (P) (Rs.)	Service quality (SQ)	CO_2 Emission (CE)
Suppliers					
1	3	75	187	84	40
2	2	77	195	76	30
3	4	85	272	27	25
4	5	86	236	110	22
5	3	74	287	94	38
6	6	62	242	102	10
7	3	73	168	82	24
8	4	92	396	63	38
9	2	77	144	55	26
10	3	69	137	61	18
11	6	54	142	122	24
12	5	57	196	75	30
13	1	77	247	80	55
14	3	61	148	121	39
15	4	69	294	125	8
16	6	94	249	76	6
17	2	88	121	114	55
18	1	78	269	65	48

5   Mathematical Formulation of the Problem with DEA

DEA based method is used for determining the efficiencies of Decision-Making Units (DMUs) on the basis of multiple inputs and outputs [32]. DMU can comprise of business firms, divisions of huge groups such as institution of higher education, schools, hospitals, power plants, police stations, tax offices, prisons, a set

of organizations etc. [33,34,35,36]. The DMU well-describe in this research work using input as well as output criteria are as follows:
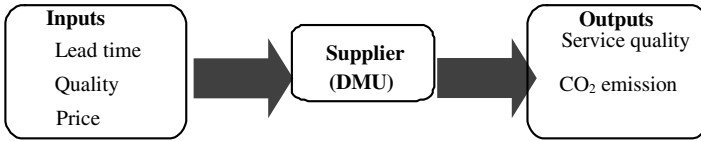
The performance of DMU is estimated in DEA by the concept of efficiency or productivity, which the proportion of weights sum of outputs (o/p) to the weights sum of (i/p) inputs [37] i.e

$$\text{Efficiency} = \frac{\text{Weighted sum of O/P}}{\text{Weighted sum of I/P}} \tag{1}$$

The two basic DEA models are the CCR (Charnes, Cooper and Rhodes) model [38] and the BCC (Banker, Charnes and Cooper) model [39], these two models distinguish on the returns to scale assumed. The former assumes constant returns-to-scale whereas the latter assumes variable returns-to-scale [32]. In the current study we use CCR model which is well-defined further down:

Suppose that there are N DMUs and each unit have I input and O outputs then the efficiency of mth unit is achieved by resolving the below model which is presented by Charnes et al [38].

$$\text{Max } E_m = \frac{\sum_{k=1}^{O} w_k Output_{k,m}}{\sum_{l=1}^{I} z_l Input_{l,m}} \tag{2}$$

$$O \le \frac{\sum_{k=1}^{O} w_k Output_{k,n}}{\sum_{l=1}^{I} z_l Input_{l,n}} \le 1; n = 1,2..., m..N$$

$$w_k, z_l \ge O; \forall k, l$$

Where

E_m is the efficiency of the mth DMU, k=1 to O, l =1 to I and n =1 to N.

$Output_{k,m}$ is the kth output of the mth DMU and w_k is weight of output $Output_{k,m}$

$Input_{l,m}$ is the lth input of mth DMU and z_l is the weight of $Input_{l,m}$

$output_{k,n}$ and $input_{l,n}$ are the kth output and l^{th} input respectively of the nth DMU, Where n=1, 2 …m...N

The fractional program shown in Equ-2 can be converted in a linear program which is shown in Equ-3.

$$\text{Max } E_m \sum_{k=1}^{O} w_k Output_{k,m}$$

$$s.t.$$

$$\sum_{l=1}^{I} z_l Input_{l,m} = 1.$$

$$\sum_{k=1}^{O} w_k Output_{k,n} - \sum_{l=1}^{I} z_l Input_{l,n} \le 0, \quad \forall n \tag{3}$$

$$w_k, z_l \ge 0; \quad \forall k, l$$

To calculate the efficiency score for each DMU we run the above program run N times. A DMU is considered efficient if the efficiency score is 1 otherwise it is considered as inefficient.

5.1   Mathematical Model

Based on the hypothetical statistics stated in Table I the DEA model of Kth DMU will be as follows:

$$
\begin{aligned}
& Max \ \ SQ_m + CE_n \\
& s.t. \\
& z_1 L_m + z_2 Q_m + z_3 P_m = 1 \\
& w_1 SQ_n + w_2 CE_n - (z_1 L_m + z_2 Q_m + z_3 P_m) \leq 0 \\
& \forall \ n = 1,...m,...18
\end{aligned} \tag{4}
$$

6   DE Algorithm

Differential evolution (DE) algorithm proposed by Stron and Price in 1995 [1], It is a type of evolutionary algorithm, used to most effective use of (optimize) the functions. It is a population set based evolutionary algorithm for global optimization. In current study we have used DE/rand/1/bin scheme and DE algorithm from reference [1].

6.1   Pseudo Code for the DE Algorithm

Table 2. The Pseudo code for the DE algorithm

1	Begin
2	Generate uniformly distribution random population $P=\{X_{1,G}, X_{2,G},..., X_{NP,G}\}$. $X_{i,G} = X_{lower} + (X_{upper} - X_{lower})*rand(0,1)$, where $i =1, 2,..,NP$
3	Evaluate $f(X_{i,G})$
4	While (Termination criteria is met)
5	{
6	For $i=1:NP$
7	{
8	Select three random vector $X_{r1,G}, X_{r2,G}, X_{r3,G}$ where $i \neq r_1 \neq r_2 \neq r_3$
9	Perform mutation operation
10	Perform crossover operation
11	Evaluate $f(U_{i,G})$
12	Select fittest vector from $X_{i,G}$ and $U_{i,G}$ to the population of next generation
13	}
14	Generate new population $Q= \{X_{1,G+1}, X_{2,G+1},..., X_{NP,G+1}\}$
15	} /* end while loop*/
16	END

6.2   Constraints Handling

For the constraint problems various methods have been suggested in literature. A survey of different methods for constraint handling can be found in [40] and [41]. In this paper Pareto-Ranking method is used for handling the constraints [42].

6.3   Parameter Setting for the DE Algorithm

In present research article we have applied DE to solve the DEA based mathematical model. The parameter settings for DE are given in Table-3.

Table 3. Parameter setting for DE

Pop size (NP)	100
Scale Factor (F)	0.5
Crossover rate (Cr)	0.9
Max iteration	3000

The program is implemented is DEV C++ and all the uniform random number is generated using the inbuilt function *rand ( )* in DEV C++. The fitness value is taken as the average fitness value in 30 runs and the program is terminate when reach to Max-Iteration. A buyer (decision maker) can effect an assessment (supplier evaluation) with the ability to choose of weight system. For this purpose with the help of program which is implemented is DEV C++, we intended to generate all the uniform random number (in between 0 to 1) using the inbuilt function *rand ( )* in DEV C++, to assist the selection of the weights for input as well as output criteria in a manner to permit the control of the result for the sustainable supplier evaluation and assessment practice.

7   Results and Discussions

In the Table-4 results of all DMUs are given. From this Table we can see that the research of efficient SSS practice can acquire a desirable cluster of competent sustainable suppliers 11, 14 and 17 using DE algorithm gives the better solution.

For the current research conducted in 18 suppliers, the results are:

1. Suppliers 11, 14 and 17, the efficiency score is 1 so these suppliers are assumed to be 100% sustainable efficient.
2. Supplier 3 is probably the most inefficient in comparison to all other suppliers.
3. Suppliers 11, 14 and 17 would be the most suitable set of suppliers (or key suppliers).
4. By using this DE algorithm, the business firms can acquire a desirable clusters of competent sustainable suppliers.

5. Combination of suppliers 11, 14 and 17 would be the desirable clusters of competent sustainable suppliers set meanwhile the business firms requiring single-item sustainable suppliers.

In the Table-5 results of all DMUs are given and fig. 3 shows the histogram of all suppliers with their efficiency score.

Table 4. Average efficiency and weights of 18 suppliers in 30 runs

Suppliers	Value of input and output weight					
	Z_1	Z_2	Z_3	W_1	W_2	**Efficiency**
1	0.5864615	0.0021643	0.0077285	0.0093171	0.00422648	0.885125
2	1	0	0	0.0096163	0.08283e-017	0.942402
3	0.0668079	0.0088852	0	0.0070293	0.33556e-01	0.0843525
4	0.0790911	0.010519	0	0.0083216	0.00248747	0.832165
5	0.107364	0.0142795	0	0.0112962	0.0219116	0.734253
6	0.0657469	0.0085730	0.0003544	0.0073778	0.00230969	0.811567
7	0.32575	0.0020776	0.0074189	0.0089439	0.58488e-01	0.822839
8	0.0683372	0.0090886	0.89E-20	0.0071902	0.0327347	0.524889
9	1	0	0	0.0096163	0.0498655	0.721226
10	0.0702566	0.0091611	0.0003788	0.0078839	0.0305688	0.638597
11	0.0321097	0.0136241	0.0008513	0.0089303	0.0031707	**1**
12	0.0984692	0.0092357	0.0015945	0.0102223	0.29418e-018	0.8689
13	0.0863947	0.0114904	0	0.0090900	0.0363442	0.636303
14	0.0208733	0.0068106	0.0043416	0.0090108	0.0315172	**1**
15	0.0747521	0.0099419	0	0.0078651	0.14697e-017	0.983144
16	0.0632637	0.0082492	0.0003411	0.0070992	0.0280174	0.468551
17	1	0	0	0.0096163	0.0625062	**1**
18	0.0774918	0.0103063	0.66829	0.0081534	0.01723e-014	0.448437

Table 5. Suppliers' efficiency

Suppliers	**Efficiency**	**Suppliers**	**Efficiency**
1	0.885125	**10**	0.638597
2	0.942402	**11**	**1**
3	0.0843525	**12**	0.8689
4	0.832165	**13**	0.636303
5	0.734253	**14**	**1**
6	0.811567	**15**	0.983144
7	0.822839	**16**	0.468551
8	0.524889	**17**	**1**
9	0.721226	**18**	0.448437
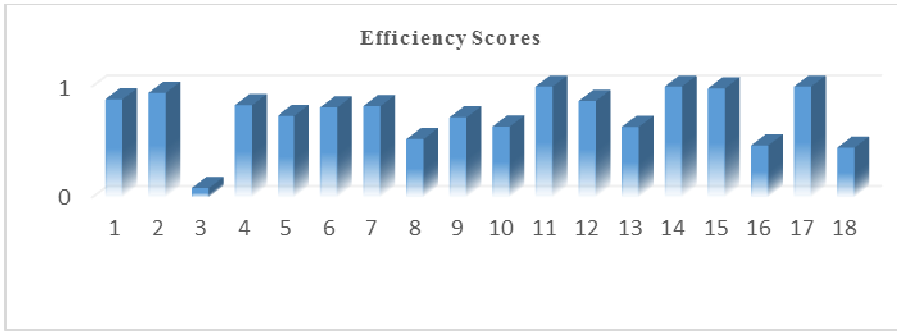
Fig. 3. Histogram of all suppliers with their efficiency score

8   Conclusion and Summary

SSS is a challenging task among thousands of potential suppliers. The present study shows DE algorithm as a tool for selecting the optimal sustainable suppliers. In this research work, we present a novel SSS approach for pulp and paper industry. The first step is to build a criterion set that comprising both Input as well as Output factors, which is appropriate for real world applications. We then present an approach to solve the multiple-criteria SSS problem with the application of DE, for DEA based mathematical model. By using this tactics, the business firms can acquire a desirable cluster of competent sustainable suppliers. In this study, the goal was the application of DE algorithm to the efficient SSS in the SCM.

The main motivation of this study was to gain an understanding of the mechanics of DE algorithm and to determine the accuracy of DE in generating the optimum solutions for the DEA based mathematical model, which is the underlying optimization problem for the aforementioned purchasing system. To our best information, this is the first time that the DE algorithm is applied to the SSS. This research article presents a DEA model for SSS practice in SCM. The key offerings of this research are précised as below: (1) SSS in SCM: to date, there are a small number of researches seeing sustainable concern in the supplier selection practice. (2) The selection criteria on the basis of sustainable concern are collected by means of the literature after that these are put in to the mathematical model for the SSS practice. (3) The present model can be carrying out for any quantity of suppliers and criteria's in the great size business firms. (4) In spite of the fact that many attempts have been made for the supplier selection, taking into consideration sustainable concern for this problem remains a demanding task. (5) Future research may explore the practice of the DE algorithm to find a solution to more difficult problem such as multi objective SSS.

References

1. Storn, R., Price, K.: Differential evolution—a simple and efficient adaptive scheme for global optimization over continuous spaces, Berkeley, CA, Tech. Rep. TR-95-012 (1995)
2. Plagianakos, V., Tasoulis, D., Vrahatis, M.: A Review of Major Application Areas of Differential Evolution. In: Chakraborty, U.K. (ed.) Advances in Differential Evolution. SCI, vol. 143, pp. 197–238. Springer, Berlin (2008)

3. Wang, F., Jang, H.: Parameter estimation of a bio reaction model by hybrid differential evolution. In: Proceedings of IEEE Congress on Evolutionary Computation (CEC 2000), pp. 410–417 (2000)
4. Joshi, R., Sanderson, A.: Minimal representation multi sensor fusion using differential evolution. IEEE Trans. Syst. Man Cybern. Part A Syst. Hum. 29(1), 63–76 (1999)
5. Ilonen, J., Kamarainen, J., Lampine, J.: Differential evolution training algorithm for feed-forward neural networks. Neural Process. Lett. 17(1), 93–105 (2003)
6. Ali, M., Siarry, P., Pant, M.: An efficient differential evolution based algorithm for solving multi-objective optimization. European Journal of Operational Research (2011)
7. Supply Chain Management in Pulp and Paper Industry, `https://www.cirrelt.ca/DocumentsTravail/2006/DT-2006-AM-3.pdf` (accessed on February 02, 2014)
8. Available and Emerging Technologies For Reducing Greenhouse Gas Emissions From The Pulp and Paper Manufacturing Industry, `http://www.epa.gov/nsr/ghgdocs/pulpandpaper.pdf`
9. Philpott, A., Everett, G.: Supply chain optimisation in the paper industry. Annals of Operations Research 108(1-4), 225–237 (2001)
10. Martel, A., M'Barek, W., D'Amours, S.: International factors in the design of multinational supply chains: the case of Canadian pulp and paper companies. Document de travail DT-2005-AM-3, Centor, Université Laval. 10 (2005)
11. Pulp and Paper Technology | Article, `http://www.pulpandpaper-technology.com/articles/id/29` (accessed on February 02, 2014)
12. Agarwal, P., Sahai, M., Mishra, V., Bag, M., Singh, V.: A review of multi-criteria techniques for supplier evaluation and selection. International Journal of Industrial Engineering Computations 2 (2011), doi:10 5267/j ijiec 2011 06 004
13. Weber, C.A., Current, J.R., Benton, W.C.: Vendor selection criteria and methods. European Journal of Operational Research 50(1), 2–18 (1991)
14. Degraeve, Z., Labro, E., Roodhooft, F.: An evaluation of vendor selection models from a total cost of ownership perspective. European Journal of Operational Research 125(1), 34–58 (1991)
15. Boer, L., Labro, E., Morlacchi, P.: A review of methods supporting supplier selection. European Journal of Purchasing & Supply Management 7(2), 75–89 (2000)
16. Holt, G.D.: Which Contractor Selection Methodology? International Journal of Project Management 16(3), 153–164 (1998)
17. Aamer, A.M., Sawhney, R.: Review of Suppliers Selection from a Production Perspective. In: Proc. IIE Conference, pp. 2135–2140 (2004)
18. Ho, W., Xu, X., Dey, P.K.: Multi-criteria decision making approaches for supplier evaluation and selection: A literature review. European Journal of Operational Research 202, 16–24 (2010)
19. Tahriri, F., Osman, M.R., Ali, A., Yusuff, R.M.: A review of supplier selection methods in manufacturing industries. Suranaree Journal of Science and Technology 15(3), 201–208 (2008)
20. Cheraghi, S.H., Dadashzadeh, M., Subramanian, M.: Critical success factors for supplier selection: An update. Journal of Applied Business Research 20(2), 91–108 (2011)
21. Jauha, S.K., Pant, M.: Recent trends in supply chain management: A soft computing approach. In: Jagdish, C., Bansal, P., Singh, K., Deep, M., Pant, A. (eds.) Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012). AISC, vol. 202, pp. 465–478. Springer, India (2013)

22. Jauhar, S.K., Pant, M., Deep, A.: An Approach to Solve Multi-criteria Supplier Selection While Considering Environmental Aspects Using Differential Evolution. In: Panigrahi, B.K., Suganthan, P.N., Das, S., Dash, S.S. (eds.) SEMCCO 2013, Part I. LNCS, vol. 8297, pp. 199–208. Springer, Heidelberg (2013)

23. Jauhar, S.K., Pant, M.: Supplier selection in SCM: A decision making approach. In: Proceeding of OPTIMA 2012, University of Delhi, India, November 29-December 01 (2012)

24. Jauhar, S., Pant, M., Deep, A.: Differential Evolution for Supplier Selection Problem: A DEA Based Approach. In: Pant, M., Deep, K., Nagar, A., Bansal, J.C. (eds.) Third International Conference on Soft Computing for Problem Solving. AISC, vol. 258, pp. 343–353. Springer, India (2014)

25. Jayal, A.D., Badurdeen, F., Dillon Jr., O.W., Jawahir, I.S.: Sustainable manufacturing: Modeling and optimization challenges at the product, process and system levels. CIRP Journal of Manufacturing Science and Technology 2(3), 144–152 (2010)

26. Floridi, M., Pagni, S., Falorni, S., Luzzati, T.: An exercise in composite indicators construction: Assessing the sustainability of Italian regions. Ecological Economics 70(8), 1440–1447 (2011)

27. Luthe, T., Schuckert, M.: Socially Responsible Investing–Implications for Leveraging Sustainable Development. In: Trends and Issues in Global Tourism 2011, pp. 315–321. Springer, Heidelberg (2011)

28. Paoletti, M.G., Gomiero, T., Pimentel, D.: Introduction to the special issue: Towards a more sustainable agriculture. Critical Reviews in Plant Sciences 30(1-2), 2–5 (2011)

29. Büyüközkan, G., Çifçi, G.: A novel fuzzy multi-criteria decision framework for sustainable supplier selection with incomplete information. Computers in Industry 62(2), 164–174 (2011)

30. Shirouyehzad, H., Lotfi, F.H., Dabestani, R.: A data envelopment analysis approach based on the service quality concept for vendor selection. In: International Conference on Computers & Industrial Engineering, CIE 2009, July 6-9, pp. 426–430 (2009), doi:10.1109/ICCIE.2009.5223823

31. LOCOG Guidelines on Carbon Emissions of Products and Services, http://www.london2012.com/documents/ locog-publications/locog-guidelines-on-carbon-emissions- of-products-and-services.pdf

32. Dimitris, K.S., Lamprini, V.S., Yiannis, G.S.: Data envelopment analysis with nonlinear virtual inputs and outputs. European Journal of Operational Research 202, 604–613 (2009)

33. Ramanathan, R.: An Introduction to Data Envelopment Analysis: A Tool for Performance Measurement. Sage Publication Ltd., New Delhi (2003)

34. Wen, U.P., Chi, J.M.: Developing green supplier selection procedure: A DEA approach. In: 2010 IEEE 17th International Conference on Industrial Engineering and Engineering Management (IE&EM), October 29-31, pp. 70–74 (2010), doi:10.1109/ICIEEM.2010.5646615

35. Vörösmarty, G., Dobos, I.: Supplier selection and evaluation decision considering environmental aspects. 149. sz. Mőhelytanulmány, HU (October 2012) ISSN 1786-3031

36. Kumar, P., Mogha, S.K., Pant, M.: Differential Evolution for Data Envelopment Analysis. In: Deep, K., Nagar, A., Pant, M., Bansal, J.C. (eds.) Proceedings of the International Conference on SocProS 2011. AISC, vol. 130, pp. 311–319. Springer, Heidelberg (2012)

37. Srinivas, T.: Data envelopment analysis: models and extensions. Production/Operation Management Decision Line, 8–11 (2000)

38. Charnes, A., Cooper, W.W., Rhodes, E.: Measuring the efficiency of decision making units. European Journal of Operational Research 2(6), 429–444 (1978)

39. Banker, R.D., Charnes, A., Cooper, W.W.: Some models for estimating technical and scale inefficiencies in data envelopment analysis. Management Science 30, 1078–1092 (1984)
40. Jouni, L.: A constraint handling approach for differential evolution algorithm. In: Proceeding of IEEE Congress on Evolutionary Computation (CEC 2002), pp. 1468–1473 (2002)
41. Coello, C.A.C.: Theoretical and numerical constraint handling techniques used with evolutionary algorithms: a survey of the state of the art. Computer Methods in Applied Mechanics and Engineering 191(11-12), 1245–1287 (2002)
42. Ray, T., Kang, T., Chye, S.K.: An evolutionary algorithm for constraint optimization. In: Whitley, D., Goldberg, D., Cantu-Paz, E., Spector, L., Parmee, I., Beyer, H.G. (eds.) Proceeding of the Genetic and Evolutionary Computation Conference (GECCO 2000), pp. 771–777 (2000)

A New Clustering Algorithm Based on Probability

Zhang Yue[*] and Zhou Chuansheng

Software College, Shenyang Normal University, 110034 Shenyang, China
baofeng0722@126.com, jasoncs@126.com

Abstract. Clustering is a hot topic of data mining. After studying the existing classical algorithm of clustering, this paper proposes a new clustering algorithm based on probability, and makes a new definition for clustering and outlier. According to the distribution characteristics of sample data, this algorithm determines the initial clustering center automatically. It also implements eliminating outliers in the process of clustering. The experiment results on IRIS show that this algorithm can clustering effectively.

Keywords: Clustering, Outlier, DBSCAN Algorithm, Mathematical expectation, Standard deviation.

1 Introduction

As a data analysis method, clustering analysis is commonly used in data mining. It can group data into classes or clusters. There is high similarity between data objects in the same cluster. [1] In different cluster, the difference between data objects is greater. So, we are able to identify the dense and sparse area, the global distribution patterns of data, and the relationship among data. [2]

Clustering algorithms are divided into four types. (1) Partition-based algorithm, e.g. k-means[3], k-medoids. This kind of algorithm needs to set cluster number. According to the similarity of objects to determine each object belongs to the nearest cluster. (2) Hierarchical clustering algorithm, e.g. single-link, complete-link, BIRCH. It includes agglomerative method and divisive method. (3) Statistics-based algorithm, e.g. EM[4]. Assuming that the data set is generated by a statistical process and the data set is described by best model. (4) Density-based algorithm, e.g. DBSCAN[5], OPTICS. The key idea is finding areas of high density data sets are separated by low density regions.

Efficient clustering algorithm should make the characteristics in the same cluster as similar as possible and different cluster objects as dissimilar much as possible. When selecting the initial clustering center, the clustering algorithm proposed in this paper implements this characteristic as far as possible. It does not select *k* objects as the center of initial clustering randomly. According to the distribution characteristics of sample data, the clustering algorithm determines the initial clustering center automatically and avoids the influence of artificial factors on the clustering results. Besides, we make a

---

[*] Corresponding author.

new definition of outlier, and implement eliminating outliers in the process of clustering.

2   Relevant Concepts

In order to comprehend the algorithm proposed in this paper, correlative conceptions are introduced as follows.

Definition1: Mathematical expectation $E(x)$. Supposing that, the distribution of a given object X is $P\{X = x_k\} = p_k, k = 1,2, ...$ If $\sum_{k=1}^{\infty} x_k p_k$ is absolute convergence, it is called $E(x)$. Namely, $E(x) = \sum_{k=1}^{\infty} x_k p_k$, $k = 1,2, ...$ If k is bigger, $E(x) = \frac{\sum_{i=1}^{N} x_i}{N}$.

Mathematical expectation is also known as mean. Generally, it characterizes the average of random variable. It is also weighted average for the random one. Besides, the center of distribution for random variable is determined by it.

Definition 2: Variance $D(X)$. Supposing that, X is a random variable. If $E\{[X - E(X)2$ exists, it is the variance for X. Namely, $D(X) = EX - E(X)2$. When k is bigger, $D(X) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$.

Variance characterizes the range of sample fluctuation. It describes the deviation between a random variable and its mathematical expectation. The bigger is variance, the deviation is it.

Definition 3: Standard deviation $\sigma(X)$. Supposing that, X is a random variable. If $D(X)$ exists, we call $\sqrt{D(X)}$ is standard deviation. Namely, $\sigma(X) = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N-1}}$.

Definition 4: Chebyshev inequivalence.[6] Supposing that, for a given object X, $E(X) = \mu$, $D(X) = \sigma^2$. For any positive number ε, if $p\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$ is established, it is called Chebyshev inequivalence.

Definition 5: Outlier. According to Chevyshev inequality, we can see that for any small positive number $\varepsilon > 0$, $p\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$ always exists. Specially, $\varepsilon = 3\sigma$, $p\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2} = \frac{1}{9}$ exists. If the group data is normal, above probability is only 0.27%. So, any given object X which satisfies $|X - \mu| \geq 3\sigma$ is called outlier. $\mu = E(x)$, $\sigma = \sqrt{D(X)}$.

3   Algorithm Description

3.1   DBSCAN Algorithm

Density-based clustering algorithm is an important branch. This algorithm can overcome the shortcoming of distance-based algorithm which is able to find quasi-circular cluster. It can form clusters with any shape. However, the most

important aspect is that dense-based clustering algorithm is not sensitive to noisy data.

DBSCAN is a typical density-based clustering algorithm. It defines cluster as the maximum set of those points which densities are connected. It is able to find any shape cluster in spatial database with noisy data. In DBSCAN algorithm, for any object in the spatial, if the count of objects within given radius *Eps* is bigger than certain given value *MinPts*, it is called core object; otherwise we call this it as border object.

From above algorithm we can learn that the value of *Eps* and *MinPts* will affect the clustering directly. The value of *MinPts* is 4 in the two-dimensional spatial cluster with small data sample. [7]Setting 1/25 of data set as the value of *MinPts* is also an effective method.[8] Parameter value also affects the determination of outlier. Inappropriate value will confuse the object in a sparse cluster with outlier. With the decrease of parameter values, the possibility for detecting outlier will reduce. It is very important to choose the value of *Eps* and *MinPts*.

3.2   Determine Initial Clustering Center

It is so difficult to implement by randomly that k objects which belong to different cluster are selected before clustering. So, determining the center of initial clustering is the key of this proposed algorithm.

For DBSCAN algorithm, there is no problem of choosing initial clustering center. The key idea in DBSCAN is that for each object of a cluster, the neighborhood of a given radius has to contain at least a minimum number of objects. The procedure for finding a cluster is based on the fact that a cluster is uniquely determined by any of its core objects. Finally, according to the distribution characteristic of data, DBSCAN algorithm outputs initial clustering results $C(C_1, C_2, \ldots C_k)$, which can further determine the initial clustering center.

Mathematical expectation represents the center of random variable probability distribution. According to Chebyshev theorem, if the amount of sample is large enough, the arithmetic mean of the sample data is approximately equal to ensemble average. That is, $lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} \xi_i - \mu\right| < \varepsilon\right) = 1$. So, we use the mean of sample data to estimate ensemble average. The mean of each cluster $E(C_i) = \frac{1}{n}\sum_{j=1}^{n} c_j$ ($i = 1,2, \ldots, k$) in initial clustering results $C(C_1, C_2, \ldots C_k)$ will be as k centers of initial clustering for probability-based clustering algorithm. At present, there is no mature theory or conclusions for the number of samples on the prediction accuracy.[9]But the amount of extracted data must be greater than the credible number of sample in mathematical statistics. Usually, it is 45. [6]

3.3   Algorithm Principle

The clustering algorithm proposed in this paper is a method based on probability. Firstly, select effective sample K' from original data set K, and satisfy the count of $K' \geq 45$. It is clustered under DBSCAN algorithm. We can obtain k initial clustering results $C(C_1, C_2, \ldots C_k)$ and the center of k initial clustering results. Secondly, select object k_i ($i = 1,2, \ldots, n$) from original data set. If $|k_j - E(C_i)| \geq 3\sqrt{D(C_i)}$($j = 1,2, \ldots, n, i = 1,2, \ldots, k$) is established, object k_j ($j = 1,2, \ldots, n$) does not belong to

cluster $C_i (i = 1,2,...,k)$. Otherwise, object $k_j (j = 1,2,...,n)$ belongs to cluster $C_i (i = 1,2,...,k)$ which makes $\min(|k_j - E(C_i)|) (j = 1,2,...,n, i = 1,2,...,k)$ is established.

As DBSCAN algorithm is insensitive to noisy data, they can be removed in the process of initial clustering. For scattered data set, if the value of parameters *Eps* and *MinPts* are small, internal data would be by mistake for noisy data under DBSCAN algorithm. So, this kind of noisy data are called suspicious outliers. Suspicious outliers cannot simply be removed directly. Further analysis is needed.

From Chebyshev inequivalence we can see that, for any small positive number $\varepsilon > 0$, $p\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2}$ always exists. Specially, $\varepsilon = 3\sigma$, $p\{|X - \mu| \geq \varepsilon\} \leq \frac{\sigma^2}{\varepsilon^2} = \frac{1}{9}$ is established. If the group data is normal, $P\{|X - \mu| \geq 3\sigma\} = 1 - P\{\mu - 3\sigma < X \leq \mu + 3\sigma\} = 1 - (\Phi(3) - \Phi(-3)) = 1 - 0.9974 = 0.26\%$. So, for a random variable X, if $|X - \mu| \geq 3\sigma$, X is outlier and needed to be removed.

According to the definition of outlier, we can obtain the absolute value of offset between a given object and each cluster in existing clustering result set. This absolute value is $|k_i - E(C_i)|$. Toward all known clusters in clustering result set, if $|k_i - E(C_i)| \geq 3\sigma$, object k_i does not belong to any cluster, then building a new cluster C_{k+1}. Otherwise, object k_i belongs to cluster C_k which makes $\min(|k_1 - E(C_i)|) (i = 1,2,...,k)$ existed. When the count of object in cluster C_k equals to 1, no matter what next object k_{i+1} is, inequivalence $|k_i - E(C_i)| \geq 3\sigma$ always come into existence. So, toward this instance, we adopt threshold of deviation δ to make analysis. The choice of δ will affect the quality of cluster directly. With reducing the value of δ, the count of cluster will increase. Contrarily, it will reduce. If δ is too large, it will increase the probability which makes outliers with normal data in the same cluster. This instance will be bad for detecting outliers. If δ is too small, it will add the similarity within clusters and divide certain cluster into some small ones. It also reduces the quality of clustering.

3.4   Clustering Algorithm

Input: data set K, threshold of deviation δ, and initialize the object count of cluster $n_i = 0 (i = 1,2,...,k)$

Output: clustering result set $C(C_1, C_2, ... C_k)$

Algorithm description:

1) Select effective sample K' from original data set K, and satisfy the count of $K' \geq 45$;

2) Obtain k initial clustering results $C'(C_1', C_2', ..., C_k')$ under DBSCAN algorithm, and the center of each initial clustering $E(C_i') = \frac{1}{n}\sum_{j=1}^{n} c_j' (i = 1,2,...,k)$;

3) Randomly choose an object k_1 in original data set K, if $|k_1 - E(C_i)| \leq \delta (i = 1,2,...,k)$, put k_1 into cluster C_i which makes $\min(|k_1 - E(C_i)|) (i = 1,2,...,k)$, and $n_i = n_i + 1 (i = 1,2,...,k)$. Recalculate $E(C_i)$ and $D(C_i) (i = 1,2,...,k)$; otherwise jump to(5);

4) If reach the end of data set K, jump to 7); Otherwise, randomly read a new object k_j , and compare k_j $(j = 1,2, ... , n)$ with the mean of each known cluster $C_i (i = 1,2, ... , k)$ in initial clustering result set;

5) If $n_i \leq 1$, $|k_j - E(C_i)| \geq \delta$, or $n_i \geq 2$, $|k_j - E(C_i)| \geq 3\sqrt{D(C_i)} (j = 1,2, ... , n, i = 1,2, ... , k)$, object k_j does not belong to cluster $C_i (i = 1,2, ... , k)$; For all known clusters in clustering result set, if object k_j $(j = 1,2, ... , n)$ does not belong to any cluster, then building a new cluster C_{k+1}, and $n_{k+1} = n_{k+1} + 1$, calculate $E(C_{k+1})$ and $D(C_{k+1})$, jump to 4);

6) Otherwise, object k_j $(j = 1,2, ... , n)$ belongs to cluster $C_i (i = 1,2, ... , k)$ which makes $\min(|k_j - E(C_i)|) (j = 1,2, ... , n, i = 1,2, ... , k)$ existed, $n_i = n_i + 1 (i = 1,2, ... , k)$, recalculate $E(C_i)$ and $D(C_i)(i = 1,2, ... , k)$, jump to 4);

7) End.

4   Experiments Results and Analysis

To demonstrate the validity of algorithm, this paper makes an experiment on the data set IRIS. It belongs to UCI which is a machine learning and intelligent system repository. IRIS is used specially to test clustering algorithms popularly.[10] It consists of 3 classes object (setosa, versicolour and virginica), 4 attributes (sepal length, sepal width, petal length, petal width) and 150 data. All experiments in this paper are implemented by Java under the environment of 1.66GHz Genuine Intel(R) CPU, 1.25GB memory and Microsoft Windows XP Professional. Two dimensions testing data store in text file.

Experiment 1 is used to demonstrate the validity of clustering algorithm without outliers. In order to observe the effects of clustering distinctly, we choose the two most sensible attributes (petal length, petal width) from IRIS, and place them into the two-dimensional coordinate system. Fig.1 describes the original distribution of data in IRIS.
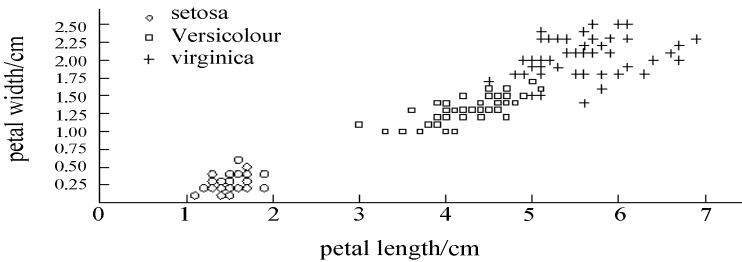
Fig. 1. Original distribution of data in IRIS

As shown in Figure1, IRIS data set is divided into three classes. Each class includes 50 data objects. But the boundary between class versicolour and virginica is indistinct. Class virginica is more distribute than class setosa and versicolour.

After preparing data set, this paper adopts DBSCAN algorithm which is provided by Weka to clustering and detect suspicious outliers. Weka is a practical machine learning tool.[11] It is programmed by Java. Weka provides a uniform interface to apply many different learning methods to any given data set, and evaluates results. In DBSCAN algorithm, the clustering result is changing with the value of parameter Eps and MinPts. Table 1 shows different value and clustering results.

Table 1. Experimental parameter and clustering results

Eps	MinPts	Cluster	Data Count	Suspicious outliers
0.11	10	C_1'	50	(6.6,2.1)
		C_2'	50	(6.9,2.3) (5,1.5)
		C_3'	45	(5.1,1.5) (5.6,1.4)
0.12	10	C_1'	50	(6.9,2.3)
		C_2'	50	(5,1.5)
		C_3'	46	(5.1,1.5) (5.6,1.4)
0.13	10	C_1'	50	(6.9,2.3)
		C_2'	50	(5.6,1.4)
		C_3'	48	
0.14	10	C_1'	50	none
		C_2'	50	
		C_3'	50	

From Table.1 we know that, Eps=0.14 and MinPts=10, clustering result is the best. When MinPts value remains the same, with Eps value reducing, suspicious outliers increase by 2 to 9. However, no matter how the value of parameters Eps and MinPts, IRIS data set is divided into three clusters under DBSCAN algorithm. Table 2 describes the mean and variance of each cluster under different parameter.

Table 2. Mean and variance of clusters

Eps		0.11	0.12	0.13	0.14
C_1'	$D(C_1')$	(0.03,0.011)	(0.03,0.011)	(0.03,0.011)	(0.03,0.011)
	$E(C_1')$	(1.464,0.244)	(1.464,0.244)	(1.464,0.244)	(1.464,0.244)
C_2'	$D(C_2')$	(0.221,0.039)	(0.221,0.039)	(0.221,0.039)	(0.221,0.039)
	$E(C_2')$	(4.26,1.326)	(4.26,1.326)	(4.26,1.326)	(4.26,1.326)
C_3'	$D(C_3')$	(0.26,0.06)	(0.28,0.059)	(0.278,0.069)	(0.305,0.075)
	$E(C_3')$	(5.52,2.066)	(5.543,2.057)	(5.523,2.033)	(5.552,2.026)

As Eps=0.14 and MinPts=10, clustering result is the best. We choose the mean $E(C_i)(i = 1,2,3)$ as initial clustering centers while Eps=0.14 and MinPts=10. Table 3 describes the comparison between IRIS and clustering result.

Table 3. Comparison between IRIS and clustering result

		setosa	versicolour	virginica
IRIS	Object count	50	50	50
Data set	EX	(1.46,0.24)	(4.26,1.32)	(5.55,2.02)
	DX	(0.03,0.01)	(0.22,0.03)	(0.30,0.07)
Clustering	Object count	50	61	39
algorithm	EX	(1.46,0.24)	(4.38,1.40)	(0.57,0.23)
	DX	(0.03,0.01)	(0.25,0.06)	(0.23,0.06)

In this experiment, we set 0.95 as deviation threshold δ. Table.3 shows that IRIS data set is divided into three classes by clustering algorithm. Each class separately includes 50, 61, 39 data objects. The distribution of versicolour obtained by clustering algorithm is slightly scattered than it in IRIS. And data in virginica generated by clustering algorithm is more intensive than it in IRIS.

5   Conclusion

This paper proposes an outlier mining algorithm based on probability, and makes a new definition for outlier. It combines probability with density-based algorithm, and needs not suppose the distribution of data set and the count of outlier. Experiment results on IRIS indicate that the algorithm proposed in this paper can detect outlier from suspicious outlier set, and put non-outlier into cluster which has similar characteristic with it.

However, this new algorithm for outlier mining based on probability still has some problems to further study duo to the limited conditions. For example, the data set which has category attributes beyond the studying scope of this paper and is a topic for future study.

References

1. Zhai, D., Yu, J., Gao, F., Yu, L., Ding, F.: K-means text clustering algorithm based on initial cluster centers selection according to maximum distance. Application Research of Computer 31(3), 713–715 (2014)
2. Xia, L.N., Jing, J.W.: SA-DBSCAN: A self-adaptive density-based clustering algorithm. Journal of the Graduate School of the Chinese Academy of Sciences 26(4), 530–538 (2009)

3. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: LeCam, L., Neyman, J. (eds.) Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability, pp. 281–297. University of California Press, Berkeley (1967)
4. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining. Post & Telecom Press, Beijing (2006)
5. Ester, M., Kriegel, H.P., Sander, J.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: Simoudis, E., Han, J.W., Fayyad, U.M. (eds.) Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, pp. 226–231. AAAI Press, Portland (1996)
6. Shen, H.: Probability and Statistics, 5th edn. Higher Education Press, Beijing (2011)
7. Yu, Y., Zhou, A.: An Improved Algorithm of DBSCAN. Computer Technology and Development 21(2), 30–33, 38 (2011)
8. Daszykowski, M., Walczak, B., Massart, D.L.: Looking for Natural Patterns in Data. Chemometrics and Intelligent Laboratory Systems 56(2), 83–92 (2001)
9. Chen, S., He, Y.J., Zhen, M.G.: NPP-oriented intelligent diagnose. Nuclear Power Engineering and Technology (3), 20–24 (2003)
10. Center for Machine Learning and Intelligent Systems at the University of California, Irvine, http://archive.ics.uci.edu/ml/datasets/Iris
11. Witten, I.H., Frank, E., Hall, M.A.: Data Mining Practical Machine Learning Tools and Techniques, 3rd edn. Morgan Kaufmann (2011)

The Comparison between IABC with EGARCH in Foreign Exchange Rate Forecasting

Jui-Fang Chang[1], Pei-Wei Tsai[2,*], Jung-Fang Chen[3], and Chun-Tsung Hsiao[1]

[1] Department of International Business,
National Kaohsiung University of Applied Sciences,
415 Chien-Kung Road, Kaohsiung City 80778, Taiwan, R.O.C.
[2] Department of Marine Information and Technology,
National Kaohsiung Marine University,
482 Jong-Jhou 3rd Road, Kaohsiung City 80543, Taiwan, R.O.C.
[3] Department of Business Administration,
National Kaohsiung University of Applied Sciences,
415 Chien-Kung Road, Kaohsiung City 80778, Taiwan, R.O.C.
{rose,afan}@cc.kuas.edu.tw,
{peri.tsai,chuntsung0307}@gmail.com

Abstract. Foreign exchange rate forecasting catches many researchers interests in recent years. Problems of the foreign exchange rate forecasting model selection and the improvement on forecasting accuracy are not easy to be solved. In this paper, the forecasting results obtained by conventional time-series models and by the Inter-active Artificial Bee Colony (IABC), which is a young artificial intelligent meth-od, are compared with each other with 4 years historical data. The sliding win-dow strategy is used in the experiment for both the training and the testing phases. In our experiments, we use continuous previous three days data as the training set, and use the training result to forecast the foreign exchange rate on the fourth day. In addition, we evaluate the forecasting accuracy with three criteria, namely, Mean Square Error (MSE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). The experimental results indicate that feeding macroeco-nomic factors to IABC as the input data is capable to produce higher accurate data in the foreign exchange rate than the conventional time-series models such as EGARCH.

Keywords: IABC, Foreign Exchange Rate Forecasting, Time-series, EGARCH.

1 Introduction

Foreign exchange rate forecasting is an important issue in finance. The forecasting result is quite sensitive to the selected forecasting model is one of the key factor re-sulting in the forecasting accuracy. In addition, the foreign exchange

---

* Corresponding author.

rate is quite sensitive to many factors such as price index, interest rates, money supply, balance of trade, and so forth. As the matter of fact, the forecasting accuracy is directly affected by the selection of the referenced variables and information fed into the forecasting model. Although the monetary theory model has higher accuracy than the random walk method [13] has been revealed, the performance of the monetary theory model is still limited [6]. Furthermore, most of the literatures focus on the long-term ex-change rate forecasting, only a few literatures discussed about the short-term, i.e. the daily exchange rate forecasting topic. It implies that the existing foreign exchange rate forecasting models are not sharp at short-term responses.

For the investors, who involve in the foreign exchange market, the profit comes from the short-term investment is much worth than from the long-term investment. Hence, the urgent need of new foreign exchange rate forecasting model for short-term prediction becomes clear. Answering to the need, we propose a new foreign exchange rate forecasting model based on Interactive Artificial Bee Colony (IABC) optimiza-tion, which is a young swarm intelligence algorithm, for the short-term foreign ex-change rate prediction. Our model requires the last 3 continuous historical data as the input for producing the predicted result for the next day.

The rest part of this paper is composed as follows: the literature review is given in section 2, our proposed method and the experiment design are described in section 3, the experiment design is depicted in section 4, the experimental results are given in section 5, and finally, the conclusion is made in section 6.

2   Literature Review

In this section, we first focus on the foreign exchange rate theory. A brief review on EGARCH is given in the next paragraph and is followed by the brief review on IABC algorithm.

2.1   Foreign Exchange Rate Theory

Many different theories of foreign exchange rate determination, including Purchas-ing Power Parity (PPP), monetary model, Interest Rate Parity (IRP), balance of pay-ment model, portfolio balance model, and etc. are proposed in finance history. The brief review of the models listed above are given as follows:

2.2   Purchasing Power Par (PPP) and Monetary Model

The PPP theory claims that the exchange rate between two countries currencies are equal to the ratio of their price levels because the Purchasing Power of a countrys currency is reflected in the countrys price level, the money price of a reference bas-ket of goods and its services. The PPP theory is generated based on the concept of the arbitrage across goods markets and the law of one price [9]. The PPP theory is a foun-dation of many other economic models, e.g. the

Monetary model. Monetary School uses the monetary supply and the demand side to define the exchanges of exchange rates. This model is proposed based on the PPP theory [12].

2.3   Interest Rate Parity Theory

The interest pate parity (IRP) theory establishes the joint between the spot currency market and the forward currency market with foreign and domestic market. The IRP is maintained by the arbitrage.

2.4   Portfolio Balance Model (PBM)

The PBM can be treated as an extension of the monetary model. PBM is proposed by Tobin, who claims that people having different assets should undertake different returns and risks. These two items should be assessed by their own returns and risks in order to determine the optimal asset portfolio. This model also treats the expected returns of different financial assets existing in different countries as the primary fac-tors which affect the exchange rate. The main factors which affect these expected returns are the interest rates of domestic, foreign financial assets and the expectation of exchange rate between domestic and foreign countries.

2.5   Balance of Payment Model

The balance of payment model indicates that the equilibrium exchange rate should be the one, which makes the surplus or the deficit of balance from the payment of the country equals to zero. If the terms described above do not equal to zero, the ex-change rate must be fluctuated [4].

Many practical exchange rate forecasting models are introduced in the past, for in-stance, balance of payment model [4], monetary model [12] and purchasing power parity model [9]. These models use a single structural model to find out which factor gives the effect on the foreign exchange rate. In recent years, researchers working in the related research field usually use the models mentioned above with macroeco-nomic factors to construct the linear model for simulating the exchange rate fluctua-tion for the exchange rate forecasting.

2.6   Time-Series Model: EGARCH

Bollerslev (1986) proposed a Glossary-ARCH (GARCH) [3] model based on Engles (1982) AutoRegressive Conditional Heteroskedastic (ARCH) class of models. Alt-hough the ARCH model works well in analysis of the heteroscedasticity difference of a time series, the prediction is still not accurate in some cases. The GARCH is a time series regression model, which is designed especially for the financial data. The GARCH model not only construct the model for the time series data, but also create models for the modeling error. This results in the

GARCH model present good results in the analysis and the prediction of the volatility. This information is quite im-portant and useful for the investors and can be treated as a leading index.

Although GARCH model provides information of the volatility, it is still not able to depict the leverage effect in the data. Hence, Nelson (1991) propose the Exponen-tial General Auto-Regressive Conditional Heteroskedastic (EGARCH) model [14] to overcome this problem. An EGARCH(p,q) model can be described as follows:

$$\log \sigma_t^2 = \omega + \sum_{k=1}^{q} \beta_k g(Z_{t-k}) + \sum_{k=1}^{p} \alpha_k \log \sigma_{t-k}^2 \ . \tag{1}$$

$$g(Z_{t-k}) = \theta Z_t + \lambda[|Z_t|t - E(|Z_t|)] \ . \tag{2}$$

where σ_t^2 indicates the conditional variance, α, β, ω, θ are coefficients, re-spectively, and Z_t comes from the generalized error distribution or the standard normal varia-ble. The sign and the magnitude of Z_t are capable to donate sep-arate effects on the volatility in Eq. (2). This is particularly useful in an asset pricing context [16]. In addition, the parameters are bounded with less restric-tions because the value of $\log \sigma_{t-k}^2$ may be negative.

2.7 Swarm Intelligence and the Interactive Artificial Bee Colony (IABC)

Swarm intelligence is an artificial intelligence technique based on the study of col-lective behavior [2]. Many of the algorithms in swarm intelligence are developed based on simulating the behaviors of the creatures in the Mother Nature. Swarm intel-ligence is symbolically made up of a population of simple agents interacting locally with one another and with their environment. Even though there is no centralized control structure indicating how individual agents should behave, local interactions between such agents often lead to the emergence of global behavior.

We have seen many successful applications using swarm intelligence methods to solve problems in optimization. For example, Particle Swarm Optimization (PSO) algorithm has successfully been used to design antennas [21] and to con-struct param-eters in neural network systems [10]; Ant Colony Optimization (ACO) algorithm has successfully been used to solve the Traveling Salesman Problem (TSP) [8] and the routing problem of networks [17]; Artificial Bee Colony (ABC) algorithm has suc-cessfully been used to solve the lot-streaming flow shop scheduling problem [15]; Cat Swarm Optimization (CSO) algorithm has successfully been used to discover proper positions for information hiding [19] and to adjust parameters for the Support Vector Machine (SVM) [11]. Moreover, many researchers have led in the idea of par-allelizing the optimization methods by splitting the artificial agents into independ-ent groups such as the Island-model genetic algorithm [20], the parallel genetic algo-rithm [1], the ant colony

system with communication strategies [5], and the parallel particle swarm opti-mization algorithm with communication strategies [7]. The paral-lelized group of artificial agents increases the accuracy and extends the global search capacity than the original structure.

The Interactive Artificial Bee Colony is proposed by Tsai et al. in 2008. [18] It is evolved from the Artificial Bee Colony Optimization (ABC), and it is devel-oped based on researchers notice that the moving pattern of the bee of original algorithm to search nectar is linear movement, and it will narrow the explore scope, so we put forward to join law of universal gravitation to improve the short-comings of the origi-nal colony algorithm. The gravitations between the onlooker and the selected em-ployed bees are concerned, thus, it can be calculated by ap-plying equation. To im-plement the IABC optimization, the procedures are given as follows:

Step 1. Initialization: Put n_e percentage of the populations into the solution space randomly, then calcu-late their fitness values, which called the nectar amounts, where means the ratio of employed bees to the whole population. In case these populations are positioned into the solution space, they will be the employed bees.

Step 2. Move the Onlookers: Calculate the probability of selecting a food source by Eq. (3), choose a food source to move to by roulette wheel selection for every onlooker bees, and then decide the nectar amounts of them. The movement of the onlooker bees follows Eq. (4).

$$P_i = \frac{F(\theta_i)}{\sum_{k=1}^{S} F(\theta_k)} . \tag{3}$$

$$x_{ij}(t+1) = \theta_{ij}(t) + \sum_{k=1}^{n} \tilde{F}_{ik_j} \cdot [\theta_{ij}(t) - \theta_{kj}(t)] . \tag{4}$$

where x_i represents the position of the ith onlookers, t denotes the iteration number, θ_k represents the randomly chosen employed bee, j is the dimension of the solution, and $\tilde{F}_{ik_j}$ is the normalized gravitation.

Step 3. Move the Scouts: Supposing the fitness values of the employed bees is limited, which means they do not be improved by a continuous predetermined number of iterations, those nectar are abandoned and these employed bees be-come the scouts. To move the scouts, Eq. (5) is applied.

$$\theta_{ij} = \theta_{jmin} + r(\theta_{jmax} - \theta_{jmin}) . \tag{5}$$

where r is a random number and $r \in [0, 1]$.

Step 4. Update the Best Food Source: Memorizing the best fitness value and the best position, which are found by the bees.

Step 5. Termination Checking: Stop when the amount of the iterations satis-fies the termination condition, and then terminate the program and output the results; on the contrary, go back to *Step 2* to restart.

2.8   Our Proposed Method and Experiment Design

To establish a new exchange rate forecasting model, we first collect the factors for the reference and then utilize IABC to find the optimum weighting distribution for con-structing the predicted daily exchange rate. Eight macroeconomic factors including Consumer Price Index, M1, M1B, Commercial Paper Rate, Federal Fund Rate, Bal-ance of Trade, Foreign Investment, and Stock Return are taken with the NTD/USD exchange rate to be the reference data. According to the collected historical data in the database, most of the factors only provides monthly records. Only the NTD/USD exchange rate has both monthly and daily records. The weekly record of Stock Return can also be found in the database. Hence, our reference information is composed of totally 13 elements including the monthly records of the eight macroeconomic factors and the NTD/USD exchange rate, the weekly record of the Stock Return, 2 daily records of the NTD/USD exchange rate and the Stock Return, and one constant, which is set to 1 in the experiment.

We use IABC to train a set of considering weightings correspond to the reference information. The goal is to output the estimated foreign exchange rate, which should approach the exact foreign exchange rate as much as it can. Hence, the prediction error is used in the fitness function to train the considering weightings. The goal of this optimization process is to produce a set of considering weighting by minimizing the prediction error. The fitness function used in IABC is listed in Eq. (6):

$$min f(W) = \sum_{t-1}^{n} |(\sum_{d=1}^{D} w_d \times v_{t,d}) - R_{real,t}| \ . \tag{6}$$

where $f(W)$ indicates the fitness function, $W = \{w_1, w_2, ..., w_D\}$ is the set of the considering weights, n denotes the number of past dates of which the reference information is taking part in the optimization process, D is the total number of reference infor-mation (D is set to 13 in our experiment), v denotes the reference information, and R_{real} stands for the real foreign exchange rate.

The trained considering weights are output as the prediction parameter and the predicted foreign exchange rate for the next date is calculated by Eq. (7):

$$R_{pd,t+1} = \sum_{d=1}^{D} w_d \times v_{t,d} \ . \tag{7}$$

where R_{pd} denotes the predicted foreign exchange rate.

In addition, the forecasting accuracy is evaluated by the Mean Absolute Error (MAE), the Mean Square Error (MSE), and the Root Mean Square Error (RMSE) by Eqs. (8)-(10):

$$MAE = \sum \frac{|\hat{S}_t - S_t|}{n} \ . \tag{8}$$

$$MSE = \sum \frac{(\hat{S}_t - S_t)^2}{n} \ . \tag{9}$$

$$RMSE = \sqrt{\sum \frac{(\hat{S}_t - S_t)^2}{n}} \ .$$

(10)

where $\hat{S}_t$ is the forecast exchange rate at day t; S_t represents the exchange rate at day t, and n is the number of data.

The historical data from the first trading day in February in 2005 to the last trading day in 2008 is included in our experiment. The collected data is examined by the basic test at the first beginning. The test result indicates that the collected data fits the criteria of the regular data and is suitable for the usage in our experiment. Since our proposed method is for forecasting the daily exchange rate and we only require the last three days' data for training, the sliding window strategy is employed in our experiment to run over totally 2922 days. Same strategies is also applied with the EGARCH model. The concept of the sliding window strategy is depicted in Fig. 1:
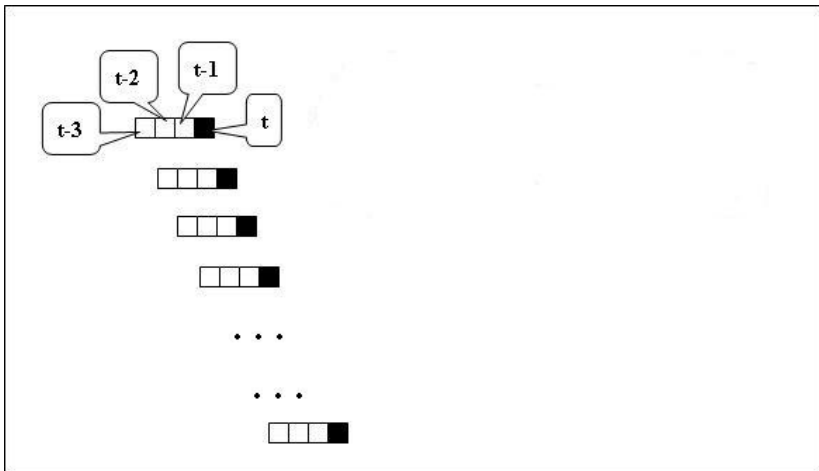
Fig. 1. The Concept of the Sliding Window Strategy

Here we give two examples to demonstrate how to use the sliding window strate-gy in the experiment. As shown in Fig. 1, if we are going to predict the exchange rate on day t, the reference data should be collected from the latest past three days, i.e. from days $(t-3)$ to $(t-1)$. On the other hand, if we want to predict the exchange rate on day $(t+1)$, the reference data should be collected from days $(t-2)$ to t.

The parameter setting for IABC is listed as follows: the population size is 16, the agents are equally divided into 4 independent groups, and the initial range is set to $[-1, 1]$. Half of the agents play the role of onlookers. The optimization process is set to run 120 iterations with 30 independent runs. The best result over all runs is output-ted as the final result.

2.9   Experimental Results

The prediction results produced by the IABC forecasting model and the EGARCH model are drawn year-by-year in Fig.2 with the actual foreign exchange rage. The vertical axis shows the foreign exchange rate; and the horizontal axis shows the date in a year. The experimental results show that the forecasting result obtained by the IABC forecasting model almost stick together with the actual foreign exchange rate from the beginning to the end in year 2005 and 2008. In year 2006 and 2007, the forecasting result obtained by the IABC forecasting model show the oscillation in some days. It may be caused by the failure to find the near best solution, but trapped in the local optimum. On the other hand, the forecasting results obtained by the EGARCH model shows more oscillation and larger oscillation amplitude over all forecasting results.
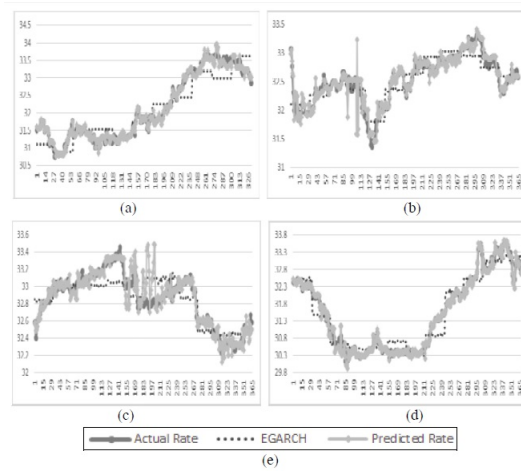
Fig. 2. The actual foreign exchange rate and the forecasting results obtained by the IABC forecasting model and the EGARCH model: (a) results in 2005, (b) results in 2006, (c) results in 2007, (d) results in 2008, (e) the graphic symbols

Table 1 shows the compare on forecasting results from IABC trained foreign ex-change rate forecasting model and the conventional EGARCH forecasting models with MAE, MSE, and RMSE.

Our proposed IABC forecasting model presents the lowest MSE and RMSE values over all experiment data. The MAE value is larger than the EGARCH model in 2006. It is caused by the failure to find the near best solution around the 100th days in 2006.

2.10   Conclusions

In this paper, we propose a new daily foreign exchange rate forecasting model with IABC algorithm. IABC plays the role to construct the forecasting exchange

Table 1. Satisfaction Capability Index of the Investment Performance

Year	Forecasting Model	MSE	MAE	RMSE
2005	IABC result	8.365×10^{-3}	6.321×10^{-2}	9.146×10^{-2}
	EGARCH result	1.237×10^{-1}	2.778×10^{-1}	3.518×10^{-1}
2006	IABC result	1.735×10^{-2}	7.389×10^{-2}	1.317×10^{-1}
	EGARCH result	4.312×10^{-2}	1.673×10^{-1}	2.077×10^{-1}
2007	IABC result	1.609×10^{-2}	7.147×10^{-2}	1.269×10^{-1}
	EGARCH result	2.905×10^{-2}	1.444×10^{-1}	1.704×10^{-1}
2008	IABC result	9.339×10^{-3}	5.792×10^{-2}	9.664×10^{-2}
	EGARCH result	8.906×10^{-2}	2.307×10^{-1}	2.984×10^{-1}

rate by finding the optimum combination and the weighting distribution from the past three continuous day's information. The forecasting results are compared with the classical conventional time-series model, e.g. the EGARCH model. The experimental results indicate that our proposed method provides the forecasting result with high accuracy. Although the vibrations of the forecasting result sometimes appears, it still doesnt cause much drop on the accuracy. The vibration may be overcame by increasing the population size or the iteration number. In the future, we plan to further use IABC to reduce the number of the referred factors. By doing so, the computational cost is able to be reduced.

References

1. Abramson, D., Abela, J.: A parallel genetic algorithm for solving the school timetabling problem. In: Appeared in 15 Australian Computer Science Conference, Hobart, Australia, pp. 1–11 (1991)
2. Beni, G., Wang, J.: Swarm Intelligence in Cellular Robotic Systems. In: NATO Advanced Workshop on Robots and Biological Systems, Tuscany, Italy (1989)
3. Bollerslev, T.: Generalized Autoregressive Conditional Heteroskedasticity. Journal of Econometrics 31, 307–327 (1986)
4. Branson, W.H.: Flow and stock equilibrium in a dynamic metzler model. Journal of Finance 31(5), 1323–1339 (1976)
5. Chang, J.-F., Chu, S.-C., Roddick, J.F., Pan, J.-S.: A parallel particle swarm optimization algorithm with communication strategies. Journal of Information Science and Engineering 21(4), 809–818 (2005)
6. Cheung, Y.W., Chinn, M.D., Pascual, A.G.: Empirical Exchange Rate Models of the Nine-ties: Are Any Fit to Survive? Journal of International Money and Finance 24, 1150–1175 (2005)
7. Chu, S.-C., Roddick, J.F., Pan, J.-S.: Ant colony system with communication strategies. Information Sciences 167(1-4), 63–76 (2004)
8. Dorigo, M., Gambardella, L.M.: Ant colony system: a cooperative learning approach to the traveling salesman problem. IEEE Transactions on Evolutionary Computation 1(1), 53–66 (1997)
9. Frenkel, J.A.: Flexible Exchange Rates, Prices, and the Role of News: Lessons from the 1970s. Journal of Political Economic 89(4), 665–705 (1981)

10. Lin, C.-J., Chen, C.-H., Lin, C.-T.: A hybrid of cooperative particle swarm optimization and cultural algorithm for neural fuzzy networks and its prediction applications. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews 39(1), 55–68 (2009)
11. Lin, K.-C., Chien, H.-Y.: CSO-based feature selection and parameter optimization for support vector machine. In: 2009 Joint Conferences on Pervasive Computing, Taipei, Taiwan, pp. 783–788 (2009)
12. MacDonald, R., Taylor, M.P.: The Monetary Model of the exchange Rate Long-run Relationships, Short-run Dynamics and how to Beat a Random Walk. Journal of International Money and Finance 13(3), 276–290 (1994)
13. Mark, N.C.: Exchange Rates and Fundamentals: Evidence on Long-Horizon Predictability. American Economic Review 85(1), 201–218 (1995)
14. Nelson, D.B.: Conditional Heteroskedasticity in Asset Returns: A New Approach. Econometrica 59, 347–370 (1991)
15. Pan, O.K., Tasgetiren, M.F., Suganthan, P.N., Chua, T.J.: A discrete artificial bee colony algorithm for the lot-streaming flow shop scheduling problem. Information Sciences 181(12), 2455–2468 (2011)
16. Pierre, S., Eilleen, F.: Estimating EGARCH-M Models: Science or Art. The Quarterly Review of Economics and Finance 38(2), 167–180 (1998)
17. Pinto, C.P., Nägele, A., Dejori, M., Runkler, T.A., Sousa, J.M.C.: Using a local discovery ant algorithm for bayesian network structure learning. IEEE Transactions on Evolutionary Computation 13(4), 767–779 (2009)
18. Tsai, P.-W., Pan, J.-S., Liao, B.-Y.: Enhanced Artificial Bee Colony Optimization. International Journal of Innovative Computing, Information and Control ICIC International 5(12(B)), 5081–5092 (2009)
19. Wang, Z.-H., Chang, C.-C., Li, M.-C.: Optimizing least-significant-bit substitution using cat swarm optimization strategy. Information Sciences 192, 98–108 (2012)
20. Whitley, D., Rana, S., Heckendorn, R.B.: The Island Model Genetic Algorithm: On Separability, Population Size and Convergence. Journal of Computing and Information Technology 1305/1997, 109–125 (1998)
21. Wu, H., Geng, J., Jin, R., Qiu, J., Liu, W., Chen, J., Liu, S.: An improved comprehensive learning particle swarm optimization and its application to the semiautomatic design of antennas. IEEE Transactions on Antennas and Propagation 57(10), 3018–3028 (2009)

Equivalence Proof of Traditional and Random Grid-Based (2, 2) Visual Secret Sharing

Shen Wang⋆, Xuehu Yan, Jianzhi Sang, and Xiamu Niu

School of Computer Science and Technology,
Harbin Institute of Technology, Harbin, China
Shen.Wang@hit.edu.cn

Abstract. Visual secret sharing (VSS) has attracted considerable attention to scientists and engineers as another branch alongside conventional cryptography to protect the sensitive visual information from several rapacious behaviors. In the literature, there are a number of several techniques used to protect the visual information, among which traditional VSS and random grid (RG)-based VSS are the primary branches. In this letter, we show, by examples, the two means are equal. In addition, the color representation of traditional VSS and RG-based VSS found it different from digital applications like images. Based on the given examples, it is demonstrated that the color representation of the two means can be the same and confirm with digital processing applications.

Keywords: Visual cryptography, Visual secret sharing, Random grid, Equivalence proof, Color representation.

1 Introduction

Visual cryptography(VC) and general secret image sharing [1][2] protects the secret by sharing the user data into different secret shares (also called shadows) and distributing them among multiple carriers. They have attracted more attention of scientists and engineers. Visual secret sharing (VSS) also called visual cryptography scheme (VCS), and Shamir's polynomial-based scheme are the primary branches in this field.

Naor and Shamir [1] firstly propose the threshold-based VSS. The main properties of the VSS by [1] are simple recovery that is the decryption of secret image is completely based on human visual system (HVS) without any cryptographic computation. However, it also suffers from meaningless shadow images, lossy recovery and pixel expansion [2].

Attributed to the traditional VSS proposed by Naor and Shamir [1], various secret image sharing schemes have been proposed in the literature. However, most of the proposed schemes remained weakness to be resolved before secret

sharing can gain possibly pervasive applications. For instance, the scheme of Giuseppe et al. [3] has meaningful shadow images, but still has lossless recovery and pixel expansion problem. Yang [4] proposes a probability-based visual sharing scheme which is lossless and has no pixel expansion, but doesn't have meaningful shadow images. Wang et al. [5] proposes a secret sharing scheme based on Boolean operation that is lossless. However, these schemes suffer from pixel expansion problem.

Since VSS by random grids (RG) could avoid pixel expansion and has no codebook (basic matrices) needed, some other researchers [6],[12],[7], [9],[11] have paid more attention to RG-based VSS. Encryption of binary secret images based on RG is firstly presented by Kafri and Keren [6], each of which is generated into two noise-like RG (shadow images or share images) that have the same size as the original secret image. The decryption operation is Boolean OR operation which is the same as traditional VSS.

Though the traditional VSS and RG-based VSS are two primary branches of VSS, however they have some similarities about the key idea. In both traditional VSS and RG-based (2, 2) VSS, the color representation is "1" denotes black pixels, "0" denotes white pixels, which may be not convenient [5],[10],[8] in digital images and common digital processing software. In most digital image formats like BMP and JPEG, and common digital image processing related software, such as Matlab and Photoshop, 0 denotes black or opaque pixel value and 1 denotes white or transparent pixel value. Different color representations will increase computation time for reversing or complementing operations (that is $0 \rightarrow 1$ or $1 \rightarrow 0$).

In the paper, we show the traditional and random grid-based (2, 2) VSS is equal by given examples of generation ideas and visual quality. In addition, the color representation of traditional VSS and RG-based VSS found it different from digital images in the previous proposed approaches. Based on the given examples, it is demonstrated that the color representation of the two means is the same with digital images.

The rest of the paper is organized as follows. Section 2 gives the preliminary techniques of traditional VSS and RG-based VSS as the basis for the work. In Section 3, the equivalence proof of the two means is introduced. Finally, Section 4 concludes this paper.

2   Preliminary Techniques

This section introduces the related works of traditional (2, 2) VSS [1] and one typical RG-based [6] VSS.

Shadow images number is 2, the binary secret image is denoted as S with pixel value $S(i,j), 1 \leq i \leq M, 1 \leq j \leq N$. The shadow images covered secret after sharing are denoted as SC. The recovered secret image from $k = 2$ shadow images is denoted as S'. In traditional VSS and RG-based VSS "1" denotes black

pixels, "0" denotes white pixels, the decryption is Boolean OR operation. Here, the secret image "HIT" with size 256×256 is used to illustrate the idea. Symbols $\&, \oplus$ and $\otimes$ denote the Boolean AND, XOR and OR operations, respectively. $\overline{x}$ is a bit-wise complementary operation of x.

2.1 Traditional (2, 2) VSS

In traditional (2, 2) VSS, a binary secret image is shared by generating corresponding two noise-like shadow images. The two noise-like shadow images are superposed to recover the secret image visually based on HVS and probability. However, less than 2 participants cannot reveal any information of the secret image

Fig. 1 shows the idea of traditional (2, 2) VSS, a certain pixel of the secret image is split into a pair of white and black subpixels in each of the two shadow images. The subpixels are randomly selected from the two columns tabulated under the certain secret pixel, which leads to the certain secret pixel is encoded into two subpixels of white-black or black-white with the same probabilities (50%). Hence, an individual shadow image gives no information about the secret image. When the subpixels are stacked, the black secret pixel will be decoded into black pixel, and the white secret pixel into one white pixel and one black pixel. Thus, the secret image could be recovered by stacking the two shadow images together. Fig. 2 is an application example of the (2, 2) VCS, though the pixel expansion is 2 and some contrast is lost, the secret image could be recognized clearly by HVS.

Fig. 1. The idea of traditional (2, 2) VC

Fig. 2. An application example of traditional (2, 2) VC

2.2   RG-Based VSS

The generation and recovery phases of one typical (2, 2) RG-based [6] VSS are described below.

Step 1: Randomly generate 1 RG SC_1

Step 2: Compute SC_2 as in Eq. 1

Recovery: $S' = SC_1 \otimes SC_2$ as in Eq. 2. If a certain secret pixel of $S(i,j)$ is 1, the recovery result $SC_1 \otimes SC_2 = 1$ is always black. If a certain secret pixel of $S(i,j)$ is 0, the recovery result $SC_1 \otimes SC_2 = SC_1(i,j) \otimes SC_1(i,j)$ has half chance to be black or white since SC_1 is generated randomly

$$SC_2(i,j) = \begin{cases} SC_1(i,j) & if\ S = 0 \\ \overline{SC_1(i,j)} & if\ S = 1 \end{cases} \tag{1}$$

$$\begin{aligned} S'(i,j) &= SC_1(i,j) \otimes SC_2(i,j) \\ &= \begin{cases} SC_1(i,j) \otimes SC_1(i,j) & if\ S(i,j) = 0 \\ SC_1(i,j) \otimes \overline{SC_1(i,j)} \ = 1 & if\ S(i,j) = 1 \end{cases} \end{aligned} \tag{2}$$

Fig. 3 shows an application example of original (2, 2) RG-based VSS, among which, the shadow images are generated by the generation phase described above and the recovered secret image is recovered by the recovery phase described above of the original (2, 2) RG-based VSS.

From the generation and recovery phases of original (2, 2) RG-based VSS described above, we can present the idea of RG-based (2, 2) VSS in Fig. 4, a certain pixel of the secret image is generated into a white or black subpixel in each of the two shadow images. The subpixels are randomly selected from the two columns tabulated under the certain secret pixel, which lead to certain secret pixel that is encoded into two subpixels with the same probabilities (50%). Hence, an individual shadow image gives no information about the secret image.

Fig. 3. An application example of original (2, 2) RG-based VSS

Secret pixel				
Matrix collections	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$
Shadow image1				
Shadow image2				
Probability	50%	50%	50%	50%
Stacking result(OR)				

Fig. 4. The idea of original (2, 2) RG-based VSS

When the subpixels are stacked, the black secret pixel will be decoded into black pixel, and the white secret pixel into white pixel or black pixel with the same probabilities (50%). Thus, the secret image could be recovered by stacking the two shadow images together.

3   Equivalence Proof

3.1   Original Idea Equivalence

From Fig. 1 and Fig. 4, if the first column of matrix collections in Fig. 1 of the traditional (2, 2) VSS is gleaned, matrix collections in Fig. 4 of the RG-based (2,

2) VSS will be gained. This means that in the generation phase of traditional (2, 2) VSS, if only the first column is chosen randomly to the two shadow images, the same recovering result could be obtained.

One example is shown in Fig. 5, in the example, the traditional (2, 2) VSS is applied to share the secret image.

First, the shadow images with size $M \times 2N$ are randomly assigned 0 or 1, the pixel expansion $m = 2$ since the pixel expansion of traditional VSS.

Second, for every pixel $S(i, j), 1 \leq i \leq M, 1 \leq j \leq N, M = N = 256$ of the secret image, the pixel values of shadow images $SC_1(i, j), SC_2(i, j)$ are decided by the first column of matrix collections in Fig. 1 of the traditional (2, 2) VSS randomly.

When the two shadow images are stacked together, the secret will be revealed in the left half part of the recovered secret image, that means if only the left part of the shadow images is assigned, there will only be the left part of the recovered secret image ,the result could be clearly shown in Fig. 6. The illustration is caused by the first column of matrix collections in Fig. 1 of the traditional (2, 2) VSS. Hence, the key idea of RG-based (2, 2) VSS is the same as traditional (2, 2) VSS.
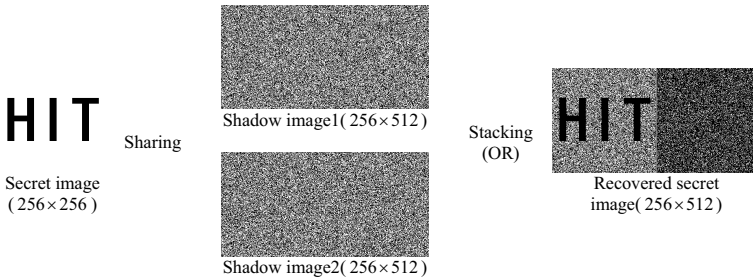
Fig. 5. An equivalence example of traditional (2, 2) and RG-based VC

3.2 Visual Quality Equivalence

Based on Fig. 3 and Fig. 6, the recovered secret images are the same based on human visual system (HVS), which means that the visual quality of the recovered secrets are the same, and the visual quality of traditional and RG-based (2, 2) VSS are equal. Furthermore, the equivalence of theoretical visual quality is analyses as follows:

The probability of pixel color is transparent (0) say $P(x = 0)$ and the same for the probability of pixel color is opaque (1). Besides, $\sum_{i=1}^{M} \sum_{j=1}^{N} X(i, j), 1 \leq i \leq M, 1 \leq j \leq N$.

The visual quality of traditional (2, 2) VSS [1] is evaluated by α_t, the relative difference in the Hamming weight, i.e., the loss in contrast between the recovered

Left part of shadow image1(256×256)

Stacking
(OR)

Recovered secret image(256×256)

Left part of shadow image2(256×256)
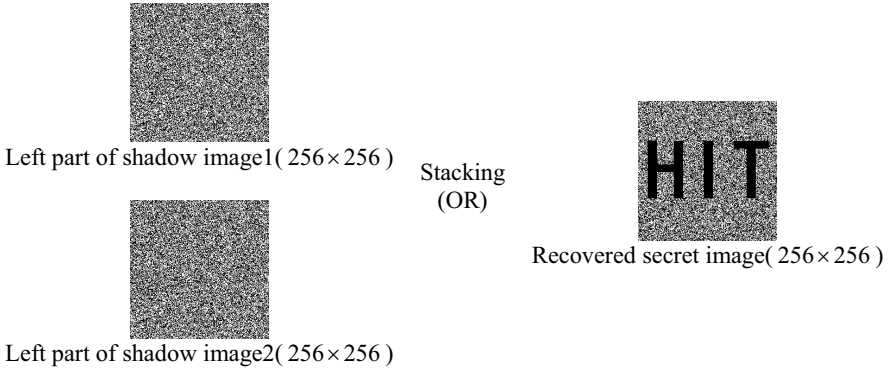
Fig. 6. Stacking result of the left part of the shadow images in Fig. 5

secret pixels that is steaming from a white and black pixel in the original secret image. Hence, the contrast of traditional (2, 2) VSS is 1/2.

The visual quality and security of RG-based (2, 2) VSS [7] is evaluated by contrast α defined as follows:

Definition 1 (Contrast): The visual quality, which will decides how well human eyes could recognize the recovered image, of the recovered secret image S' corresponding to the original secret image S is evaluated by contrast defined as follows:

$$\alpha = \frac{P_0 - P_1}{1 + P_1} = \frac{P(S'[AS0] = 0) - P(S'[AS1] = 0)}{1 + P(S'[AS1] = 0)} \tag{3}$$

where α denotes contrast, P_0(resp.,P_1) is the appearance probability of white pixels in the recovered imageS'in the corresponding white (resp., black) area of original secret image$S$. $AS0$ (resp., $AS1$) is the white (resp., black) area of original secret image S, $AS0 = \{(i,j)\,|S(i,j) = 0, 1 \leq i \leq M, 1 \leq j \leq N\}$

The visual quality of RG-based (2, 2) VSS is evaluated by contrast, which is light contrast defined based on average light transmission of all the recovered secret pixels. Hence, the contrast of RG-based (2, 2) VSS is $\frac{P_0 - P_1}{1 + P_1} = \frac{\frac{1}{2} - 0}{1 + 0} = \frac{1}{2}$

Additionally, from Fig. 1, if the secret pixel is white (0), in the recovered secret pixels(01 or 10) the light transmission will be 0.5; if the secret pixel is black(1), in the recovered secret pixels(11) the light transmission will be 0. Hence, if the **Definition1** is applied for traditional (2, 2) VSS, the contrast of traditional (2, 2) VSS will be $\frac{P_0 - P_1}{1 + P_1} = \frac{\frac{1}{2} - 0}{1 + 0} = \frac{1}{2}$ which is the same as RG-based (2, 2) VSS.

From the contrast experimental results, the same situation could also be gained. The contrast calculated by **Definition1** of Fig. 3 is 0.49948, and contrast calculated by **Definition1** of Fig. 6 is 0.49979, which are very similar to each other, and close to theoretical value 0.5.

Based on the above analyses, it is demonstrated that the traditional (2, 2) VC and RG-based (2, 2) VSS have the same visual quality.

The security and visually recognizable of both traditional VSS and RG-based on VSS [1] is defined as follows:

Definition 2 (Security and visually recognizable): The recovered secret image S' could be recognized as the corresponding original secret image S if $\alpha > 0$ when $t \geq k$. The VSS is secure if $\alpha = 0$ when $t < k$ which means no information of S could be recognized through S'.

From **Definition1,** the security and visually recognizable of traditional VSS and RG-based on VSS is evaluated by the contrast, which means that the equal contrast will lead to equal security and visually recognizable, since they are defined or decided by the contrast of shadow images and recovered secret.

3.3   Further Discussion

The color representation of traditional and RG-based VSS is different from digital images. Furthermore, Yan et al. [8] have proposed a powerful RG-based VSS based on Boolean operations. This scheme has the same color representation method with digital images, where "1" denotes white pixels, "0" denotes black pixels

Herein, a $(2, 2)$ threshold scheme is used as an example firstly to show the idea of Yan's scheme [8] . The generation and recovery phases are described below "1" denotes white pixels, "0" denotes black pixels

Step 1: Randomly generate 1 RG SC_1

Step 2: Compute SC_2 as in Eq. 4.

Recovery: $S' = SC_1 \& SC_2$ as in Eq. 5. If a certain secret pixel of $S(i,j)$ is 0, the recovery result $SC_1 \& SC_2 = 0$ is always black. If a certain secret pixel of $S(i,j)$ is 1, the recovery result $SC_1 \& SC_2 = SC_1(i,j) \& SC_1(i,j)$ has half chance to be black or white since SC_1 are generated randomly

$$SC_2(i,j) = \begin{cases} SC_1(i,j) & if\ S(i,j) = 1 \\ \overline{SC_1(i,j)} & if\ S(i,j) = 0 \end{cases} \tag{4}$$

$$\begin{aligned} S'(i,j) &= SC_1(i,j) \& SC_2(i,j) \\ &= \begin{cases} SC_1(i,j) \& SC_1(i,j) & if\ S(i,j) = 1 \\ SC_1(i,j) \& \overline{SC_1(i,j)} = 0 & if\ S(i,j) = 0 \end{cases} \end{aligned} \tag{5}$$

In addition, the idea applied in Yan's scheme, that is the color representation method is the same with color representation method of digital images not only could be applied in RG-based VSS, but also can be extended to traditional VSS. Herein, the traditional $(2, 2)$ VSS is used as an example. Fig. 7 is an extended example of $(2, 2)$ VSS, Here "0" denotes black pixels, "1" denotes white pixels. A certain pixel of the secret image is split into a pair of white and black subpixels in each of the two shadow images. When the subpixels are stacked (corresponding to Boolean AND operation), the black secret pixel will be decoded into black pixel, and the white secret pixel into one white pixel and one black pixel. Thus, the secret image could be recovered by stacking the two shadow images together. Fig. 8 is an extended application example of the traditional $(2, 2)$ VSS, the secret image could be recognized clearly by HVS.

Secret pixel							
Basic matrices		$\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$			$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$		
Matrix collections	$\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$		$\begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$		$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	
Shadow image1							
Shadow image2							
Probability	50%		50%	50%		50%	
Boolean AND result	$(0 \quad 1)$		$(1 \quad 0)$	$(0 \quad 0)$		$(0 \quad 0)$	
Stacking(&) result							

Fig. 7. An extended example of traditional (2, 2) VC
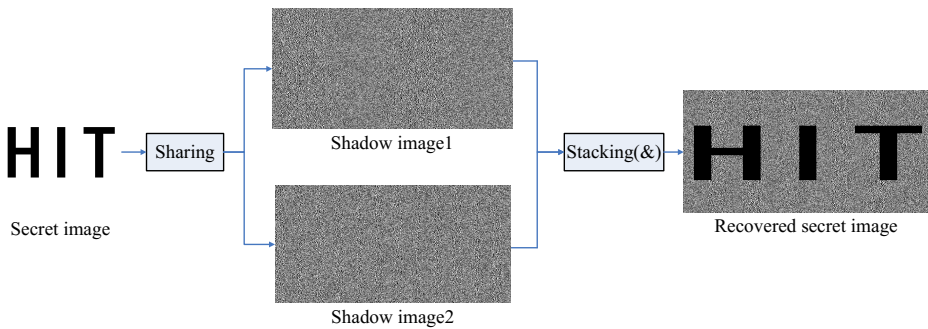
Fig. 8. An extended application example of traditional (2, 2) VC

4   Conclusion

In this paper, the traditional VSS and RG-based VSS found equal by the given examples. In addition, the color representation of traditional VSS and RG-based VSS found it different from digital images. Based on the given examples, it is showed that the color representation of the two means is the same with digital images. The presented work will be useful for understanding the relationship between the traditional VSS and RG-based VSS, and worthwhile in more digital applications.

References

1. Naor, M., Shamir, A.: Visual cryptography. In: De Santis, A. (ed.) EUROCRYPT 1994. LNCS, vol. 950, pp. 1–12. Springer, Heidelberg (1995)
2. Weir, J., Yan, W.: A comprehensive study of visual cryptography. In: Shi, Y.Q. (ed.) Transactions on DHMS V. LNCS, vol. 6010, pp. 70–105. Springer, Heidelberg (2010)
3. Ateniese, G., Blundo, C., De Santis, A., Stinson, D.R.: Extended capabilities for visual cryptography. Theor. Comput. Sci. 250(1), 143–161 (2001)
4. Yang, C.-N.: New visual secret sharing schemes using probabilistic method. Pattern Recognit. Lett. 25(4), 481–494 (2004)
5. Wang, D., Zhang, L., Ma, N., Li, X.: Two secret sharing schemese based on Boolean operations. Pattern Recognit. 40(10), 2776–2785 (2007)
6. Kafri, O., Keren, E.: Encryption of pictures and shapes by random grids. Optics Letters 12(6), 377–379 (1987)
7. Shyu, S.J.: Image encryption by random grids. Pattern Recognition 40(3), 1014–1031 (2007)
8. Yan, X., Wang, S., El-Latif, A.A.A., Niu, X.: Visual secret sharing based on random grids with abilities of AND and XOR lossless recovery. IEEE Trans. on Circ. and Sys. for Video Tech. 21(11), 1693–1703 (2011); Accepted for publication in Multimedia tools and application (2013), doi: 10.1007/s11042-013-1784-2,2013
9. Wu, X., Sun, W.: Improving the visual quality of random grid-based visual secret sharing. Signal Processing (2012)
10. Wang, Z., Arce, G.R., Di Crescenzo, G.: Halftone visual cryptography via error diffusion. IEEE Trans. Inf. Forensics Security 4(3), 383–396 (2009)
11. Guo, T., Liu, F., Wu, C.K.: Threshold visual secret sharing by random grids with improved contrast. Journal of Systems and Software (2013)
12. Chen, T.-H., Tsao, K.-H.: Threshold visual secret sharing by random grids. Journal of Systems and Software 84(7), 1197–1208 (2011)

Part III

Smart Living Technology

Assistive Listening System Using a Human-Like Auditory Processing Algorithm

Po-Hsun Sung[1], Jhing-Fa Wang[1,2], and Hsien-Shun Kuo[1]

[1] Department of Electrical Engineering, National Cheng Kung University, Taiwan
[2] Department of Digital Multimedia Design, Tajen University, Pingtung County, Taiwan
`{phsung,shun7957}@gmail.com, wangjf@mail.ncku.edu.tw`

Abstract. Enhancing the quality of hearing perception in noisy environments plays a significant role to improve life quality of elderly persons and hearing impaired people. Accordingly, this study presents a human-like auditory processing (HAP) algorithm to enhance the speech signal in low signal-to-noise ratio (SNR) and non-stationary noise environments. The proposed algorithm comprises two modules, namely Cochlear Wavelet Transform (CWT) and AM-FM Demodulation (AFD); mimicking the human peripheral auditory processing system and the human cortical auditory processing system, respectively. The performance of the proposed HAP algorithm is evaluated in accordance with the ITU Perceptual Evaluation of Speech Quality (PESQ) standard. The results show that the proposed algorithm improves the speech quality by 16.9 % on average. In the other words, the algorithm has significant potential for assistive listening in noisy environments.

Keywords: Auditory processing, wavelet transform, coherent demodulation, spectro-temporal modulation filter, non-stationary signal, cortical representation, assistive listening, speech enhancement.

1   Introduction

The speech signal received by hearing impaired is often corrupted by the communication channel and degraded by background additive noise. As a result, speech enhancement techniques are required to improve the quality and intelligibility of the spoken words. Various methods have been proposed in the literature for suppressing the effects of background noise, including spectral subtraction methods [1] spectral amplitude estimation methods, [2] statistical model-based approaches [3], and subspace-based algorithms [4]. However, these methods generally distort the speech signal because the background noise is non-stationary and not easily estimated. Accordingly, the problem of developing more sophisticated speech enhancement techniques has attracted increasing attention in recent years.

The speech analysis-modification-synthesis (AMS) method proposed by Potamianos and Maragos [5] has been successfully applied to a broad range of speech processing problems. The AMS method consists of three modules, (1) the analysis module, where the speech is transformed into a speech spectrum using the short-time

Fourier transform (STFT), (2) the modification module, where the speech spectrum undergoes modulation domain processing to reduce the noise spectrum, and (3) the synthesis module, where the output speech is reconstructed using the inverse STFT. The modification part of AMS decomposes the complex speech signal into two components, namely the slow time-varying amplitude envelope (referred to as the modulator) and the rapid pressure fluctuation (defined as the carrier). The modulator and carrier represent amplitude modulation (AM) and frequency demodulation (FM), respectively. Some previous studies use modulation domain processing to filter out the unwanted modulation components for speech enhancement. For example, Paliwal *et al.* [6] applied a spectral subtraction algorithm to compensate the noisy modulation spectrum degraded by additive noise. Qin and Atlas [7] proposed a coherent modulation filtering approach using an instantaneous frequency estimation method. However, the methods proposed in [6-7] exist an inherent ambiguity in separating the modulator and the carrier. Accordingly, the assistive listening performance is inevitably limited

The human auditory system, comprising a peripheral auditory processing system and a cortex auditory processing system, is a highly sophisticated sensory system capable of picking up the target signals even in noisy environments [8]. In the peripheral auditory processing system, the cochlea in the inner ear decomposes the signal into multiple frequency bands through the basilar membrane [9]. To approximate the human auditory system response, Slaney [10] developed a cochlear model to represent the sound in the time-frequency domain. In the proposed model, the acoustic stimuli are processed through a cascade of logarithmic spacing gammatone filterbanks which mimic the cochlear filtering function in the auditory system. Li [11] proposed an auditory-based transform which converts the time-domain signals into a set of gammatone wavelet kernel functions, which simulate the basilar member motion response.

In accordance with the findings of the studies described above[8-11], this paper proposes a human-like auditory processing (HAP) algorithm for assistive listening device applied in low SNR and non-stationary noise environments. This system comprises both a peripheral auditory processing system and a cortical auditory processing system. The peripheral auditory processing system approximates the function of the human cochlear, while the cortical auditory processing system mimics the demodulation mechanism of the human auditory cortex. The experimental results show that HAP provides a better speech quality in noisy environments than current state-of-the-art speech enhancement algorithms.

2   Human-Like Auditory Processing Algorithm

Figure 1 presents a block diagram of the proposed HAP algorithm. As shown, the noisy speech signal is decomposed by a cochlear wavelet transform (CWT) module into sub-band signals, which are represented by a cochlear spectrogram for visualization purposes. The sub-band signals are then processed by the cortical

auditory processing system, which first demodulates the signals using an AM-FM demodulation (AFD) module and then filters the demodulated signals through modulation processing. Finally, the demodulated signals are synthesized by means of an inverse CWT operation in order to reconstruct the enhanced speech.
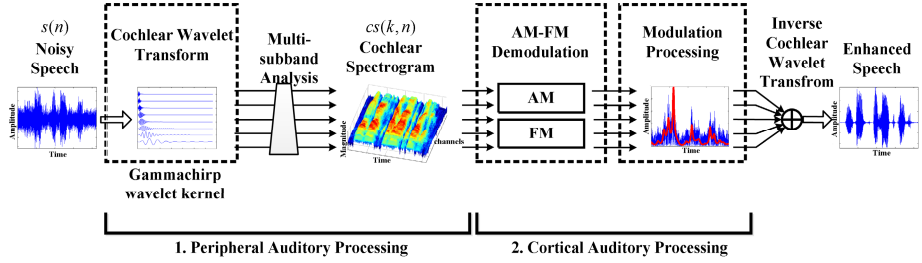
Fig. 1. Functional block diagram of the proposed human-like auditory processing algorithm

2.1 Peripheral Auditory Processing

The proposed CWT module extends the auditory transform [11] to mimic the mechanical response of the human cochlea using the complex-valued gammachirp function proposed by Irino [12]. The CWT module can be realized by convoluting the normalized input signal with an analyzed cochlear wavelet kernel function. The wavelet kernel function adopts the gammachirp function to characterize the level-dependent asymmetric shape of the auditory filter. The CWT module can be implemented as the form or convolution

$$\mathbf{CS}[k,n] = \psi_{a,b}[k,n] * s[n], \quad k = 0,1,...,K-1 , \tag{1}$$

where $\mathbf{CS}[k,n]$ is the complex-valued output of the discrete CWT, and the magnitude of the complex-valued output can be defined as the auditory spectrogram, a is the scale factor of each frequency bin k; b is the sample location ; K is the of number of frequency bin k; n is the sample index. The cochlear wavelet kernel function can be defined as follows:

$$\psi_{a,b}[k,n] = \frac{1}{\sqrt{a_k}} (\frac{b-n}{a_k})^{\alpha-1} \exp\left[-j2\pi\beta \cdot ERB(f_{center}) \cdot (\frac{b-n}{a_k})\right] \cdot$$
$$\exp\left[j2\pi f_{center}(\frac{b-n}{a_k}) + j\gamma\ln(\frac{b-n}{a_k}) + \phi\right]u(t) \tag{2}$$

where α and β are parameters for describing the envelope of the gamma distribution; f_{center} is the critical center frequency; γis the chirp term; ϕ is the initial phase; the scale factor a is associated with the central frequency of filters and b is the time location variable.

The term $ERB(f_{center})$ is the equivalent rectangular bandwidth varied as a function of the center frequency and defined as

$$ERB(f_{center}) = 24.7 + 0.108 \cdot f_{center} .$$

(3)

The frequency distribution of the scale factor a can be selected based on linear or the logarithmic scale. This reason for selecting ERB is because it performs a close estimate of auditory filter bandwidths. The magnitude of the CWT output can be represented as the cochlear spectrogram shown in Fig. 1. The cochlear spectrogram represents the time-frequency information to mimic the hearing content received by the ear [13]. Moreover, the acoustic features, such as pitch and spectro-temporal modulation frequency, can be further analyzed based on the cochlear spectrogram.

2.2   Cortical Auditory Processing

The cortical auditory processing system integrates the amplitude-frequency demodulation (AFD) modules and modulation processing to get the enhanced sub-band signals. Respectively, the AFD and the modulation processing approach the perception mechanisms and auditory attention and in the auditory cortex. The demodulation methods can be roughly categorized as inherent and coherent demodulation techniques. The incoherent approach splits the analytical signal into an envelope and a carrier by using the Hilbert transform. In contrast, the coherent approach first estimates the carrier of the signal and to calculate the modulator by multiplexing the conjugate carrier with the transmitted signal. This work used incoherent approach to extract the envelope of sub-band signal, namely sub-band AM signal. In the heavy noisy conditions, the sub-band AM signals are corrupted by the noise interference. The modulation processing technique is applied to extract the accurate AM signal and filter out the noise interference. The complex-valued output of CWT can be expressed as follow:

$$\begin{aligned}
\mathbf{CS}(k,n) &= cs(k,n) + jH\{cs(k,n)\} \\
&= \|\mathbf{CS}(k,n)\| \cdot \exp[j\varphi(k,n)]
\end{aligned}$$

(4)

where $cs(k,n)$ is the real part of complex-valued output of CWT; $H\{\cdot\}$ denotes the Hilbert transform; $\|\cdot\|$ denotes absolute value; φ is the phase of $\mathbf{CS}(k,n)$.

The AM sub-band signals can be represented as the magnitude of the cochlear spectrogram in each sub-band, shown as below

$$\mathbf{AM}(k,n) = \|\mathbf{CS}(k,n)\|,$$

(5)

In order to remove the high modulation components induced from the noise interference. The autoregressive model (AR) based mode, namely frequency domain linear prediction (FDLP), is used to extract the smooth envelope of the sub-band and reject the uncorrelated excitation. The FDLP can be defined as a technique for autoregressive modeling of the Hilbert envelopes in speech processing [14]. Furthermore, each AM sub-band signal can be passed through the zero phase shifting

low pass filter, the cut-off frequency was selected based the auditory temporal modulation characteristics [15]. The enhanced AM can be crossed the original carrier term to get the enhanced cochlear sub-band signals as follows:

$$\mathbf{CS}_{enh}(k,n) = \mathbf{AM}(k,n)_{enh} \cdot \exp[\,j\varphi(k,n)]\,. \tag{6}$$

Furthermore, the enhanced speech can be synthesized using the inverse CWT.

$$\tilde{s}[n] = \frac{1}{K}\sum_{k=1}^{K}\sum_{b=1}^{N}\frac{1}{\left|a_k\right|^{3/2}}\mathbf{CS}_{enh}[k,n]\cdot\psi_{a,b}[k,n]\,. \tag{7}$$

3   Results

The proposed HAP algorithm applied in assistive listening system was used to improve the speech quality and intelligibility. The performance of HAP algorithm was estimated using noisy speech sentences. 30 sentences spoken by three men and three women (5 sentences each) selected from the IEEE sentences corpus [16] were mixed with four different types of environmental noise (airport, bubble, car and street) taken from the AURORA database [17]. For each sentence and each environmental noise, the mixing process was performed using our different SNR levels of -5, 0, 5, and 10 dB. Thus, a total of 480 (i.e., 30 x 4 x 4) noise corrupted sentences were obtained. (Noted that the sample with SNR levels of 0, 5, 10 dB were extracted directly from the NOIZEUS database compiled by P. Loizo [18].)

The experiments commenced by investing the effect of the number of sub-bands in the CWT module on the quality of the reconstructed speech. The overall performance of the proposed speech enhancement algorithm was then evaluated by the quality of the processed noisy samples with the results obtained using three baseline methods, namely multi-band spectral subtraction (MBSS) [19], Hilbert envelope demodulation [7], and coherent modulation filtering [20].

3.1   Reconstruction Quality with Numbers of Sub-bands

The quality index was used to evaluate the performance of speech reconstruction. The subjective listening test, such as the Mean Opinion Scores (MOS), provides reliable results for assessing the speech quality. However, it is time consuming and requires the trained listeners. An objective quality evaluation method, the perceptual evaluation of speech quality (PESQ), was used to assess the quality of the speech in this paper. The PESQ has been used successfully employed for objective evaluation, and is one recommended by the ITU-T for speech quality assessments [18]. The PESQ score ranges from 1.0 to 4.5 for most cases. The high score of PESQ indicates a better quality, and score of 4.5 is defined as 100% of the quality score in this experiment. Fig. 5 showed relationship between the reconstruction quality score and the selection of sub-band numbers. In this circumstance, the IEEE speech sentences, the corpus of spontaneous Japanese [21] and self-recorded Mandarin Chinese corpus, are used to evaluation the reconstruction quality.
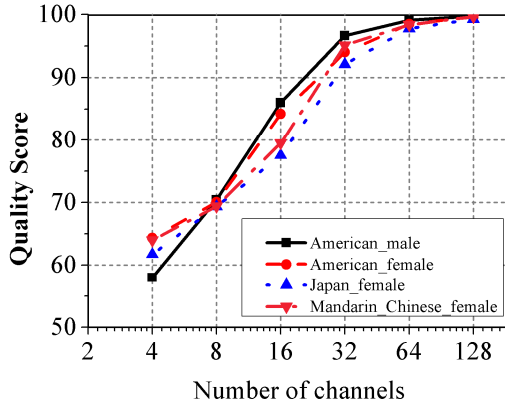
Fig. 2. The reconstruction quality is related with the selection numbers of analyzed sub-bands

3.2   Performance Evaluation of Assistive Listening Systems

The proposed approach for assistive listening system is compared with the baseline algorithms, which are the classical speech enhancement algorithm, the multi-band spectral substraction (MBSS) [19], and the demodulation based speech enhancement methods. The demdoulation based approaches can be sepratedly as inherent method, Hilbert envelope demodulation [7], and coherent method, coherent modulation filtering methos [20]. Fig. 3 demonstrates that applying the proposed algorithm can assist the enhancement of speech interfered by additive noise.
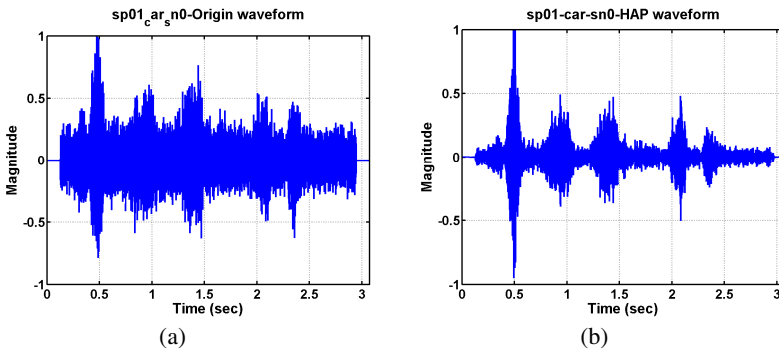
Fig. 3. The panels show (a) the waveform of noisy speech (b) the enhanced speech waveform processed by HAP algorithm

The quality evaluations of noise-corrupted speech were processed through four types of speech enhancement algorithms. Fig. 4 showed the average score of enhanced speech under four types of noisy conditions, such as that at airport, bubble, car and street with different SNR levels. The results show that the proposed HAP algorithm has comparative performance in quality.
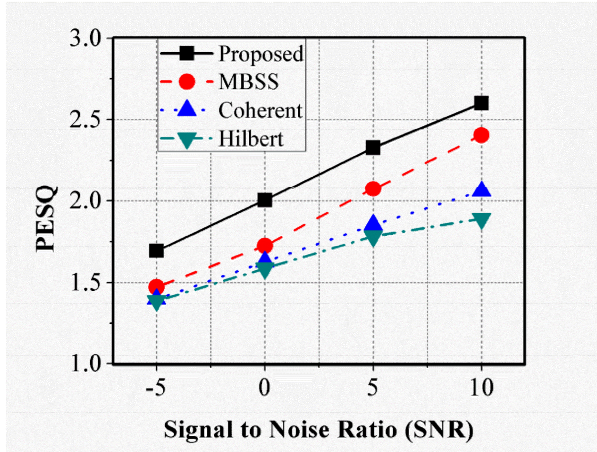
Fig. 4. Speech quality (PESQ) scores form the enhanced speech obtained from four kinds of speech enhancement algorithms at different SNRs

4   Discussion

The assistive listening device is to improve the communication and environment perception capability of hearing-impaired people. However, the classical speech enhancement algorithms, such as spectral subtraction [1] and a minimum mean-square error (MMSE) [2], usually degraded the quality and intelligibility of the speech and become even worse at low or negative signal-to-noise ratio (SNR) environments, especially at the SNR < 10 dB. The reason is because the classical speech enhancement algorithms focus on the estimation of noise by mathematical model. However, the noise is usually non-stationary, transient, and volatile with ambience. The classical speech enhancement processes usually produce the music noise, which is induced from the excessive suppression of noisy interference. However, the neurophysiological theory of auditory processing emphasizes the target sound source based on psychoacoustic characteristics instead of mathematic estimate [8, 15]. It means the human auditory system uses certain important strategies in processing acoustic signal.

The AM-FM representation is typically used for modeling the transmission information, such as speech, optics and radio. In human auditory processing, the perception of speech in noise is similarly to encode the AM-FM characteristics of the target source with noise interference. However, it is not easy to decompose the sound into AM-FM components perfectly because of the inherent ambiguity. Human ear decompose the sound into different frequency band using basilar membrane in the cochlea and discriminate speech from the non-speech sounds based on the perception of spectro-temporal modulations in the auditory cortex. It inspired us to construct a HAP algorithm to mimic the auditory perception mechanism of human ear. Based on the assumption, the proposed CWT module realizes the sub-band decomposition with an auditory frequency response, and increases the demodulation accuracy in noisy

conditions. The results in Fig. 2 also demonstrated that the CWT and its inverse term can reconstruction the speech well within 64 channel sub-bands. The modulation processing in the cortical auditory processing is used to extract the modulator and the carrier precisely. The modulator amplitude was calculated by the autoregressive model with long temporal segments (>200ms), which were selected because they fit the temporal modulation frequency range of the human voice [15]. This avoids the artificial distortion induced from the inaccurate phase information. Fig.4 demonstrates the proposed algorithm has better improvement performance at speech quality. This is because the amplitude envelope modification by using the proposed algorithm is useful for increasing the speech quality. However, the carrier still exists the noise components and non-continuous term. In order to evaluate the HAP algorithm in the real environment, we assembled an earphone with microphone and receiver which are connected with control box to execute the HAP algorithm. Based on the prototype, we can hear the enhanced sound through the earphone directly and distinguish the difference of the noisy speech and enhanced speech.
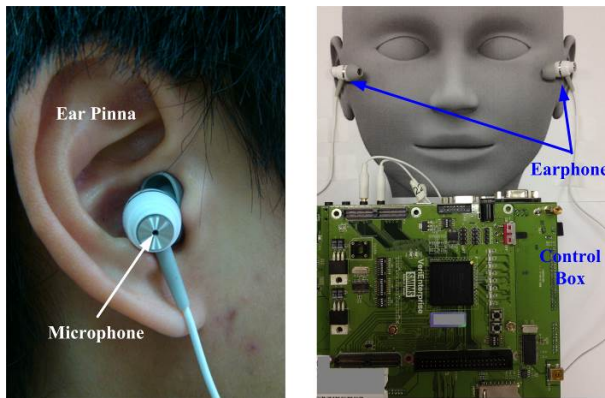
Fig. 5. The prototype of the proposed assistive listening device based on HAP algorithm

5   Conclusion

This work involved developing a human-like auditory processing approach for assistive listening system. To approximate the peripheral and cortical auditory processing in the human auditory system, this study proposed the cochlear wavelet transform, which simulates the cochlea in the inner ear to decompose the acoustic signal into sub-bands. The AFD and modulation processing were applied to encode the AM-FM characteristics of target source. Finally, the enhanced speech can be reconstructed through the proposed inverse cochlear wavelet transform. The objective evaluation results indicated the speech quality was improved by 16.9 % on average. When tested in the in the negative SNR environments (a SNR of 5dB), the proposed HAP algorithm continued to outperform the baselines. Future studies can further extend the monaural HAP algorithm to binaural HAP processing [36] for enhancing the human voice under heavy noisy conditions and reverberation environments.

References

1. Boll, S.F.: Suppression of Acoustic Noise in Speech Using Spectral Subtraction. IEEE Trans. Acoust. Speech, Signal Processing 27, 113–120 (1979)
2. Ephraim, Y., Malah, D.: Speech Enhancement using A Minimum-Mean Square Error Short-Time Spectral Amplitude Estimator. IEEE Trans. Acoust. Speech, Signal Processing 32, 1109–1121 (1984)
3. Ephraim, Y.: Statistical-Model-based Speech Enhancement Systems. Proc. IEEE 80, 1526–1555 (1992)
4. Hermus, K., Wambacq, P., Van Hamme, H.: A Review of Signal Subspace Speech Enhancement and its Application to Noise Robust Speech Recognition. EURASIP J. Adv. Sig. Pr. 2007, 045821 (2007)
5. Potamianos, P., Maragos, P.: Speech Analysis and Synthesis using an AM–FM Modulation Model. Speech Commun. 28, 195–209 (1999)
6. Paliwal, K., Wójcicki, K., Schwerin, B.: Single-Channel Speech Enhancement using Spectral Subtraction in the Short-Time Modulation Domain. Speech Commun. 52, 450–475 (2010)
7. Qin, L., Atlas, L.: Coherent Modulation Filtering for Speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 4481–4484. IEEE Press, Las Vegas (2008)
8. Khalfa, S., Bougeard, R., Morand, N., Veuillet, E., Isnard, J., Guenot, M., et al.: Evidence of Peripheral Auditory Activity Modulation by the Auditory Cortex in Humans. Neurosci. 104, 347–358 (2001)
9. Patterson, R.D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., Allerhand, M.: Complex Sounds and Auditory Images. In: Proc. 9th International Symposium on Hearing, pp. 429–446. Pergamon, Oxford (1992)
10. Slaney, M.: Lyon's Cochlear Model. Apple Technical Report (1988)
11. Qi, L.: An Auditory-based Transform for Audio Signal Processing. In: IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, pp. 181–184. IEEE Press, New York (2009)
12. Irino, T., Patterson, R.D.: A Dynamic Compressive Gammachirp Auditory Filterbank. IEEE Trans. Audio, Speech, Language Process. 14, 2222–2232 (2006)
13. Sung, P.H., Thompson, W.R., Wang, J.N., Wang, J.F., Jang, L.S.: Computer-assisted Auscultation: Patent Ductus Arteriosus Detection based on Auditory Time Frequency Analysis. J. Med. Biol. Eng. (2014), doi:10.5405/jmbe.1689
14. Athineos, M., Ellis, D.P.W.: Frequency-Domain Linear Prediction for Temporal Features. In: IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE Press, St. Thomas (2003)
15. Woolley, S.M.N., Fremouw, T.E., Hsu, A., Theunissen, F.E.: Tuning for Spectro-Temporal Modulations as a Mechanism for Auditory Discrimination of Natural Sounds. Nat. Neurosci. 8, 1371–1379 (2005)
16. IEEE. IEEE Recommended Practice for Speech Quality Measurements. IEEE Trans. Audio Electroacoust., 225–246 (1969)

17. Pearce, D., Hirsch, H.G.: The Aurora Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions. In: Proc. ISCA ITRW ASR, Paris, France (2000)
18. Yi, H., Loizou, P.C.: Evaluation of Objective Quality Measures for Speech Enhancement. IEEE Trans. Audio, Speech, Lang. Processing 16, 229–238 (2008)
19. Kamath, S., Loizou, P.: A Multi-Band Spectral Subtraction Method for Enhancing Speech Corrupted by Colored Noise. In: IEEE International Conference on Acoustics, Speech and Signal Processing, p. IV-4164. IEEE Press, Orlando (2002)
20. Schimmel, S., Atlas, L.: Coherent Envelope Detection for Modulation Filtering of Speech. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 221–224. IEEE Press, Philadelphia (2005)
21. The Corpus of Spontaneous Japanese,
    http://www.ninjal.ac.jp/english/products/csj/

An Adaptive Harmony Search Algorithm with Zipf Distribution

Shih-Pang Tseng and Jaw-Shyang Wu

Department of Computer Science and Entertainment Technology,
Tajen University, Pingtung, Taiwan
{tsp,jswu}@tajen.edu.tw

Abstract. Harmony search (HS) can be applied to various optimization problems and easy to implement. In this paper, we try to improve HS by change the reference probability distribution of harmony memory. Zipf distribution is used to balance the intensification and diversification. In addition, we propose the adaptive mechanism to avoid setting the new parameter. Experimental results show that the improvement is effective on the high dimensional numerical function optimization problem.

1 Introduction

Optimization trys to find the best solution within an allowed solutions set to achieve some objectives for many theoretical and practical problems. These optimization problems can be categorized into two kinds: continuous and discrete. In the continuous optimization problem, the solutions can be encoded with real-valued variables. Otherwise, the solutions can be encoded with the discrete variables in the discrete optimization problems. Whether the continuous or discrete ones, The allowed solutions set, usually denoted by *search space* contains a great number of possible solutions, it is not practical to explore the whole search space for many problems. For these optimization, to find the optimal solution in the search space is impractical by an affordable computing cost. Metaheuristics [1] are developed for these kinds of optimization problems. A metaheuristic is an approximate algorithm to explore the search space as efficiently and effectively as possible. A metaheuristic try to find the sub-optimal, or near-optimal solution by an affordable computing cost. In the past 30 years, there were many metaheuristics proposed, such as genetic algorithms [2], ant system[3], and so on. In the first ten years of 21 century, a new metaheuristic, Harmony search (HS) [4], is proposed and successfully applied in theoretical and physical domains, such as solving Sudoku [5], cell-phone network [6], Ground Water Modeling [7] clustering web documents [8], and so on. Harmony search is inspired by the process of music to search for a perfect state of harmony.

In HS, the solution of this problem is call by the *harmony*. The harmony memory(HM) is used to store a set of harmonies. HS constructs new harmony by referring these harmonies in harmony memory. Only one new harmony (solution) in one iteration of HS. If the new harmony is better than the worst one in the HM, the new one would be inserted into the HM to replace the worst one.

A harmony is coded to a vector of attributes in HS. In constructing new harmony, for each attribute of harmony, there are three possible choices to decide its value, as following:

1. **Usage of harmony memory**: pick the attribute value from the harmony selected randomly in HM.
2. **Pitch adjusting**: like the above, but add a little change to the attribute value.
3. **Randomization**: randomly assign a value to the attribute.

The usage of harmony memory ensures that good harmonies are considered as elements of new harmony. In order to use HM effectively, a parameter $r_{accept} \in [0,1]$, called harmony memory considering (or accepting) rate, is used to define the probability of usage of harmony memory in constructing a new harmony. The pitch adjusting makes a little change to the attribute's value from harmony memory. The parameter $r_{pa} \in [0,1]$ is called pitch rate which defines the probability of pitch adjusting after the usage of harmony memory. This try to tune the value to find a better solution. The randomization increases the diversification of the solutions to avoid premature optimization.

1.1   Motivation

The balance between intensification and diversification [1] is the most important to all metaheuristics. In general, the diversification can prevent the premature optimization and get better final solution. Moreover, the diversification usually reduces the converging speed and needs more computation time. The randomization of HS can be used to increase the diversification. However, the probability of randomization, $1 - r_{accept}$, is a sensible parameter of HS. It should be small enough to ensure the convergence. For this reason, it is a better way to increase the size of harmony memory, and another way should be involved to enhance the intensification.

So we can increase the size of harmony memory to handle the diversity of HS. And we do not consider the intensification factor to the size of harmony memory. It makes the size of harmony memory not a sensitive parameter. Thus we can use a larger size of harmony memory to keep more solutions in the process of harmony search. In original HS, all harmonies in harmony memory are with the same reference probability in constructing a new harmony. The distribution of reference probabilities is an uniform distribution. In this paper, we redesign the structure of harmony memory. And the distribution of reference probabilities is changed to Zipf distribution. Consequently, the better harmony is with more probability, and the worse harmony is with less probability. In addition, the difference of reference probabilities between better and worse harmonies can be clearly control by a single parameter z.

It is not easy to decide the algorithm parameters values of the metaheuristics.[9] For the balance between intensification and diversification, the Zipf distribution and one new parameter z are involved in HS. In our previous work, ZHS [10], we

applied various z on different numeric optimization functions. It shows that ZHS is beter than HS on the balance between intensification and diversification. However, it is still a shortcoming to find a suitable value of z for different functions in ZHS. It is the best way to set the value of z adaptively. It means that the z would be encoded as one part of the harmony, and the value of z should be adjusted for different functions.

In this paper, a Adaptive Zipf harmony search (AZHS) is proposed to more effectively and efficiently solve the numerical function optimization problem. The remainder of the paper is organized as follows. Section 2 briefly describe the Zipf distribution. Section 3 provides a detailed description of AZHS we proposed. Performance evaluation of the AZHS is presented in Section 4. Conclusion is given in Section 5.

2   Zipf Distribution

Zipf distribution is derived by the Zipf's law which originally states the skewness of natural language [11][12]. George Kingsley Zipf proposed the fact that the frequency of one word is proportional to the power inverse of its rank. The similar relationship is discovered in some economic domains, such as the income distribution. In recent years, the Zipf distribution is used to describe the assess behavior of Internet, such as the visits to web sites[13]. In this work, the Zipf distribution is used to control the reference behavior to harmony memory, and to balance between intensification and diversification.

Zipf distribution is a discrete probability distribution. N denotes the number of elements. k_i is the rank of element i, k_i is an integer and $1 \leq k_i \leq N$. The important parameter of Zipf distribution is z which can adjust the inequality. z is a real number and $z \geq 0$. The probability of the element i, $P(i)$, is proportional to the z power inverse of its rank, k_i.

$$P(i) \propto \frac{1}{k_i{}^z} \tag{1}$$

$$P(i) = \frac{k_i{}^{-z}}{\sum_{j=1}^{N} k_j{}^{-z}} \tag{2}$$

Figure 1 shows the probabilities of Zipf distribution with different z values. When $z = 0$, the Zipf distribution is equivalent to the uniform distribution in this special case. When z is larger, the difference between the first one and the last one is larger.

3   Proposed Method

In traditional HS, harmony memory is only a set of harmonies. All harmonies in harmony memory are with the same reference probability in constructing a new
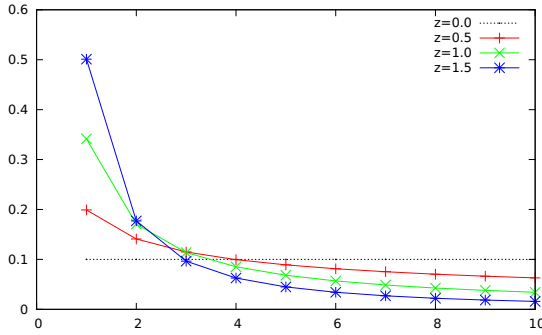
Fig. 1. Zipf distribution with different z values

harmony. On the other hand, the harmony memory is organized to a sorted list in AZHS. The best harmony is in the head of sorted list, and the worst one is in the tail. The main difference between ZHS and traditional HS is the reference probability distribution to the harmonies in harmony memory. The reference probabilities of harmonies in harmony memory is related to their position, rank, in this sorted list. The reference probability of each harmony is proportional to the z power inverse of its rank. When a new harmony has been constructed, it would be inserted into the appropriate position in the sorted list if it is better than the worst one. And then the worst harmony should be removed from the harmony memory. HS can be considered as a special case of AZHS, when z is always zero. Figure 2 shows the structure of Zipf harmony memory which size is 10.

h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}
$p_1 = 0.34$	$p_2 = 0.17$	$p_3 = 0.11$	$p_4 = 0.085$	$p_5 = 0.068$	$p_6 = 0.056$	$p_7 = 0.048$	$p_8 = 0.043$	$p_9 = 0.037$	$p_{10} = 0.034$

Fig. 2. An example of Zipf Harmony memory, $size = 10$

Algorithm 1 shows the detail of AZHS. In the beginning to construct a new harmony, a new z would be assigned for the following steps. The assignment of z is similar to the assignment of attribute values in HS. There are three ways to assign z value.

1. **Usage of harmony memory**: pick the z value from the harmony selected randomly in HM.
2. **Pitch adjusting**: like the above, but add a little change to the z value.
3. **Randomization**: randomly assign a value to z.

In this step, the reference probability distribution of HM is uniform, due to the z is not assigned. After thr z is assigned, the attributes of new harmony would be assigned in sequence according the z value. If the z value is larger, the constructing process would prefer the better harmonies in HM. If the z value is smaller, the reference probability distribution of HM is more uniform in the constructing process.

Algorithm 1. Adaptive Zipf Harmony Search

Initiate Harmony Memory
for $i = 1$ to $max\,number\,of\,iterations$ **do**
  if $rand() < r_{accept}$ **then**
    Choose a z value from Harmony Memory
    if $rand() < r_{pa}$ **then**
      Adjust the z value
    end if
  else
    Assign a random z value
  end if
  for $j = 1$ to $number\,of\,dimensions$ **do**
    if $rand() < r_{accept}$ **then**
      Choose a value from Harmony Memory for this attribute
      if $rand() < r_{pa}$ **then**
        Adjust the value
      end if
    else
      Assign a random value to the attribute
    end if
  end for
  Insert new harmony into Harmony Memory
end for
Return the best harmony

---

4 Experiment Result

In this section, we evaluate the performance of the proposed algorithm, AZHS and compare with original HS. The empirical analysis was conducted on an virtual machine in CHT Hicloud(http://hicloud.hinet.net/) with 4 CPUs and 8GB of memory using CentOS 6.4 running Linux kernel 2.6.32. Moreover, all the programs are written in C++ and compiled using g++ (GNU C++ compiler).

In this research, we choose five benchmarks of numerical function optimization, F1 $\sim$ F5. These benchmark functions are well-known and used in many researches. The function F1 is the De Jongs function 1, sphere function. Next, the function F2 is the axis parallel hyper-ellipsoid function. The function F3 is the rotated hyper-ellipsoid function The function F4 is the De Jongs function 2, Rosenbrocks valley. The Rastrigins function is the function F5. The F1, F2

and F3 are unimodal; and the F4 and F5 are multimodal. The formulae and ranges of benchmark functions are shown in Table 1. All these benchmarks are the minimization problems.

In our experiments, all the simulations are carried out for 30 runs. The r_{accept} is always set to 0.9 . The probability of pitch adjusting, r_{pa}, is 0.3 . The Zipf distribution parameter z is varied from 0 to 10.0. We choose two different harmony memory sizes, 50 and 100.

Table 1. BENCHMARKS FOR NUMERICAL FUNCTION OPTIMIZATION

	Functions	Range
F1	$f_1(\mathbf{x}) = \sum_{i=1}^{n} x_i{}^2$	$-100 \leq x_i \leq 100$
F2	$f_2(\mathbf{x}) = \sum_{i=1}^{n} ix_i{}^2$	$-5.12 \leq x_i \leq 5.12$
F3	$f_3(\mathbf{x}) = \sum_{i=1}^{n} \sum_{j=1}^{i} x_j{}^2$	$-65.536 \leq x_i \leq 65.536$
F4	$f_4(\mathbf{x}) = \sum_{i=1}^{n-1} 100(x_{i+1} - x_i{}^2)^2 + (x_i - 1)^2$	$-2.048 \leq x_i \leq 2.048$
F5	$f_5(\mathbf{x}) = 10n + \sum_{i=1}^{n} x_i{}^2 - 10\cos 2\pi x_i$	$-100 \leq x_i \leq 100$

To simplify the discussion of the simulation results, we use the following conventions. Let v denote either the value of final solution, Δ_v the enhancement of v_ϕ with respect to v_ψ

in percentage. In other words, Δ_v is defined as follows:

$$\Delta_v = \frac{v_\phi - v_\psi}{v_\psi} \times 100\% \tag{3}$$

where v is either the value of final solution for our experiments on numerical function optimization problems. The v_ϕ and v_ψ are the result of two different algorithms. For example, the HS and the AZHS

Table 2 shows the comparison of HS and AZHS when the dimensions of benchmarks is set to 50, and the evaluations is set to 100000. In general, the AZHS is apparently better than the HS, except one. Because all benchmark are minimization problem, the negative Δ_v means the AZHS is better. The z^* denotes the average z value of the best harmony in different tries. Table 3 shows the experimental result when the dimensions of benchmarks is set to 50, and the evaluations is set to 100000. In general, the AZHS is significantly better than the HS in all cases.

Table 2. Experimental results of 30 dimensions with 100000 evaluations

	HM	HS	AZHS	Δ_v	z^*
F1	50	0.001039	0.000599	-42	5.91
F2		0.0164	0.0104	-37	5.92
F3		2.53	1.66	-35	5.97
F4		56.0	35.1	-37	6.90
F5		0.277	0.208	-25	6.03
F1	100	0.00417	0.00057	-86	6.28
F2		0.0562	0.0098	-82	5.92
F3		9.51	1.59	-83	5.19
F4		41.5	46.9	13	5.64
F5		0.562	0.135	-76	6.12

Table 3. Experimental results of 60 dimensions with 500000 evaluations

	HM	HS	AZHS	Δ_v	z^*
F1	50	0.549	0.016	-97	6.17
F2		11.3	0.576	-94	6.12
F3		1956.2	91.6	-95	6.70
F4		152.6	134.9	-11	6.04
F5		20.5	14.0	-31	5.81
F1	100	0.925	0.017	-98	5.85
F2		19.080	0.547	-97	5.53
F3		3068.4	85.0	-97	5.66
F4		136.5	134.8	-1	5.42
F5		21.8	15.0	-31	6.24

5   Conclusion

Harmony search (HS) is a relatively new metaheuristic in the past 10 years. We proposed the Adaptive Zipf harmony search (AZHS) which used the Zipf distribution to be the reference probability distribution of harmony memory. In addition, we involve the adaptive mechanism to set the z parameter automatically. It makes the better balance between intensification and diversification. We applied the AZHS on the numerical function optimization problem. In elementary, the experimental result shows that AZHS is a significantly enhancement of HS.

References

1. Blum, C., Roli, A.: Metaheuristics in combinatorial optimization: Overview and conceptual comparison. ACM Comput. Surv. 35(3), 268–308 (2003)
2. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc. (1989)
3. Dorigo, M., Maniezzo, V., Colorni, A.: The ant system: Optimization by a colony of cooperating agents. IEEE Transactions on Systems, Man, and Cybernetics-Part B 26, 29–41 (1996)
4. Geem, Z.W., Kim, J.H., Loganathan, G.V.: A new heuristic optimization algorithm: Harmony search. Simulation 76(2), 60–68 (2001)
5. Geem, Z.W.: Harmony search algorithm for solving sudoku. In: Apolloni, B., Howlett, R.J., Jain, L. (eds.) KES 2007, Part I. LNCS (LNAI), vol. 4692, pp. 371–378. Springer, Heidelberg (2007)
6. Zhang, R., Hanzo, L.: Iterative multiuser detection and channel decoding for ds-cdma using harmony search. Signal Processing Letters 16, 917–920 (2009)
7. Ayvaz, M.T.: Simultaneous determination of aquifer parameters and zone structures with fuzzy c-means clustering and meta-heuristic harmony search algorithm. Advances in Water Resources 30, 2326–2338 (2007)
8. Mahdavi, M., Chehreghani, M.H., Abolhassani, H., Forsati, R.: Novel meta-heuristic algorithms for clustering web documents. Applied Mathematics and Computation 201(1-2), 441–451 (2008)
9. Geem, Z.W., Sim, K.B.: Parameter-setting-free harmony search algorithm. Applied Mathematics and Computation 217(8), 3881–3889 (2010)
10. Tseng, S.P., Lin, W.W.: Improving harmony search by zipf distribution. In: 2012 Sixth International Conference on Genetic and Evolutionary Computing (ICGEC), pp. 115–118 (August 2012)
11. Zipf, G.K.: Selected Studies of the Principle of Relative Frequency in Language. Harvard University Press (1932)
12. Zipf, G.K.: Human Behavior and the Principle of Least Effort. Addison-Wesley, Reading (1949)
13. Adamic, L.A., Huberman, B.A.: Zipf's law and the Internet. Glottometrics 3, 143–150 (2002)

LED Lighting Applications for Digital Life

Lih Wen Hwang

Department of Computer Science and Information Engineering,
Tajen University, Pingtung, Taiwan
`lwhwang@tajen.edu.tw`

Abstract. LED lighting based on solid state IC technology works seamlessly with computer and/or computer network. The integration of LED lighting with microprocessor (MCU or PSoC) at system levels proliferate many new applications toward a digital life style. Meanwhile, due to its long life span, compact-ness and efficient energy consumption LED lightings are environmental friend-ly. In this article we introduce four new applications and show how LED light-ing may advance in future "green" digital life.

Keywords: LED, digital life, LED lighting systems and applications.

1 Introduction

LED (light-emitting diodes) has long played a role in consumer electronics as an indicator. It is now emerging as a potential technology for everyday commercial and residential lighting applications. As incandescent light bulbs are phased out because they use too much energy, and concerns over mercury affect sales of compact fluorescent bulbs, LED lighting application is receiving increasing attention from consumers and governments seeking to reduce energy costs and lower their carbon footprint[1][2][3].

While LEDs themselves are not made with silicon, silicon chips are an integral component of its fixtures. To be driven properly, modern high power LED lighting fixtures are armed with MCU (micro control unit) or PSoC (programmable system on chip) for added intelligence, functionality and integration [1]. This represents a significant market opportunity for chip makers and the use of MCU or PSoC for LED lighting creates a frontier for engineering design and for new applications.

Taking advantage of the chip's user interface, communication, battery status monitoring and etc., the followings show uses for a MCU or PSoC in smart LED lighting fixtures for four examples of different applications and perceive the intersection of energy, automation and the digital lifestyle.

2 LED Lighting Applications

2.1 Intelligent LED Lighting System

Enabled by smart LED fixtures, the following two examples show intelligent lighting systems can combine software and networking technology to tackle large collaborated networking applications to save energy and cost.

A. LED lighting system manufacturers create a mesh network of fixtures which can be managed by software. Each lighting fixture is equipped with an MCU or PSoC, Zigbee wireless networking chip, and sensors for detecting light levels and when a person enters a room. Test results show with warehouse-like space the system has cut costs by 90 percent over HID lights.

B. Other than being a super energy efficient way to light up a room, LED lighting systems are also a way to connect to the internet. Scientists from Germany's Fraunhofer Institute have devised a way to encode a visible-frequency wireless signal in light from our plain old desk lamps and other light fixtures with an MCU or PSoC. In the near future, jumping on the internet might be as simple as flipping on your light switch.

2.2   Solar Powered LED Lighting System

Using solar energy in the remote areas is crucial, for it not only saves energy, but solves the problem of inconvenience in wiring the power grid. Taking advantage of solar energy's merits that is clean, environment protection and inexhaustibility and the LED's merits that are high efficiency, long life and haze-penetrating capability, LED lighting will be a perfect match with solar cell.

In the IEEE ICIEA 2009, Zhaohui and etc. [2] proposed LED lighting controller that adopts a PSoC and improve the traditional perturbation and observation (P&O) algorithm. The same LED lighting controller also regulating the LED brightness with environmental light and extends the lifetime of LED.

The essence of the controller is a negative feedback system. The controller can change the objective value according to the environmental light intensity, fulfilling the requirement of saving energy. Meanwhile, the controllers save the parameters in the EEPROM and communicate with the computer and the network.

2.3   Smart Greenhouse LED Lighting System

LED can be used to grow plant in hydroponics' agriculture, greenhouse lighting, herb gardens and home gardens. Well calculated LED lighting brings the perfect balance of light to supplement and control the plants photosynthesis. Three light properties are especially important when working with plant: amount, direction, and spectral quality. All three parameters can be adjusted automatically through MCU of LED driver.

LED provides the solution to the problems associated with traditional greenhouse lighting. Conventional grow lights use up to 1,000 watts of power to operate and must be replaced every year. Each LED light uses less than 9 watts of power and is rated for 100,000 hours of life (12+ years). Over the lifetime of the LED significant savings occur in electricity and replacement costs alone. This energy savings be-comes more and more important as energy costs continue to rise.

For commercial growing of plant or vegetables where the cost is the major concern, one might create the LED lighting recipe for the control of light intensity, spectra, and duration for minimum energy requirements. Overall, LED lights for Plants have the following advantages that can result in lower operating costs:

A. LED color can be fine-tuned to the spectra most useful to plants. Using only the required spectra will increase energy efficiency and reduce operating costs.

B. Long life time when compare with other light sources.

C. Work with low voltage, reducing the costs of high voltage wiring and protection systems;

D. Dimmable (0% ~ 100%) to adapt to plant life stages to further save electric energy;

E. Lack of radiant heat, which results in a cooler environment near the plants.

F. Can be paired with secondary optics in the form of either lenses or reflectors that direct the light only to those areas needed, resulting in a higher degree of radiation utilization and a uniform intensity.

G. Supplying only the plant required spectra leaves out some spectra required by insects for reproduction, supplying a barrier to pests.

2.4   UV-LED Lighting System

While white LED for lighting and LCD displays applications are undergoing rapid technological advancing due to market demanding, UV-LEDs are benefit from these fast developments of these trends, because UV and white LEDs are technically similar. In principle, LEDs can be made at any wavelength from 250 nm (UV) to 570 nm (greenish yellow) can be manufactured by adjusting the semiconductor composition.

Unlike white LEDs' lighting impacts, the advancement of UV LEDs is offering a wide range of advantages to a diverse number of applications. Dental curing instruments and counterfeit detection applications were early adopters of UV LED technol-ogy. The performance, cost and durability benefits combined with recent enhance-ments in life span are causing UV LEDs to be integrated into a rapidly growing num-ber of niche applications.

If UV-LED efficiency keeps on improving to the levels of white light LEDs, they would be more than twice as efficient as mercury lamps and offer a wider range of

UV wavelengths. The followings are a list of applications classified by their wavelengths ranges [4]:

A. UV-A (315~400 nm): polymer curing, ink printing, counterfeit detection and superficial/cosmetic sterilization,

B. UV-B (280~315 nm): forensic and bodily fluid detection and analysis, protein analysis, drug discovery and medical light therapy,

C. UV-C (100-280 nm): sterilization of surface areas and water, UV ID verifica-tion, barcodes.

2.5   Discussion

The objective of LED lighting system is to reduce the maintenance cost over the life of the LED lighting product and substantially reduce the carbon foot print associated with lighting in addition to use less power for better light quality.

The ultimate design goal for a smart LED lighting system shall include the following 3Cs:

- CO_2: Reduce CO_2 footprint and comply with standard
- Cost: Reduce total operation cost (TOC) and return on of investment (ROI)
- Color: Exploring LED lighting "color" and quality factors and implement innova-tive illumination application.

To do so, a smart LED lighting system, for example, may include control technologies of occupancy sensing, scheduling, tuning, daylight harvesting, and demand response to reduce peak power charges etc. Fig.1. shows a grid-connected solar power smart LED lighting system [5] [6] includes sensors and IT software solutions for controlling.
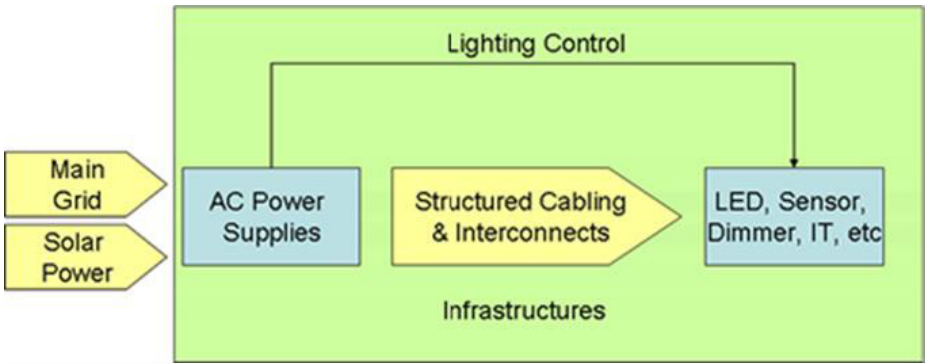
Fig. 1. Lighting system control

As IC integrated into a LED subsystem, LED lighting fixture can be programed and therefore communicate with other control devices to form a networked technology.

Further IT software solutions shall include, such as energy analysis and audit, energy management, project design and implementation, maintenance and operation, monitoring and evaluation of saving, contracts management and equipment supply. Fig.2. shows an energy management system (EMS) which is an information configuration software system with three major function requirements:

- Define, program, modify, and update lighting control system database;
- Windows based, capable of running on either central server or a remote client over TCP/IP connection;
- Allow user to define following within desired space: Zones, Scenes, Sub-spaces, Time-clock, Control station devices, Switching panels, Switch legs.
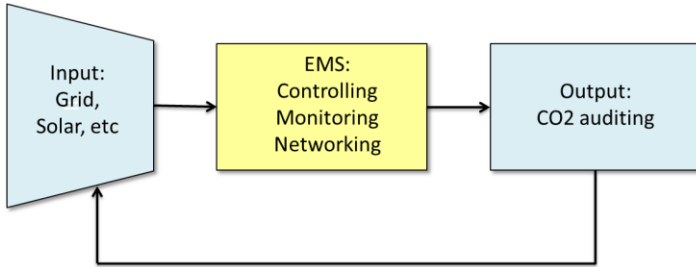
Fig. 2. Energy management system

2.6   Conclusion

Recent LED technological advancements (including higher efficiency, higher output power and lower cost) are moving the LED market segment into a whole new level of lighting product innovation and performance [7].

From designers' perspectives, LED lighting as light sources can go beyond the replacement of traditional light source [3], by providing new capabilities including the control of spectrum, color temperature, polarization, and temporal modulation. We can see in the future "smart" LED light sources act more like a computer, can be programmed and can be as a node in a network to offer multi-dimensional controllability that will enhance the functionality and performance of light sources in a wide range of applications.

References

1. Microchip: Adding Intelligence to Lighting Applications, LED Lighting Design Guide (September 2008)
2. Jiang, H., Wang, Y., Zhang, H., Yuan, Z.: The High-lightness LED Controller Using Solar Energy Basing on PSoC. In: IEEE Conference, The Proceedings of ICIEA 2009 (2009)
3. Kim, J.K., Schubert, E.F.: Transcending the replacement paradigm of solid-state lighting. Optics Express 16(26), 21835–21842 (2008)
4. Halliday, R.: Key benefits of next-gen UV LED technology, `http://www.lumex.com`
5. Zhou, Y., Narendran, N.: Photovoltaic-powered light-emitting diode lighting systems. Optical Engineering 44(11), 111311 (2005)
6. Sinha, A.K., Samantaray, S.K., et al.: Cost benefit analysis of solar powered LED based lighting system. In: Third International Conference on Power Systems, India (December 2009)
7. Navigant: Adoption of Light-Emitting Diodes in Common Lighting Applications. U.S. Dept. of Energy, `http://apps1.eere.energy.gov`

The Exclusive Challenge in Pervasive Learning Technology – Points from the Tiger

Tzong-Song Wang[1] and Yun-Chung Lin[2]

[1] Institute of Cultural and Creative Industries, Ping-tung, Taiwan, R.O.C.
`tswang@tajen.edu.tw`
[2] Department of MIS, Tajen Institute of Technology, Ping-tung, Taiwan, R.O.C.
`yuchung@tajen.edu.tw`

> It is not enough that you should understand about applied science in order that your work may increase man's blessings. Concern for man himself and his fate must always form the chief interest of all technical endeavors, concern for the great-unsolved problems of organization of labor and the distribution of goods -- in order that the creations of our mind shall be a blessing and not a curse to mankind. Never forget this in the midst of your diagrams and equations.
>
> Albert Einstein
> from an address at Cal Tech (1931)

Universal computing is a global goal worthy of the most serious consideration of engineering, e-commerce, and educational leaders. For the benefits of the information age to effect general humanity, it is essential that computing become more widely embraced across the diverse groups of society, to confront the increasing problem of the digital divide. Currently, disenfranchisement, disinterest, and overt contempt for technology still abounds (Macionis, 2001). Innovators in engineering and e-commerce need to devote some of their creative energies to systematically making the "universal computing society" a practical reality rather than a myth. The markets for e-commerce, let alone the benefits of telecommunication for humanity, require continuing energy to educate the majority of society about tools and visions for change.

Research on computer and web usage, even in highly industrialized regions of the world, continue to show that much work remains to be done to bring computer access to many groups of people, let alone to achieve universality, pervasiveness, or ubiquitousness (The World Bank, 2000). Societies of engineers, educational leaders, and nations need to invent new ways to work together to include groups of people who either remain disinterested or disenfranchised about the universal computing paradigm.

Some engineering and e-commerce leaders continue to market the power the evolving cyberworld. Cyberchange leaders have continued to endorse enthusiastically

the vision of a world of computing that affects everyone, whether they like or not, or even if they are not aware of it. The concept that a full motion digitized video technology can bring the world's people into an interactive, peaceful and productive whole is a marvelous vision, however.

As university professors in education and technology, the writers have enjoyed the benefits (one in the US and the other in Taiwan) of upper middle class access to computers and servers. Advanced degrees and the support of continuing education can provide a means to make the vision of a world of universal and ubiquitous computing a reality. But, the writers have noticed with soul-searching concern that their own daughters, let alone friends of "less economically dominant" perspectives, ages, or races do not seem to either know about the brave new positive future, or that they just don't see evidence that this vision will apply to them. Everyone is not engaged in e-commerce or universal computing.

1   Facing the Disenfranchised in the Universal Computing Vision

Harasim, Hiltz, Teles, and Turoff (1995) indicated that the Internet has been able to bridge the constraints and to break the barriers of space and time, enhancing learning without walls and just in-time learning. In addition, the connection through computer networks has made it convenient to reach different people and different information networks in the connected outside world. The writers are concerned that their daughters and friends are not in either overt or covert ways disenfranchised from participating freely, voluntarily, or enthusiastically in the body politic that seems to be in charge of the pace of cyberchange. However, Turkle (1988) found that women continue to be less receptive to use of computing and online resources than men, providing the writers with concern that their daughters me face disenfranchisement in the future.

The vision of a world in which text, audio, and video are unoppressively available—i.e., pervasive—to provide users with the knowledge, experience, and wisdom for full participation in a democratic global society holds much promise. But, it is critically important for advocates of increasing the pace of cyberlearning to continue to realize that many children, seniors, rural or urban people, and members of less advantaged groups are yet disenfranchised or very disinterested, if not hostile.

Russell described the pervasive computing trend as moving towards increasingly ubiquitous connected computing devices in the environment, particularly through wireless technologies and the Internet. His view of pervasive computing devices has been a different paradigm than current perception of personal computers. Russell conceived of a world of ubiquitous as follows:

> …"very tiny—even invisible—devices, either mobile or embedded in almost any type of object imaginable, including cars, tools, appliances, clothing and various consumer goods - all communicating through increasingly interconnected networks. According to Dan Russell, director of the User Sciences and Experience Group at IBM's Almaden Research Center, by 2010 computing will have become so naturalized within the environment that people will not even realize that they are using computers. Russell and other researchers

expect that in the future *smart* devices all around us will maintain current information about their locations, the contexts in which they are being used, and relevant data about the users. The goal of researchers is to create a system that is pervasively and unobtrusively embedded in the environment, completely connected, intuitive, effortlessly portable, and constantly available. Among the emerging technologies expected to prevail in the pervasive computing environment of the future are wearable, smart homes and smart buildings. Among the myriad of tools expected to support these are: application-specific integrated circuitry (ASIC); speech recognition; gesture recognition; system on a chip (SoC); perceptive interfaces; smart matter; flexible transistors; reconfigurable processors; field programmable logic gates (FPLG); and microelectromechanical systems (MEMS). (p. 1)

Other leading technological organizations are exploring pervasive computing. Xerox's Palo Alto Research Center (PARC), IBM's project Planet Blue, Carnegie Mellon University's Human Computer Interaction Institute (HCII), and the Massachusetts Institute of Technology (MIT) have projects that envision a future of ubiquitous computing devices as freely available and easily accessible as the air to be breathed (Russell, 2001).

Potential for micronized devises to be implanted in animals or people, whether like products to be identified with electronic bar codes, or, even more complex, like miniaturized servers to free users from the need to be connected to the control of centralized servers, can raise significant concern people who may have been in the past only casually interested in computing (Russell, 2002). Who gets the opportunity for having these marvelous devices implanted? Who pays for the option? Who is not important enough to be included? How can social justice and equity issues be assured in knowing that the wealthy and informed will be the first recipients?

Computers have become valuable tools people use to access information and communicate with others around the world. For many people in the industrialized world, computers are everywhere, and virtual communities may becoming pervasive, but far from universal. The landscape for connecting people has been changing as technology advances and computers extend into most every facet of daily living in the advantaged nations. Some level of computer literacy is requisite for most professional leaders to be successful in today's world of competing economic and international relationships (Craig, 2002).

The World Bank (2000) reported that personal computers became commonplace in high income countries, including the US, the nations of Western Europe, Japan, and Australia. But, by contrast, personal computers were rare in the poorest countries of the world, particularly in Asia and Africa. Peterson, Wunder, and Mueller (1999) made the following comments about the global economy and capitalism:

The tendency of modern capitalism to pursue short-term gain to the long-term detriment of the environment, workers, and economic stability has largely been controlled in the First World through governmental regulation. Each of the industrial capitalistic nations of the First World has learned through experience to control the most dangerous for of speculation and the most questionable financial practices. Elsewhere, however, much of the global economy is out of control. In Russia and the Second World nations, the retreat from socialism and

the move toward capitalism has resulted in the extreme wealth of a few and increasing crime rates. Many of the masses in these countries are out of work, out of home, and out of luck. More recently, the rapidly expanding boom economies of Southeast Asia have experienced sharp downturns that have resulted in economic chaos and human suffering. (p. 268-269)

With so many people unable to participate in online commerce or online educational resources, it is hardly reasonable to assume that the personal computing movement may be described as at the stage of pervasiveness, let alone universal. It is an intellectually exciting vision that personal computing is or will soon become less and less dependent on group-controlled servers and thereby move computing to become ubiquitous. But, the nagging question of who will be included and excluded prevails.

2   Eight Potentially Disenfranchised Groups with Limited Influence in Cyberchange

According the McGinn, McCormick, (1999) and Macionis (2001), telecommunicating has special appeal to workers who want to hold a job while still caring for small children or aging parents. But this makes telecommuting more popular with women who take great responsibility with family care, but this pattern has potential to perpetuate gender inequality at work in which women may advance more slowly than workers who work all day in the office.
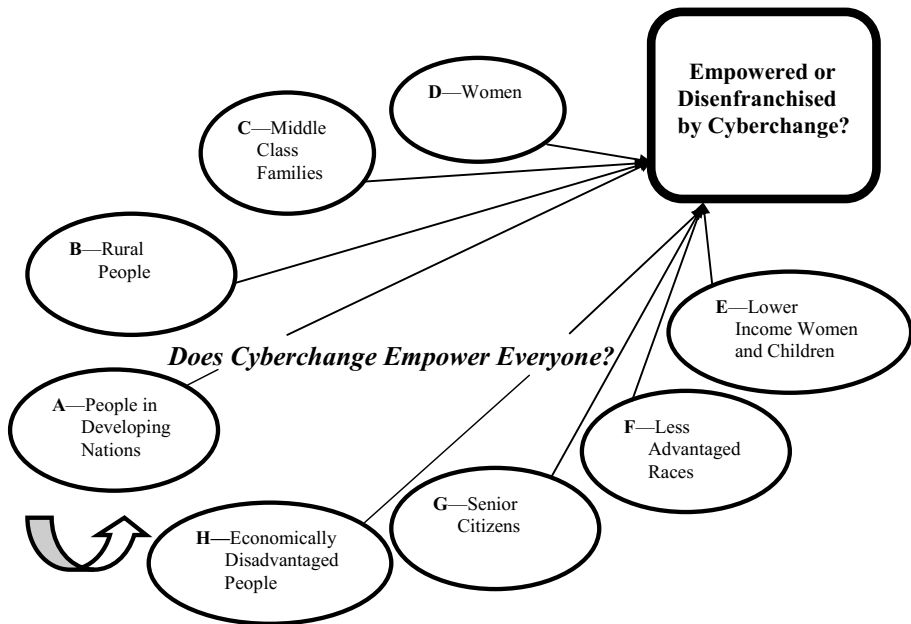
Fig. 1. The Potentially Disenfranchised Groups in Cyberchange

Attitudes, behaviors, conceptions, and motivations have been selected by some researchers as the variables affecting the adoption of Web-based learning technology. It is important to realize that the relationships among attitudes to use, intentions to use, and actual usage in the field of information technology are interactive in the development of distance learning. Information technology acceptance, as well as intentions to use, and actual usage is interdependent. These concepts are all very important research issues in the Management Information System field (Chau, 1996; Davis, 1989, 1993; Davis, Bagozzi, & Warshaw, 1989; Venkatesh, 1999). When these attitudes are further complicated by stereotypes among groups and economic and power differences, the attitudes become difficult to change. Yet, the challenges for advocates for a world of pervasive computing need to understand that attitudes either support or resist even the best intended changes.

Albert Einstein's conception even in 1931 was that scientists and technologists should have a deep concern for the fate of humanity, keeping this as the "chief interest of all technical endeavors, concern for the great unsolved problems of organization of labor and the distribution of goods -- in order that the creations of our mind shall be a blessing and not a curse to mankind" (Einstein, 1931). In any serious deliberation about the utility or general value of innovation and the "magic" of telecommunications, there should be an equal commitment of scholars to the macro issues of providing responsible inclusion of others to understand how well intended technological developments will improve the fate of humanity. George Bernard Shaw (1903) wrote, "Liberty means responsibility. That is why most men dread it."

3   Gender as a Primary Variable in Receptivity to Cyberlearning

In their research with a population of a university population in Taiwan, the writers' used a modification of the Technology Acceptance Model (TAM) of Davis, Bagozzi, and Warshaw (1989) to determine which variables were most significant among the array of factors related to attitudes toward computer and Web uses. The writers were frankly surprised to find—even in Taiwan with an international reputation for being the center of the semi-conductor industry—the two variables of (1) **gender** and (2) **intention to use online courses** were the most significant in identifying receptivity to diffusion of online courses. Women were less receptive or experienced than men.

This finding encouraged the writers to speculate on how attitudes of women may be negatively at work in the global information society, the importance of further understanding the reticent phenomena, and the implications for selecting and nurturing women as educators and leaders interested in a positive rate of adoption and adaptation of online resources in global education.

Davis theorized in the TAM that an individual's actual system usage is determined by behavioral intention, which is in turn determined by perceived usefulness and perceived ease of use. The TAM also has become a powerful model for predicting user acceptance of technology and has been studied in more than 100 studies focusing on the validation of the model across different settings (Lucas & Spitler, 1999).
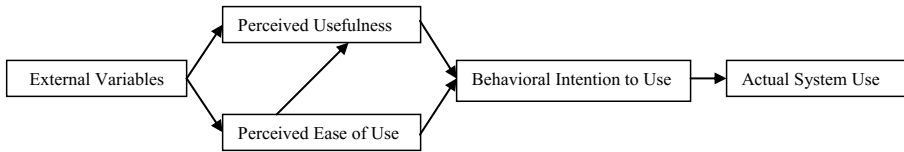
Fig. 2. Technology Acceptance Model (TAM) (Davis et al.

In the writers' research in Taiwan, the writers followed the work of Lee (2001) who developed a modification of the Technology Acceptance Model and used this structural equation modeling to decipher learner behaviors related to adopting Web-based learning technology in adult and higher education. Some empirical studies have examined attitudes on a wide range of topics, such as user-attendance management (Frayne & Latham, 1987), computer skill acquisitions (Gist, Schwoerer, & Rosen, 1989; Mitchell, Hopper, Daniels, George-Falvy, & James, 1994), and user-acceptance of technology (Agarwal, Sambamurthy, & Stair, 2000).

Several theoretical frameworks have been proposed and empirically tested for the adoption of information technology. Following are research studies related to acceptance of information technology. The Technology Acceptance Model has been one of the most influential theories in accounting for information system and technology acceptance of end users (Davis, 1989; Davis et al., 1989). The TAM concept was developed from the social psychology theory of Technology Reasoned Action (TRA) (Ajzen & Fishbein, 1980; Fishbein & Ajzen, 1975), and it explained user acceptance of a technology based on user attitudes.

Based on prior literature reviews, attitudes, behaviors, conceptions, and motivations have been selected by some researchers as the variables affecting the adoption of Web-based learning technology. It is important to realize that the relationships among attitudes to use, intentions to use, and actual usage in the field of information technology are interactive in the development of distance learning. Information technology acceptance, as well as intentions to use, and actual usage is interdependent. These concepts are all very important research issues in the Management Information System field (Chau, 1996; Davis, 1989, 1993; Davis, Bagozzi, & Warshaw, 1989; Venkatesh, 1999).

Studies in the US have reported gender differences as a contributing factor in self-efficacy. Miura (1987), for example, found males to have significantly higher computer self-efficacy than females in sample of undergraduate students. Males also scored higher on perceived relevance of computer skills to future career, interest in knowing how a computer works, and intentions to take computers courses.

The writers conducted research in an available population of adult men and women pursuing technical careers in Taiwan in order to test assumptions from the literature that women, even in technology-rich Taiwan, would be less receptive to use of online courses than their men counterparts. The findings of the study were surprising and disappointing. The Taiwanese women had less receptivity than men to use of online resources, just had been found in studies related to gender in the US.

4   Gender Findings in the Taiwanese Study–Cyberspace Lessons from a Tiger

The writers used the TAM adaptation to assess Taiwan technical university students in one institution (Tajen Technological Institute, the place of work of writer Wang) to examine specific factors as predictors of which adult learners will likely continue to be receptive to Web-based learning opportunities. This study also contained different perspectives from prior TAM research, including Web-based learning system features and psychological correlations with external variables. The research focused on determining which attitudes and perceptions predicted user acceptance and use of Web-based learning technology.

4.1   Assessing Technical University Students in a Land of the Crouching Tiger

The target population of the writers' Taiwanese study contained all postsecondary students who had basic computer-related literacy courses and were enrolled in three different departments at Tajen Institute of Technology in Taiwan during the fall semester of 2002. The assumption was that students in technology institute computer classes would have basic experience in using a Web browser and the Internet. This student experience allowed these respondents to have the minimum requirement for using the study survey that was designed to evaluate respondent attitudes about future participation of students in Web-based learning technology. In addition, this study adopted intact groups, such as computer class cohort clusters, to complete a short training session for Web-based learning.

Cluster sampling was used to select computer classes for the study. There were 23 such classes in the fall of 2002. Students from different academic majors of study who were taking computer courses participated in this study. The departments of Management Information Systems, Healthcare Administration, and Pharmacy each had computer-related courses being taken by their students. Also, the instructors of these courses demonstrated Web-based learning instruction for students, as well as hands-on learning trials. The students were over 18 years old and were from the above academic departments. For these three academic departments, there were 11 classes within the higher education division and 12 classes within the continuing education division.

These classes were numbered, and from a table of random numbers, three were selected to represent the continuing education subgroup, and three were selected to represent the higher education subgroup. For the higher education division, the selected three classes each had approximately 50 students enrolled. For the continuing education division, the selected three classes also each had approximately 50 students enrolled. To sum up, there were a total of six classes selected by cluster sampling from 23 available classes. Also, two classes of the remaining 17 classes first took part in a field test to assure the reliability and validity of instrument. Krejcie and Morgan (1970) designed a table for determining sample size from a given population and deciphered the relationship between sample size and total population. According to the population data they provided, of a population size of 1,300 students who enrolled

in basic computer-related literacy courses in the fall academic year of 2002, the sample size should be approximately 297.

The 1,320 student population included 638 students in higher education and 682 students in continuing education. Because of the relative equality in the size of these groups and the use of cluster sampling procedures, the number of students in the research sample was 265 respondents out of a total number of 312 students in the sample size. From the higher education division there were 135 and 130 were from the continuing education division.

Even with strong Taiwanese parental advocacy that daughters select a technical career—such as management information systems, health care management, and pharmacy—the women in the writer's research were two times more reticent about using online resources or taking online courses than men counterparts. To advance the pace of appropriate use of online education, this gender bias research needs to be a high priority across the world. A plan needs to be developed to both educate women about the benefits of a cyberworld and to provide them with the knowledge and skills to overcome their reticence and to lead in the use of cyberlearning for improving global education

5   Variables in the TAM

Davis (1989) identified user attitudes as directly affecting behavioral intentions to use the information technologies. Behavioral intentions, in turn, directly influenced actual usage. The core constructs of the TAM concept as modified by the writers for a study of university technical university degree students in Taiwan included prior computing experience (PCE), computing skills (CSs), perceived usefulness (PU), perceived ease of use (PEOU), and behavior intention to use (IU, with IU1 related to intention to use Web resources for supplemental learning, and IU2 related to intention to use the Web for distance education courses).

The following Figure 3 shows that among the TAM variables in the writers' study in Taiwan, gender had the most significance in predicting positive perceptions toward usefulness of computing and its ease of use, as well as the intention to use the Web for distance education courses. The probability levels of .002, .005, and .022 on the Figure were significant at the .05 level.
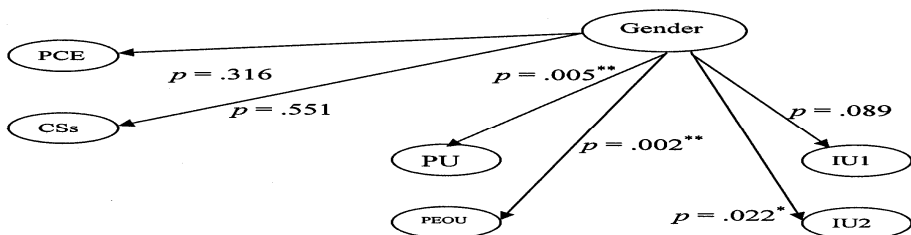
Fig. 3. Gender as the most significant variable to predict receptivity

It is interesting to speculate how the order of significance in the above constructs may or may not be influenced by cultural context. The writers look forward to dialogue with East and West colleagues about this issue. Adult students across the world who have become able to embrace the information technology can make it a part of their learning culture.

The writers modified the TAM instrument five sections (from A to E). First, the perceived usefulness (PU) and the perceived ease of use (PEOU) related to use of Web-based learning technology were placed into a multiple-choice format in section A. Section B asked questions about the intentions of using Web-based learning either as a supplementary learning tool or as an entire online distance education method. Sections C and D asked about attitudes toward computing and attitudes toward Web-based learning with the combined scales of computer attitudes scale (CAS) and Web-based attitudes scale (WAS).

The importance of attitudes and beliefs for learning to use computers has been widely acknowledged. In order to evaluate the perceptions of computer and Web-based environments simultaneously, it was necessary to develop a survey that included these two attitudes scales, such as the Computer Attitude Scale (CAS) and Web-based Attitude Scale (WAS).

A number of instruments have been developed to measure attitude towards computers (Brock & Sulsky 1994; Cambre & Cook 1985; Jones & Clarke 1994). The Computer Attitude Scale, developed by Loyd and Loyd (1985), consists of four subscales: anxiety of computers, liking of computers, confidence in computers, and usefulness of computers, which has been tested numerous times for validity and reliability.

Due to the fact that the TAM computer attitude scale could not survey the perception of computer and Web-based environments simultaneously, it was essential the study of the writers to develop a survey that included these two attitude issues. Therefore, there were two scales, the computer attitude scale (CAS) and the web-based attitude scale (WAS), in this study to recognize the perception of computer and Web-based environments.

Furthermore, the Internet and World Wide Web (WWW) are appropriate computer technologies to create a Web-based environment. In this study, the survey structure of Web-based environment was based on the structure of the Internet and WWW technologies. Similarly, the WAS also contained these four subscales in which respondents were asked to respond to a 16-item Web-based attitude scale survey concerning attitudes toward using the Internet, WWW, and their experience in order to cover the attitude components such as affect, cognition, and behavior.

For computing to become ubiquitous, both genders need to be advocates of the changes. The climate for adult web uses in the twentieth first century is important area concern for e-commerce or e-service. It takes time for people to accept an attitudinal

position toward web uses. Adults who have felt comfortable when using the web as instructional tool are more likely to integrate the daily activities (Riel, 1993).

6   Rural and Urban Differences

The market for high technology is necessarily in the locations of the world that have the most dense fiber connections. The fiber optic network has evolved in the largest population centers of the world. If there were more Internet connections in Tokyo than in all of Africa during the past decade, then the promotion, support, and consumption of online resources would certainly have been in Japan. The less economically self-sufficient and rural populations of the world are left disenfranchised when urban markets and unregulated capitalism determine the directions for technology development.

Keller and Klein (1999) wrote the following conclusion about factors in less developed regions that contribute to economic development trends:

> The factors contributing to the cycle of underdevelopment create structural problems that make development difficult. Outside interventions can either help of hinder the process. They can inadvertently reinforce the problems while purporting to provide solutions. Trade, aid, private investment, and technical assistance act positively on an economy when local elites have achieved stable economic and political decision-making processes. Such stability results from a social consensus…. Thus, in addition to economic factors, any analysis of developing world problems must account for the underlying social factors that may explain why some countries achieve ongoing growth and improved standards of living while others do not. (p. 67)

The urban and rural split has contributed to the digital divide. Even in South Dakota in the US Midwest, the high speed Internet II fiber optic network is only available in cities and towns. The writers have been painfully aware that the slow speed of reliance on twisted copper telephone wire makes the use of rural telephone service a frustrating experience when graphics, pictures, and video are a part of communications. The urban setting, with its higher speed networking services, is advantaged and may allow itself in the protection of its members to think that cyberspace is universal. But, in truth, the services of the Internet are far from universal or pervasive in most rural parts of the world.

6.1   A Graphic Display of the Eight TAM Instrument Variables

The TAM instrument constructs, as modified by the writers, has eight primary variables as in Figure 4 below.

Primary Receptivity Constructs Related to Use of Online Resources
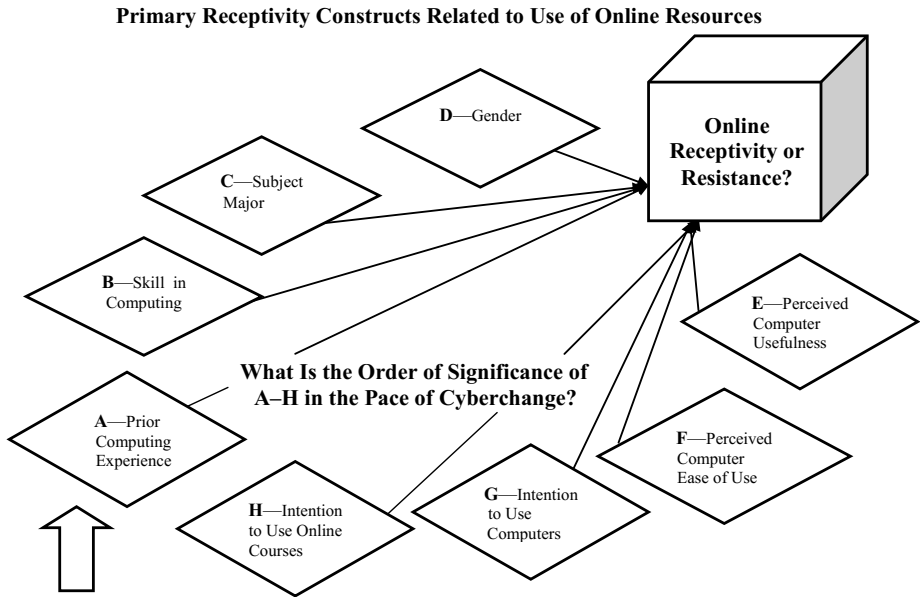
Fig. 4. The Writers' Modified TAM Receptivity Constructs Related to Use of Online Resources

In Figure 4 above, the order of significance—from most to least significant—of the variables in the writers' Taiwanese study was as follows: (1) D—Gender, (2) H—Intention to Use Online Courses, (3) G—Intention to Use Computers, (4) F—Perceived Computer Ease of Use, (5) E—Perceived Computer Usefulness, (6) A—Prior Computing Experience, (7) B—Computing Skill Levels, (8) C—Subject Major.

7   The Impact of Socio-Economic, Race, and Age Differences

The increasingly complex knowledge and skills required for full participation in the cyberlearning and e-commerce world require levels of education and outlay of funding for extended and continuing education that are significantly out of reach of many middle class families, single parent families, and seniors, let alone for the working poor. The clever and omnipresent marketing of the large computer conglomerates exhausts even the most cheerful advocate for increasing the pace of cyberlearning. But, for the poor, disadvantaged, isolated, and unempowered, the challenge to be a part of the freedom of the body politic in influencing the direction or pace of the cyberworld is hardly on their minds.

The notion that ordinary people might share in the further development of a cyberculture as planners and developers seems almost ridiculous in the face of today's dominance of Microsoft and a small number of people who think of pervasiveness in terms of making the majority of the citizens of the world into consumers. Guell (2003) had the following argument about how US society is trapped by big business:

When a business treats us badly, most of us get a high degree of satisfaction by announcing that we will never be back… When a representative of a monopolistic company makes you mad, you know the representative knows that you have alternatives; you are stuck... For capitalism to function, these situations work best when there is both a carrot of high profits and a stick of bankruptcy to keep firms working in the consumer's best interest. Without such incentives, a company's profit motive tends to work against consumers rather than in their best interests. (p. 332)

As the mega-corporations have increasingly dominated the cyber world, the potential for the disenfranchised to increase in number is likely. Kimmelman reported (2000) on consumer groups and the digital divide, giving the following grim report about the status of the "disadvantaged and disenfranchised" in the computer world:

Gene Kimmelman, co-director of the CU's (Consumers Union) Washington office, said the (CNN) report shows that "vulnerable groups are harmed by their lack of access to technology" and policymakers "should begin to seek cost-effective avenues to address this deprivation." (p. 2)

In the scope of this paper, the very significant issues in age and race discrimination, which have been written about extensively by other, will not be addressed. But, it should be emphasized that matters of economic discrimination based on race also contribute to defeating the Bill Gates' dream of having pervasive computing among people of a very large upper middle class.

The persistent digital divide between and among groups of the world who have and those who have not access to computing resources begs for constant consideration by educators and policy makers who profess to be "change agents." The director of the Consumer Federation of America, Mark Cooper, made the following conclusions (2000) about the need for policy on assuring that the cyberworld will focus on people, education, and resources, and not on corporate tax breaks:

While computer ownership and Internet use continue to grow, the "digital divide" that separates those Americans connected to the Internet from those who not persists and is not likely to disappear any time soon,… The gap puts millions of Americans at a disadvantage in our increasingly "online" society. The more important the Internet becomes, the more serious the problems will be, unless steps are taken to close the gap;… a detailed national survey of 1900 respondents found that 47% of the respondents do not have access to the Internet at home. The "disconnected" are much more likely to be lower income, older, and minority households.   (p. 1)

The educational systems of the world continue to set goals, design tests, and work their efforts at developing a meritocracy to select the best minds for membership into the ranks of the educated and into the laboratories of scientists and engineers. Then, in a peer culture of people of similar backgrounds, society encourages the educated to work their magic through research. It is no surprise that specialization and departmentalization create exclusive societies that are responsible for leadership. The

very nature of the narrowness of professional organizations, however, creates a continuing need for these groups to be vigilant in seeking ways to reengage themselves in the larger diverse society.

The university education establishment has long hoped for a rigorous liberal arts education to reinforce members of all specialized knowledge groups to remember to be a part of the whole of society. In matters of contemporary technology development, it remains to be seen whether or not the general and liberal arts education is adequate in producing highly specialized professionals who can both produce important innovations, while yet sharing in the responsibility to communicate visions and better practices to the public at large. The synthesis of specialized knowledge and the pursuit of the integration of knowledge and diverse people of the world continue to be significant challenges for educated people with a social conscience.

It is a challenge for educators to understand engineers, for computer analysts to relate to poets, for the rich to empathize with the poor, and for the races to live in harmony. But, the brave new cyberworld depends on the capability of a vision of universal computing to be embraced by the majority of the groups of the world. As the highly-specialized research on information technology continues, it may be very useful and essential for professional societies to assure that in most every conference and journal there is an opportunity to share perspectives, visions, needs, and promising innovations with colleagues and citizens from other specialties and walks of life.

Castells wrote from a sociological point of view on the emerging global network society. He concluded that the new global informational economy is capitalistic in which sources of productivity and competitiveness for firms, regions, and nations depend, more than ever, on knowledge, information and technology processing. He maintained that this new economy is potentially more exclusionary than the industrial economy if it is not adequately regulated. It reaches out to the world but excludes the majority unevenly through switching on some links and switching off others. Castells believes that the Third World has become increasingly diversified internally, the First World has generated exclusion, and an emergent Fourth World of exclusion is populated largely by women and children (Castells, 2002).

In the highly industrialized nations of the world, achievement motivation for competing in global markets consumes most of the energies of educated and specialized people. In the US, Japan, and Taiwan, from the perspective of the writers, most professional people work at a frantic pace to remain informed about changing knowledge and fruitful areas for research. For many educated professionals, keeping up with change in an area of micro-specialization precludes the time necessary for reflecting on the macro-societal issues that are of concern to less fortunate members of society.

How emerging technologies may or may not be relevant to matters of redistribution of wealth, the happiness of men or women in the family, or to human rights and dignity are hardly matters that very busy professionals have the time to consider. Yet, as the great thinkers of the world, including Einstein, have cautioned, the future of human innovation needs always to be judged on its service to broader humanity. All the commitments to sixty-hour work weeks, efforts at cram schools, struggles to finish online or campus-based degree programs, or sacrifices to meet economic goals will

mean little if they do not help to ameliorate the problems of poverty, discrimination, and isolation of many people in the world. E-service, E-commerce, and E-learning will be best when they demonstrate that they are essential to world peace, the expansion of the middle classes, and the creation of an inclusive global society.

8   Summary

Regarding the vision of universal pervasive computing, there is a need for systematic, reasoned, and humanistic thinking about the problem of disenfranchisement of so many parts of society. The large scope of the problem should cause significant pause for advocates of the further development of the cyberworld.

The number of people who yet have either inferior service or no access needs to be a major concern to research and development leaders in engineering and e-commerce. The challenge for cyberspace leaders is to recognize and resolve the problem of the disenfranchised groups in the world who have not yet been touched by the concept of universal pervasive computing. Large numbers of people remain inaccessible to e-commerce and disenfranchised from participating in the new world technology vision.

The continued dominance of urban markets controls the direction of the free enterprise development of computing. Less populated and smaller markets in rural parts of the world go unserved or have to endure much slower, inferior, and non-competitive access. This disparity among groups is a reality in industrialized nations, as well as developing countries. Women continue to be significantly less receptive to the cyberlearning vision than men. Studies in the US and Asia continue to show that cyberspace advocates have yet to promote a concept of a better society that many women will advocate, let alone engage.

Professional engineering and education societies need systematic plans to increase potential for inclusiveness in the practical implementation of the universal pervasive computing vision. Innovation unheeded by major parts of society defeats the efforts of exceptional researchers. To further develop a universal cyberworld, technology advocates need to include groups that have been disenfranchised or disinterested.

National and global interventions are necessary to recognize these sociological issues and to make policies that will distribute access for the achievement of the vision. Professional engineering, education, and technological societies need to accept responsibility to be inclusive, rather than exclusive, in developing a universal computing vision that can be have pervasive and ubiquitous character.

> The question is not whether "big is ugly," "small is beautiful," or technology is "appropriate." It is whether technologists will be ready for the demanding, often frustrating task of working with critical laypeople to develop what is needed or whether they will try to remain isolated, a luxury I doubt society will allow any longer.
>
> Robert C. Cowan, 1980

References

Brock, D.B., Sulsky, L.M.: Attitudes toward computers: Construct validation and relations to computer use. Journal of Organization Behavior 15, 17–23 (1994)

Cambre, M.A., Cook, D.L.: Computer Anxiety: Definition, Measurement, and Correlates. Journal of Educational Computing Research 1(1), 37–54 (1985)

Castells, M.: An Introduction to the Information Age. City, 6–16 (May 1997). An introductory address to the conference on information and the city. School of Geography, Oxford University 22(3), 96

Chau, P.Y.K.: An empirical assessment of a modified technology acceptance model. Journal of MIS 13(2), 185–204 (1996)

Cooper, M.: Consumer Federation America Reports (2000),
http://www.cnn.com/2000/TECH/computing/10/11/digital.divide
(retrieved October 14, 2003)

Cowan, R.C.: Technology Review 22(6), 62 (1980)

Craig, K.: Faculty development efforts in support of Web-based distance education among the schools of education within a university system. In: Start Teaching National Conference (2002)

Davis, F.D.: Perceived usefulness, perceived ease of use and user acceptance of information technology. MIS Quarterly 13(3), 319–339 (1989)

Davis, F.D., Bagozzi, R.P., Warshaw, P.R.: User acceptance of computer technology: a comparison of two theoretical models. Management Science 35, 982–1003 (1989)

Davis, F.D.: User acceptance of information technology: System characteristics user perceptions and behavioral impacts. International Journal of Man-Machine Studies 38, 475–487 (1993)

Einstein, A.: From an address of Cal. Harper Row Publishers, New York (1931)

Guell, R.C.: Issues in Economics Today. McGraw-Hill/Irwin, New York (2003)

Harasim, L., Hiltz, S.R., Teles, L., Turoff, M.: Learning networks: A field guide to teach and learning online. MIT, Cambridge (1995)

Jones, T., Clarke, V.A.: A computer attitude scale for secondary students. Computers and Education 22(4), 315–319 (1994)

Kelleher, A., Klein, L.: Global Perspectives: A Handbook for Understanding Global Issues. Prentice-Hall, Upper Saddle River (1999)

Kimmelman, G.: Consumers Union Reports (2000),
http://www.cnn.com/2000/TECH/computing/10/11/digital.divide
(retrieved October 14, 2003)

Loyd, B.H., Loyd, D.E.: The reliability and validity of instruments for the assessment of computer attitudes. Educational and Psychological Measurement 45, 903–908 (1985)

Macionis, J.J.: Welcome to cyber-society. In: Macionis, J.J., Benokraitis, N.V. (eds.) Seeing Ourselves: Classic, Contemporary, and Cross-Cultural Readings in Sociology, 5th edn., pp. 62–67. Prentice-Hall, Upper Saddle River (2001)

McGinn, D., McCormick, J.: Your next job. Newsweek 38, 43–51 (1988)

Miura, I.T.: The relationship of computer self-efficacy expectations to computer interest and course enrollment in college. Sex Roles 16(5/6) (1987)

Peter, R.D., Wunder, D.F., Mueller, H.L.: Social Problems: Globalization in the Twenty-First Century. Prentice-Hall, Upper Saddle River (1999)

Riel, M.: Global education through learning circles. In: Harasim, L. (ed.) Global Networks. MIT Press, Cambridge (1993)

Russell, D. (2001), `http://searchnetworking.techtarget.com/sDefinition/0,290660,sid7_gci759337,00.html` (retrieved on December 12, 2003)

Shaw, G.B.: Man and Superman. The University Press, Cambridge (1903)

The World Bank, World Development Report; The State in a Changing World. Oxford University Press, New York (1997)

Turkle, S.: Computational reticence: Why women fear the intimate machine. In: Hopkins, P.D. (ed.) Sex/Machine: Readings in Culture, Gender, and Technology, pp. 365–380. Indiana University Press, Bloomington (1988)

Venkatesh, V.: Creation of favorable user perceptions: exploring the role of intrinsic motivation (1999), `http://www.mbs.umd.edu/is/venkate/myhomepage/misq99/misq1.htm` (retrieved October 14, 2002)

Human Fetus Health Classification on Cardiotocographic Data Using Random Forests

Tomáš Peterek[1], Petr Gajdoš[1,2], Pavel Dohnálek[1,2], and Jana Krohová[3]

[1] IT4Innovations, Centre of Excellence
[2] Department of Computer Science, FEECS
[3] Department of Cybernetics and Biomedical Engineering, FEECS
VŠB - Technical University of Ostrava,
17. listopadu 15, 708 33 Ostrava, Czech Republic

Abstract. Pregnancy and fetus development is an extremely complex biological process that, while generally successful and without complications, can go wrong. One of the methods to determine if the fetus is developing according to expectations is cardiotocography. This diagnostic technique's purpose is to measure the heartbeat of the fetus and uterine contractions of its mother, usually during the third trimester of pregnancy when the fetus' heart is fully functional. Outputs of a cardiotocogram are usually interpreted as belonging to one of three states: physiological, suspicious and pathological. Automatic classification of these states based on cardiotocographic data is the goal of this paper. In this research, the Random Forest method is show to perform very well, capable of classifying the data with 94.69% accuracy. A comparison with the Classification and Regression Tree and Self-organizing Map methods is also provided.

Keywords: random forest, CTG, fetus, SOM, decision tree.

1 Introduction

Although the quality of prenatal care is still improving and the fetus mortality rate decreases every year, the death of a fetus in it's later stages of development is still a very real possibility. Therefore, fetus observation and actual state evaluation is very important throughout the whole gestation. The prenatal care is one of the most difficult medical branches as fetuses are incapable of describing their feelings and problems to the obstetrician. Therefore, the obstetricians can observe only the physical manifestations using tools like the ultrasound, electrocardiography or echocardiography. In 1960, a new examination method known as the cardiotocography (CTG) was invented. CTG has an irreplaceable role in the prenatal care and can be used from the 27th week of the pregnancy. Examination by CTG is usually recommended to pregnant women that suffer from diabetes, high blood pressure and other diseases that can endanger the fetus. CTG can measure the heart activity of a fetus, it's movements and the

mother's uterine contractions simultaneously. There are two main principles: invasive examination (intrapartum) and noninvasive examination (antepartum). The gynecologists and obstetricians observe especially four basic parameters:

1. Fetal heart rate baseline (BL) - physiological values in the range of 110-150 beats per minute (bpm). BL values higher than 150 bpm or lower than 110 bpm could mean tachycardia or bradycardia, respectively.
2. Acceleration (ACC) - an increase of the BL larger than or equal to 15 bpm observable for at least 15 seconds. The acceleration should occur at least twice every 15 minutes. Acceleration observed at night can be considered pathological, but they can also be a natural response to the movement of the fetus.
3. Deceleration (DCL) - a decrease in the BL larger than or equal to 15 bpm observable for at least 15 seconds. Combined with uterine contractions, decelerations may indicate fetal hypoxia (oxygen supply deprivation).
4. Variability - fluctuations in the BL that are not evaluated as either acceleration or deceleration. When observing the variability parameter, sleep and daytime activities must be taken into account as the parameter can differ based on the two time periods.

   The standardized method of evaluating the actual fetus state was determined by the Federation of International Gynecology and Obstetrics (FIGO) in 1986. According to this standard, CTG records can be classified to one of three basic groups: physiological, suspected and pathological, also known as normal, indeterminate and abnormal.

1. **Physiological:** Physiological state of a CTG record is indicated by the BL of 110 to 160 bpm with baseline variability of 5 to 25 bpm, at least 2 accelerations greater than 14 bpm during a 15 minute period, no decelerations at all or no decelerations exceeding 14 bpm for longer than 15 seconds and the presence of a moderate tachycardia (161-180 bpm) or bradycardia (100-109 bpm). However, moderate tachycardia and bradycardia are considered to be normal only under the condition that the observed baseline variability and accelerations do not vary. If the subject being monitored meets these criteria, further monitoring may be deemed unnecessary.
2. **Suspicious:** The BL from 100 to 109 bpm and from 161 to 180 bpm is in the suspicious category of an antenatal CTG. There is a reduced baseline variability of less than 5 bpm for at least 40 minutes but less than 90 minutes, or a variability greater than 25 bpm. At least two accelerations of 15 bpm and more lasting at least 15 seconds should occur every 15 minutes. The deceleration is variable with depth less than 60 bpm and with duration less than 60 seconds. A CTG record is considered suspicious if exactly one of these criteria is met. Further assessment is required by a specialist.
3. **Pathological:** A CTG record is considered pathological if the BL exceeds 180 bpm or drops below 100 bpm, the variability drops below 5 bpm for a period longer than 90 minutes, an atypical variable or late decelerations occur and last over 30 minutes, a single prolonged deceleration occurs and

lasts longer than 3 minutes, the heart of the fetus being monitored starts beating with a sinusoidal pattern for more than 10 minutes or a prolonged bradycardia occurs. Two or more occurrences of these symptoms are a reason for the immediate attention of a specialist.

The full description of CTG records and its classifications are largely beyond the scope of this paper. We encourage the reader to further study the matter, for instance, in [1,2] or [3].

2   Dataset

The dataset used in this paper was created at Faculty of Medicine in the University of Porto and consists of 2126 records. Each record was manually evaluated by three obstetricians and consequently classified to one of the three groups. The original CTG records were analyzed by SiSPorto 2.0, a software for automatic CTG record processing. From each of the records, 21 attributes (features) that are related to fetal heart rate, fetal movement and uterine contraction were extracted. The dataset contains 1655 physiological records (77.8%), 295 suspicious records (13.8%) and 176 records considered to be pathological (8.2%) [4].

3   Previous Work

Many studies focused on the same or similar field of research were done in the past and many more are still undergoing extensive work. Such studies usually focus objectives related to analyzing CTG records such as automatic baseline determination [5] [6], classification of the type of an acceleration and declaration [7] and feature extraction [8]. Papers focused on CTG record classification describe several methods capable of dealing with the problem with varying success. The accuracies achieved by a neural network (NN) model and the adaptive fuzzy interference system (ANFIS) were compared in [9]. While neural networks are a recurring topic in this area of research, proven by the use of the modular neural network in [10] or the back-propagation neural network in [11], other methods are applicable as well. Linear Discriminant Analysis and Decision Trees were compared with neural networks, for example, in [12].

The following subsections provide a brief description of the methods used and evaluated in this study. The choice of the methods is based on our previous research where the methods were successfully applied to various classification problems. Similarly to [12], the study compares a neural network model with the Decision Tree paradigm.

3.1   Classification and Regression Tree

The Classification and Regression Tree (CART) was first designed and described by Breiman in [13]. This structure uses binary divisions to grow and expand itself, resulting in a tree structure of conditions (yes/no questions) and outcomes. First,

Table 1. The ANOVA values for each of the features

Attribute name	Physiological	Suspicious	Pathological
Baseline value (BL)	131.97 ± 9.45	141.68 ± 7.88	131.68 ± 9.43
Acceleration (AC) [frequency]	0.004 ± 0.004	0 ± 0.001	0 ± 0.001
Fetal movement (FM) [frequency]	0.008 ± 0.041	0.008 ± 0.042	0.02 ± 0.08
Uterine contraction (UC)	0.005 ± 0.003	0.002 ± 0.003	0.004 ± 0.004
Light deceleration (DL)	0.002 ± 0.003	0.001 ± 0.002	0.004 ± 0.004
Severe deceleration (DS)	0 ± 0	0 ± 0	0 ± 0
Prolonged decelerations (DP)	0 ± 0	0 ± 0	0.001 ± 0.001
Percentage of time with abnormal short term variability (ASTV)	42.44 ± 15.48	61.90 ± 11.76	64.54 ± 14.45
Mean value of short term variability (MSTV)	1.43 ± 0.81	0.63 ± 0.66	1.57 ± 1.19
Percentage of time with abnormal long term variability (ALTV)	5.02 ± 11.5	29.03 ± 20.26	22.84 ± 33.99
Mean value of long term variability (MLTV)	8.70 ± 5.82	8.02 ± 3.75	3.58 ± 4.09
Histogram width (Width)	73.41 ± 36.35	49.15 ± 39.51	78.34 ± 49.07
Low freq of the histogram (Min)	91.07 ± 27.25	113.29 ± 31.01	83.98 ± 34.17
High freq. of the histogram (Max)	164.48 ± 17.77	162.45 ± 16.33	162.33 ± 21.67
Number of histogram peaks (Nmax)	4.16 ± 2.86	3.31 ± 3.10	4.44 ± 3.31
Number of zeros in histogram (Nzeros)	0.33 ± 0.69	0.24 ± 0.80	0.34 ± 0.67
Histogram mode (Mode)	138.25 ± 13.42	146.55 ± 10.98	114.60 ± 26.14
Histogram mean (Mean)	135.09 ± 13.01	144.75 ± 10.51	112.97 ± 22.78
Histogram median (Median)	138.45 ± 12.60	147.07 ± 10.49	119.56 ± 19.33
Histogram variance (Variance)	17.49 ± 22.29	7.21 ± 18.25	50.73 ± 60.58

all values of the input data are stored in the root node of the tree. The optimal distribution of the particular predictor into two parts is achieved by calculating the criterion statistics that determines the homogeneity of a node. This criterion statistics differs based on the type of tree being used. For regression trees, the criterion that minimizes the mean square error is used. It is necessary to find such a split of variable Y that will have the smallest root mean square deviation values of y_i in the potential node from the average of these values:

$$\overline{y}_t = \frac{1}{N} \sum y_t \tag{1}$$

$$Q(T) = \frac{1}{N_t} \sum_{i=1}^{N_t} (y_i - \overline{y}_t)^2 \tag{2}$$

where N_t is the number of observations in the node t and y_t are the values of the dependent variable in the node t. The criterion statistics computed for

the classification tree is calculated according to the Gini index (equation 3) or entropy (equation 4).

$$GI = \sum_{c=1}^{J} p_{ct} \left(1 - p_{ct}\right) = 1 - \sum_{c=1}^{J} p_{ct}^2 \tag{3}$$

$$H = -\sum_{c=1}^{J} p_{tc} \log_2 p_{tc} \tag{4}$$

In practice, the Gini index is the measure used most often. Its value becomes zero if the final node contains only a single class of the input variable while reaching the maximum value if the final node contains the same number of samples in each class. The entire growth of a classification tree can be defined as the succession of the following steps:

1. The data set is divided into training and testing sets.
2. For continuous variables, the values of each predictor are sorted in the ascending order.
3. The criterion statistics is calculated for all distributions of variables Y into two possible nodes.
4. After the distribution, the predictor with the smallest criterion statistics value is selected.
5. Based on the result in step 2, the set is divided into two daughter nodes t_l and t_2.
6. Repeat from step 2 for both nodes until a termination condition is met.
7. Prune the resulting tree

If a node contains only one observation or if all it's observations have the same values, the tree growth is suspended. The growth of a decision tree can also be parametricized by setting the number of branching tree, setting the maximum observation in the leaf nodes or determining a threshold for the mean square error or the percentage of incorrectly classified samples. Simplifying the resulting complex tree ("pruning") is the process of finding the the optimal size of the tree by using a tree complexity criterion (equation 5). If T_0 is a tree and T_1 is the tree after pruning, the complexity of the tree can be expressed as:

$$C_\alpha \left(T_1\right) = DT_1 + \alpha \left|T_1\right| \tag{5}$$

where $\left|T_1\right|$ is the number of terminal nodes and DT_1 is the error of T_1. The α parameter, $alpha \geq 0$, expresses the relationship between the tree size and it's accuracy. The goal is to find α that minimizes C_α.

3.2   Kohonen Self-organizing Neural Network

In following paragraphs, we will shortly describe the Kohonen self-organizing neural networks (self-organizing maps – SOM). The first self-organizing networks were proposed in the beginning of 70's by Malsburg and his successor

Willshaw. SOM was proposed by Teuvo Kohonen in in the early 1980s and has been improved by his team since. The summary of this method can be found in [14].

The self-organizing map is one of the common approaches on how to represent and visualize data and how to map the original dimensionality and structure of the input space onto another – usually lower-dimensional – structure in the output space.

The basic idea of SOM is based on the human brain, which uses internal 2D or 3D representation of information. We can imagine the input data to be transformed to vectors, which are recorded in neural network. Most neurons in cortex are organized in 2D. Only the adjacent neurons are interconnected.

Besides of the input layer is in SOM only the output (competitive) layer. The number of inputs is equal to the dimension of input space. Every input is connected with each neuron in the grid, which is also an output (each neuron in grid is a component in output vector). With growing number of output neurons, the quality coverage of input space grows, but so does computation time.

SOM can be used as a classification or clustering tool that can find clusters of input data which are more closer to each other. The Kohonen algorithm is defined as follows:

1. **Network initialization**
   All weights are preset to a random or pre-calculated value. The learning factor η, $0 < \eta < 1$, which determines the speed of weight adaptation is set to a value slightly less than 1 and monotonically decreases to zero during learning process. So the weight adaptation is fastest in the beginning, being quite slow in the end.
2. **Learning of input vector**
   Introduce k training input vectors $V_1, V_2, \ldots, V_k$, which are introduced in random order.
3. **Distance calculation**
   An neighborhood is defined around each neuron whose weights are going to change, if the neuron is selected in competition. Size, shape and the degree of influence of the neighborhood are parameters of the network and the last two decrease during the learning algorithm.
4. **Choice of closest neuron**
   We select the closest neuron for introduced input.
5. **Weight adjustment**
   The weights of closest neuron and its neighborhood will be adapted as follows:
$$W_{ij}(t+1) = W_{ij}(t) + \eta(t)h(v,t)(V_i - W_{ij}(t)),$$
   where $i = 1, 2, \ldots, dimX$ a $j = 1, 2, \ldots, dimY$ and the radius r of neuron's local neighborhood is determined by adaptation function $h(v)$.
6. **Go back to point 2 until the number of epochs e is reached.**

To obtain the best organization of neurons to clusters, a big neighborhood and a big influence of introduced input are chosen in the beginning. Then the

Table 2. Confusion matrix and statistics for the CART algorithm

	Physiological	Suspicious	Pathological	Recall
Physiological	1588	69	14	95.03
Suspicious	59	225	5	77.86
Pathological	8	1	157	94.58
Precision	95.95	76.27	89.28	**92.66**

primary clusters arise and the neighborhood and learning factor are reduced. Also the $\eta \to 0$, so the changes become less significant with each iteration.

3.3   Random Forests

The RF method belongs to the group model category of algorithms. The entire principle of the group models lies in the assignment of the problem to several classifiers whose outputs are combined to give the final answer. This approach, also known as *majority voting*, usually provides better classification results than standalone classifiers. Generally there are many ways of determining the the final answer from individual votes. For instance, regression often uses averaging of the results while classification simply picks the answer that was most frequent. As CARTs, the underlying structures (members) for RF, are deterministic, it is obvious that training each CART in the forest with the same dataset would result in the same answer with the same error. Therefore, training dataset needs to be randomized so that each CART handles a different set of information.

The RF algorithm was also designed by Breiman with [15] being the original paper where the entire algorithm was introduced and described. By using CARTs as members of the RF, this method is applicable for both classification and regression problems where it can describe the importance of individual variables and detect outliers. Each RF consists of N CARTs. Any such predictive forest can be expressed as $h(X_1, \Theta_1), \ldots, h(X_N, \Theta_N)$, where h is a function and Θ_N are the predictors. After splitting the input dataset into the testing and training part, bootstrap selections from the training set take place. A bootstrap selection is a repeated random selection of the same number of values of a given observation. The selections are performed independently of each other, leading to some of them being present in training sets for more than one individual CART while others may not be included in any training set at all. Selections not included in the training sets are used to estimate the tree error (the out-of-bag estimation).

4   Experiments and Results

The experiments were performed using the 10-cross-validation method. For all classifiers, a table is provided with their results. The tables show the respective confusion matrices along with summarizing statistics of Precision and Recall. Bold numbers in the tables signify total recognition accuracy and can be used as a comparable measure across methods.

Table 3. Confusion matrix and statistics for the SOM algorithm

	Physiological	Suspicious	Pathological	Recall
Physiological	1619	33	3	97.83
Suspicious	111	182	2	61.70
Pathological	27	27	122	69.32
Precision	92.15	72.21	96.06	**90.54**

Table 4. Confusion matrix for RF with 100 trees and 15 variables

	Physiological	Suspicious	Pathological	Recall
Physiological	1620	28	7	97.89
Suspicious	58	232	5	78.64
Pathological	8	7	161	91.48
Precision	96.09	86.89	93.06	**94.69**

First of all, the regular CART was tested in order to obtain results that will serve as a reference. The standard MATLAB implementation of the method with default settings was used. Its results are shown in Table 2. For RF, the optimal settings have to be found experimentally. Throughout the experiments, the number of trees in the forest was gradually increased (10, 100, 200, 300, 500, 1000, and 2000), and so was the number of features in a tree (2, 5, 10, and 15). While Table 4 shows only the most successful setting in terms of recognition accuracy, Figure 1 shows the precision values for each class for the particular numbers of trees in the forest and all the numbers of variables per tree. The settings with only 2 features per tree was considered a failure since the pathological state recognition precision was about 85%, and the suspected state recognition success rate was only about 75%. A higher count of variables per tree brings better classification performance. RF provided the best results when there were 100 trees in the forest and each tree was trained using 15 features. SOM was set to compute a 10×10 hexagonal grid and 500 epochs.

CART performed rather well in terms of recall where it achieved the best result for the pathological fetus state. Its recall for the physiological state was, however, outmatched by both RF and SOM. In terms of precision, RF and CART provided very similar results for the physiological state. Regarding the total recognition accuracy, RF proved to be the most suitable method of the three with its result of 94.69%. While the recognition accuracy of SOM and CART could be considered satisfactory, the deficit against RF is noticeable and suggests that the methods may not be as suitable in practice. As usually is the case, the lower accuracy has an advantage in terms of computation speeds. SOM, being the least successful in recognizing activities with 90.54% success rate, is also the fastest one to classify an incoming vector. RF, on the other hand, was the slowest. However, as the training part of all three algorithms can be done in advance, all three methods can be considered more than sufficiently fast.
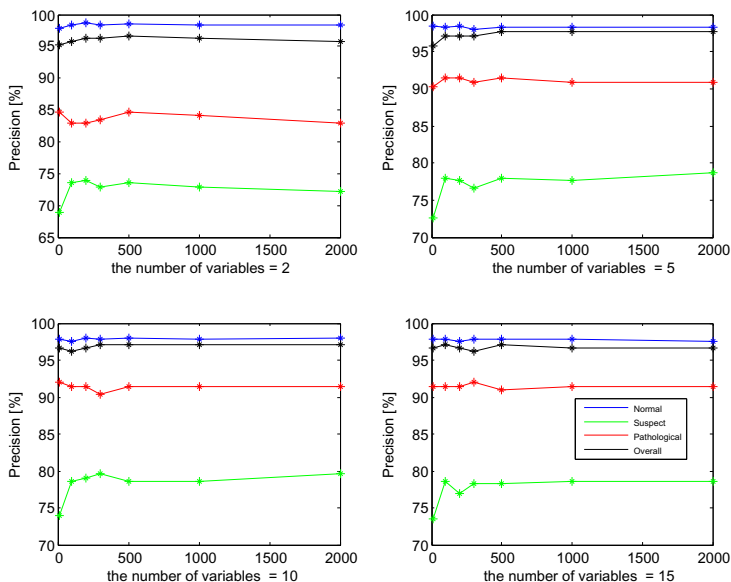
Fig. 1. RF precision for different numbers of trees and variables. The x-axis contains the numbers of trees in the forest.

5   Conclusion

In this paper, a method capable of classifying the standard three fetus states from a cardiotocogram with 94.69% accuracy was presented. The accuracy rivals and often exceeds that reported in other papers, making the Random Forest method a reasonable choice to experiment with in any further research in this area. In further research, an intra-patient approach to the problem (that is, the ability of the methods to generalize upon the training data) needs to be tested and evaluated. Accuracy can also still be improved which could be achieved by modifying RF from a general-purpose classifier to a CTG-specific one. Great potential lies in massively parallel approaches as these could speed up the training process significantly, perhaps making it possible to retrain the methods online.

References

1. Robinson, B.: A review of nichd standardized nomenclature for cardiotocography: the importance of speaking a common language when describing electronic fetal monitoring. Reviews in Obstetrics and Gynecology 1(2), 56 (2008)
2. Alfirevic, Z., Devane, D., Gyte, G.M., et al.: Continuous cardiotocography (ctg) as a form of electronic fetal monitoring (efm) for fetal assessment during labour. Cochrane Database Syst. Rev. 3 (2006)
3. Macones, G.A., Hankins, G.D., Spong, C.Y., Hauth, J., Moore, T.: The 2008 national institute of child health and human development workshop report on electronic fetal monitoring: update on definitions, interpretation, and research guidelines. Journal of Obstetric, Gynecologic, & Neonatal Nursing 37(5), 510–515 (2008)
4. Frank, A., Asuncion, A.: UCI machine learning repository (2010)
5. Chen, C.-Y., Chen, J.-C., Yu, C., Lin, C.-W.: A comparative study of a new cardiotocography analysis program. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2009, pp. 2567–2570 (September 2009)
6. Kupka, T., Wrobel, J., Jezewski, J., Gacek, A.: Evaluation of fetal heart rate baseline estimation method using testing signals based on a statistical model. In: 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS 2006, August 30-September 3, pp. 3728–3731 (2006)
7. Marques de Sa, J.P., Reis, L.P., Lau, J.N., Bernardes, J.: Estimation and classification of fetal heart rate baselines using artificial neural networks. In: Computers in Cardiology 1994, pp. 541–544 (September 1994)
8. Chudacek, V., Spilka, J., Lhotska, L., Janku, P., Koucky, M., Huptych, M., Bursa, M.: Assessment of features for automatic ctg analysis based on expert annotation. In: 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC, August 30-September 3, pp. 6051–6054 (2011)
9. Ocak, H., Ertunc, H.: Prediction of fetal state from the cardiotocogram recordings using adaptive neuro-fuzzy inference systems. Neural Computing and Applications, 1–7 (2012)
10. Jadhav, S., Nalbalwar, S., Ghatol, A.: Modular neural network model based foetal state classification. In: 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), pp. 915–917 (November 2011)
11. Zhou, H., Ying, G.: Identification of ctg based on bp neural network optimized by pso. In: 2012 11th International Symposium on Distributed Computing and Applications to Business, Engineering Science (DCABES), pp. 108–111 (2012)
12. Huang, M., Hsu, Y.: Fetal distress prediction using discriminant analysis, decision tree, and artificial neural network. Journal of Biomedical Science and Engineering, 526–533 (2012)
13. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and Regression Trees. Wadsworth and Brooks, Monterey (1984)
14. Kohonen, T.: Self-Organizing Maps, 2nd (extended) edn. Springer, Berlin (1997)
15. Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (2001)

Part IV

Signal Recognition and Image Processing

Evolutionary Weighted Ensemble for EEG Signal Recognition

Konrad Jackowski[1,2], Jan Platos[1], and Michal Prilepok[1]

[1] IT4Innovations, VSB Technical University of Ostrava, Czech Republic
{jan.platos,michal.prilepok}@vsb.cz
[2] Department of Systems and Computer Networks, Wroclaw University of Technology,
Wroclaw, Wyb. Wyspianskiego 27, Poland
konrad.jackowski@pwr.edu.pl

Abstract. Recognition of an EEG signal is a very complex but very important problem. In this paper we focus on a simplified classification problem which consists of detection finger movement based on an analysis of seven EEG sensors. The signals gathered by each sensor are subsequently classified by the respective classification algorithm, which is based on data compression and so called LZ-Complexity. To improve overall accuracy of the system, the Evolutionary Weighted Ensemble (EWE) system is proposed. The parameters of the EWE are set in a learning procedure which uses an evolutionary algorithm tailored for that purpose. To take full advantage of information returned by sensor classifiers, setting negative weights are permitted, which significantly raises overall accuracy. Evaluation of EWE and its comparison against selected traditional ensemble algorithm is carried out using empirical data consisting of almost 5 hundred samples. The results show that the EWE algorithm exploits the knowledge represented by the sensor classifiers very effectively, and greatly improves classification accuracy.

Keywords: EEG classification, machine learning, ensemble of classifiers, evolutionary algorithms.

1   Introduction

Electroencephalography (EEG) plays a big role in the diagnosis of brain diseases, as well as in Brain Computer Interface (BCI) system applications that help disabled people to use their mind to control external devices. Both research areas are growing today.

EEG records the activity of the brain using several sensors. Different mental tasks produce indiscernible recordings which differ because different brain actions activate different parts of the brain. The most difficult part is the definition of an efficient method or algorithm for detection of the differences in recordings belonging to different mental tasks. When we define such algorithm, we are able to translate these signals into control commands for an external device, e.g. prosthesis, wheelchair, computer terminal, etc.

1.1   EEG Classification Problem

The EEG classification problem may be divided into several parts. The first problem is recording the EEG signals. We may use from 0 to hundreds of sensors as well as low or high sampling frequency, etc. The second part of the problem is preprocessing the data. EEG signals are mostly filtered using low- or high-pass filters to remove any unwanted noise. The last part is comparison of the signals themselves. Many algorithms have been suggested and tested on various types of EEG data.

Lee et al. presented a Semi-supervised version of NMF (SSNMF) which jointly exploited both (partial) labeled and unlabeled data to extract more discriminative features than standard NMF. Their experiments on EEG datasets in the BCI competition confirm that SSNMF improves clustering as well as classification performance compared to standard NMF [1].

Shin et al. have proposed a new generative model of a group EEG analysis, based on appropriate kernel assumptions on EEG data. Their proposed model finds common patterns for a specific task class across all subjects as well as individual patterns that capture intra-subject variability. The validity of the proposed method has been tested on the BCI competition EEG dataset [2].

Dohnalek et al. have proposed a method for signal pattern matching based on NMF; they also used short-time Fourier transform to preprocess EEG data and Cosine Similarity Measure to perform query-based classification. This method creates a BCI capable of real-time pattern recognition in brainwaves using low cost hardware, with a very cost efficient way of solving the problem [3].

1.2   Ensemble Classifier Systems

The main objective while designing a classification algorithm is to ensure its highest accuracy. This task is not trivial as the quality of the system is affected by many factors. There are a plethora of possibilities for elevating accuracy such as the application of feature extraction, data filtering, and fine-tuning classifiers, to mention just a few. Nonetheless, we are going to focus on another option, i.e. the application of ensemble classifier systems [4]. This concept was introduced by Chow in 1965 in his research [5], where he proves that the weighted fusion of independent classifiers can result in an optimal classification system being designed. Naturally, there are several additional issues which have to be dealt with while creating an ensemble system.

Firstly, the set of classifiers must be diverse, as the fusion of similar classifiers cannot bring any advantage. There are many ways of ensuring classifier diversity such as: collecting heterogonous classifiers (i.e. those which have different classification models), dataset partitioning, boosting [6], bagging [7, 8], or diversifying classifiers by feeding them with different features [9]. The last option is automatically applied in our case of EEG recognition as each classifiers process signals from separate sensors.

The second issue is selecting the appropriate fusing methods. The most popular one is Majority Voting [10]. Its main advantage is its very simple decision-making formula, which is based on counting votes casted by classifiers. On the other hand,

one of its major drawbacks is neglecting the qualities of voters as all of them contribute to decision-making to the same degree regardless of their accuracy. It has been shown that this approach can lead to reduced ensemble quality by spoiling its decision by weak or irrelevant classifiers.

There are two ways of dealing with this problem: (1) classifier selection, or (2) vote weighting. In the first case, only subsets of classifiers form the ensemble. There are many ways of making the selection. In the simplest one, classifiers are ordered according to their accuracy and the best ones are chosen. In the weighting strategy, all the classifiers join the ensemble, but their contribution to decision-making is weighted. The common approach suggests assigning the weights with voter accuracy [11]. Alternatively, weights can be set by a heuristic ensemble learning procedure [12] which treats learning as an optimization task which aims to minimize the misclassification rate. Among others, one can find the application of an evolutionary algorithm for that purpose [13,14].

Summarizing the discussion, we propose the application of an Evolutionary Weighted Ensemble designed with the following assumptions:

1. EWE is to exploit knowledge collected in a set of diversified elementary classifiers which make a decision by analysis data gathered by seven different EEG sensors;
2. Responses of elementary classifiers are to be weighted to adjust their contribution to decision-making;
3. Setting the weights is to be realized in a learning procedure which aims to minimize the misclassification rate. An evolutionary algorithm is to be used for that purpose.

The outline of this work is as follows: In the next section we present the EEG classification problem. Section 2 consists of information related to the EWE ensemble classifier model. Section 3 provides details on the EAE learning algorithm. Section 4 presents the evaluation results of EAE on the basis of gathered empirical material. Section 5 concludes the paper.

2   Problem Statement and EWE Model

2.1   Basic Terms

A classification algorithm Ψ assigns an object to one of M predefined classes [15].

$$\Psi : X \to \mathbf{M} \tag{1}$$

The decision is made based on an analysis set of features, denoted as x, which belongs to d-dimensional feature space X.

$$x = \left\{ x^{(1)}, ..., x^{(d)} \right\} \in \mathrm{X} \in R^d \tag{2}$$

Creating a model of the classification algorithm is done by a process of setting the appropriate model parameters in the course of training. For that purpose, a learning

set is used, i.e. a set of pairs consisting of a features and corresponding class labels [16].

$$LS = \{(x_1, j_1), (x_2, j_2), \ldots, (x_N, j_N)\} \tag{3}$$

2.2   Ensemble Classifier

In the case of the ensemble classifier system, it is also assumed that the decision is made by a set of elementary classifiers Π [11][17].

$$\Pi^{\Psi} = \{\Psi_1, \Psi_2, \ldots, \Psi_K\} \tag{4}$$

Where Ψ_k is the k-th elementary classifier in the set.

There are many variants of the ensemble decision-making formula. The most popular one is Majority Voting, which is based on counting votes cast by classifiers for each class.

$$\overline{\Psi}(x) = \arg\max_{i=1}^{M} \sum_{k=1}^{K} \delta(\Psi_k(x), i) \tag{5}$$

Where $\overline{\Psi}$ denotes the ensemble classifier, and δ states for the Kronecker delta.

2.3   EAE Decision-Making Formula

A more sophisticated approach allows weighting of the classifier responses (7)

$$\overline{\Psi}(x) = \arg\max_{i=1}^{M} \sum_{k=1}^{K} w_k \delta(\Psi_k(x), i), \tag{6}$$

where w_k are the weights assigned with the k-th classifier.

Regardless of the procedure of weights calculation, it is usually implicitly assumed that the weights have only positive values which fall within a range between 0 and 1. In the first version of EWE we decided to make the same assumption. Nevertheless, we decided to implement a second version of EWE, where the weights can get values from a broader range between -1 to 1. A detailed explanation of the reasons and consequences will be provided in section 4 along with a discussion of the obtained results.

2.4   Objective Function

EAE makes decisions based on a formula (6). We decided to set weights values in the course of the learning procedure in order to minimize the misclassification rate of the ensemble evaluated over the learning set according to (7).

$$Q^{\Psi}(LS) = \frac{1}{N} \sum_{n=1}^{N} L\left(\underset{i=1}{\overset{M}{\arg\max}} \sum_{k=1}^{K} w_k F_k(x_n, i), j_n \right) \tag{7}$$

where $L(a, b)$ denotes a loss function [15].

3   EAE Learning Algorithm

Searching for optimal values of weights is an optimization problem which cannot be solved by analytically. Therefore, we decided to apply an evolutionary-based algorithm [18]. It processes the population of possible solutions encoded in chromosomes which represent a set of weights. EWE algorithm is presented in Listing 1.

```
Listing 1 EWE training algorithm
Input: LS - learning set
S - population size
G - number of generations
Π - set of individual classifiers
Begin
  Initialize population
  For t = 1 to T do
    Evaluate Population over LS
    Select Elite
    Select parents
    Mutation
    Crossover
    Create offspring population
  End For
End Begin
```

Initialize population

At the beginning, the population of individuals is created and chromosomes are filled with random numbers with two optional constraints:

1. EAE v. 1 – chromosome constituents fall within range <0,1>,
2. EAE v. 2 – chromosome constituents fall within range <-1,1>.

Evaluate Population

Evaluation of the individuals is an operation which determines the further behavior of the algorithm. Because EWE aims to minimize the misclassification rate, a formula (7) is used for that purpose.

Select elite

In order to maintain stability of the learning process, the two best individuals with the smallest misclassification rate are selected. They are not affected by any genetic operators and are automatically transferred to the offspring population.

Select parents

Not all individuals in the population participate in creating the offspring population. The selection is based on their fitness and is done in a probabilistic manner, i.e. the chance of selection is directly proportional to the fitness of the individual. EAE uses a standard ranking selection procedure.

Mutation

The mutation operator processes the subset of the selected parents by adding some random noise to the chromosome constituents. This procedure aims to maintain diversity of the population and elevates the possibility of exploring new areas of the solution space. Keeping in mind the constraints put on weights values (section 2.3), two versions of the mutation operator were implemented.

Crossover

The standard one-point crossover operator exchanges data between two selected parents and forms two child individuals.

Create offspring population

At the end of the each generation, offspring is created by merging the elite, mutated individuals and children created by the crossover operator. The new population substitutes the previous one and the entire process is repeated.

4   Experiments

This section presents the results of the evaluation of EAE on the gathered empirical material. The following objectives of the experiments were defined:

1. to examine how creating an ensemble classifier can improve classification accuracy,
2. to examine how classifier selection and classifier weighting affects the accuracy of the ensemble,
3. to examine if EWE can outperform all elementary classifiers,
4. to compare the effectiveness of EAE with other ensemble methods,
5. to check how constraints put on weights determine EAE accuracy.

4.1   Dataset

The data for our experiments were recorded in our laboratory. We used 7 selected channels from the recorded data. The signal data contain records of the movement of one finger from four different subjects - persons. Each subject pressed a button with their left index finger. The sampling rate was set to 256~Hz. The signals were band pass filtered from 0.5~Hz to 60~Hz to remove unwanted frequencies and noise from the outside world. The data were then processed so that we could extract each movement from the data as well as 0.3 seconds before the movement and 0.3 seconds after the movement. The start and end position of the movement was marked during data recording.

The preprocessed data contain 4606 data trails - 2303 data trails with finger movement and 2303 trails without finger movement. We divided this set of trails into seven groups - one group for each sensor. Each group contains part of the training and the testing data part. The testing part contains 75% of trails with finger movement and 75% of trails without finger movement. The rest of the unused trails - with and without finger movement - were used for the training part. The training part for one sensor contains 492 trails - 246 data trails with finger movement and 246 trails without finger movement. The testing part contains 166 trails - 83 trails with finger movement and 83 trails without finger movement. The training and testing part was used for further model validation.

4.2   Experimental Framework

After recording and filtering the EEG data, we applied polynomial curve fitting for data smoothing. The fitting removes noise from the data and fits the data trend. Then we applied Turtle graphics to describe the polynomial curve with angles of rotation the turtle must do and we store this description of this movement in text files. Then we applied an LZ-complexity similarity measurement. The LZ-complexity uses a compression as an approximation of the similarity for the two data files. It uses the principle from the Lampel Ziv compression method where a dictionary of phrases is created for each texted file and these dictionaries are then compared according to the number of identical phrases. More information can be found in our previous papers [19, 20].

For each sensor from the data we created one classifier which answers the question of whether the tested sample is with or without movement. The LZ-similarity returns a value between 0 and 1, where 0 means no similarity and 1 identity. These values are gathered for each pair of files, i.e. the query and each file from the learning set. These values are then ordered according to the similarity values and the classifier calculates the ratio between the movement and no-movements samples in an interval. If the number of movement samples is larger than the number of samples without movement, the classifier decides that the query is a movement. For this purpose we find an interval for which the classifier works best on the training data. These intervals are created for data with movement and data without movement separately. These two intervals are then combined in a classifier according to the rule that we take the interval which has a higher ratio of movement or non-movement samples. The results of each classifier are then stored for the Ensemble classifier, which tries to combine the results of the classifiers for each sensor into a final value.

The EAE classifier was implemented in the MATLAB framework using Optimization Toolbox. Two versions of the EAE algorithm were implemented with two optional constraints put on weights, as described in section 2.3, i.e. EWEv1 (with weights between 0 and 1), and EWEv2 (with weights between -1 and 1).

For the comparative analysis the following traditional ensemble methods were additionally implemented: Majority Voting with all classifiers in the ensemble (MV), Majority Voting with the four best elementary classifiers selected (MV_SC), Quality-based Weighted Fuser (QWF) which uses the fusing function (6) with weights set

proportionally to the classifier accuracy, and Quality-based Weighted Fuser with the four best classifiers (QWF_SC).

4.3 Results

The results of the classifier evaluation of the testing set are presented in Figure 1. Elementary classifiers are denoted with labels starting with "Sensor cl. #".
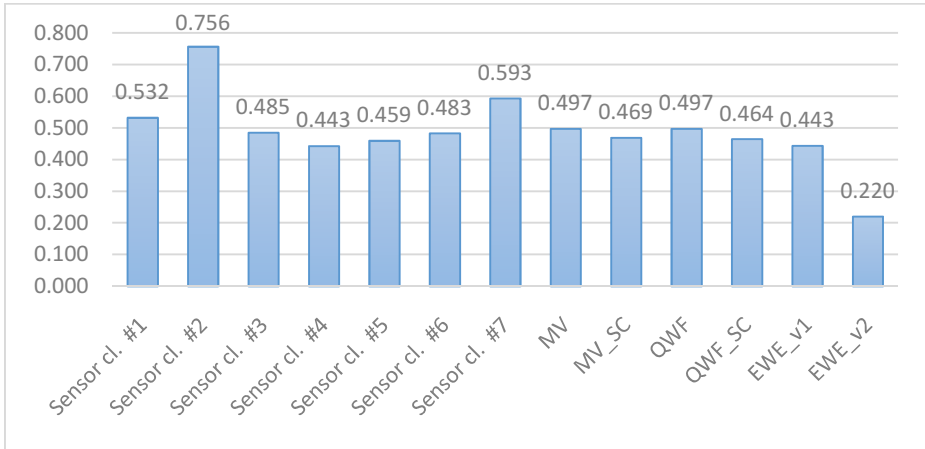
Fig. 1. Misclassification rate for EEG recognition task

Observations and discussion

1. Apparently, the set of sensor classifiers consisted of weak classifiers. Three of them (#1, #2, and #7) gained a misclassification rate higher than 50%, which makes them worse than random classifiers. The other four classifiers achieved better results, very close to 50%. The best one (#4) was able to achieve only 44% percent of misclassifications.
2. The majority voting results were not good, either (almost 50% of errors), but that should not surprise us. This kind of algorithm treats all the elementary classifiers in the same manner. Therefore, its decision can be spoiled by irrelevant and weak votes.
3. The same results as MV were obtained by QWE, which seems to be surprising because QWE has access to information about elementary classifier accuracies. An interpretation of its behavior could be as follows: all seven sensor classifiers feature a similar quality which is used for weights calculation. Therefore the weights are very similar too and do not differentiate the voters significantly. Therefore, QWE made a decision in a very similar way to MV.
4. The classifier selection procedure applied in MV_SC and QWF_SC helps only to a small degree. It allows a decrease in the error rate of about 3 percentage points in both cases. This proves that eliminating the worst classifiers reduces the possibility of spoiling a decision with irrelevant voters, but, the error rate of the ensembles still falls in the middle of the results gained by its constituent voters.

5. A slightly better result was obtained by the EWE.v1 algorithm with weights in the range between 0 and 1. It gained a 43% misclassification rate, the same result as that gained by the best of the elementary classifiers. Naturally, the result is not satisfactory, but it proves that the proposed EWE classifier is able to take everything from the available weak classifiers.
6. Comparison of the EWE.v1 and elementary classifiers triggered the conclusion that limiting classifier weights in the range 0 to 1 does not allow us to take full advantage of the knowledge collected in the set of classifiers, especially when they are weak ones. If some classifiers insist on making the wrong decision, instead of reducing their contribution to decision-making to 0, we would rather incorporate their voice with negative values. Such reasoning encouraged us to implement a second version of our algorithms (EWE.v2) with weights in a range between -1 to 1. This approach allowed us to reduce the error rate by more than half and gain only 22% of misclassifications (!).

5   Conclusions

The paper presented the application of ensemble classification algorithms to EEG signal recognition. We described popular approaches, along with a discussion of their advantages and disadvantages. An analysis of the preliminary test results allowed us to develop a new Ensemble Weighted Ensemble algorithm. A weighted voting strategy was applied with the possibility of setting the negative voters' weights. That feature allowed for a significant improvement in accuracy by more effective exploitation of the weak classifiers. An evaluation of the system was carried out on the basis of the collected empirical material. The obtained results are promising and encourage us to work on further extensions, i.e. the application of more advanced classifier fusion methods.

References

1. Jain, A.K., Duin, P.W., Mao, J.: Statistical Pattern Recognition: A Review. IEEE Trans. on PAMI 22(1), 4–37 (2000)
2. Chow, C.K.: Statistical independence and threshold functions. IEEE Transaction on Electronic Computers, EC-16, 66–68 (1965)

3. Ruta, D., Gabrys, B.: Classifier Selection for Majority Voting. Information Fusion 6, 63–81 (2005)
4. Breiman, L.: Bagging predictors. Mach. Learn. 24(2), 123–140 (1996)
5. Freund, Y.: Boosting week learning algorithm by majority. Information and Computation 121(2) (1995)
6. Freund, Y., Schapire, R.E.: Experiments with a New Boosting Algorithm. In: Proceedings of the International Conference on Machine Learning, pp. 148–156 (1996)
7. Akhand, M.A.H., Murase, K.: Ensembles of neural networks based on the alteration of input feature values. Int. J. Neural Syst. 22(1), 77–87 (2012)
8. Kuncheva, L.I.: Combining Pattern Classyfiers. Wiley-Interscience, New Jersey (2004)
9. Hashem, S.: Optimal linear combinations of neural networks. Neural Networks 10(4), 599–614 (1997)
10. Gabrys, B., Ruta, D.: Genetic algorithms in classifier fusion. Applied Soft Computing Journal 6(4), 337–347 (2006)
11. Wozniak, M., Jackowski, K.: Some remarks on chosen methods of classifier fusion based on weighted voting. In: Corchado, E., Wu, X., Oja, E., Herrero, Á., Baruque, B. (eds.) HAIS 2009. LNCS, vol. 5572, pp. 541–548. Springer, Heidelberg (2009)
12. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, New York (2001)
13. Alpaydin, E.: Introduction to Machine Learning. The MIT Press (2004)
14. Back, T., Fogel, D., Michalewicz, Z.: Handbook of Evolutionary Computation, Oxford Univ. Press (1997)
15. Lee, H., Yoo, J., Choi, S.: Semi-supervised nonnegative matrix factorization. IEEE Signal Processing Letters 17(1), 4–7 (2010), doi:10.1109/LSP.2009.2027163
16. Shin, B., Oh, A.: Bayesian group nonnegative matrix factorization for eeg analysis. CoRR abs/1212.4347, 1–8 (2012)
17. Dohnálek, P., Gajdoš, P., Peterek, T., Penhaker, M.: Pattern recognition in EEG cognitive signals accelerated by GPU. In: Herrero, Á., Snášel, V., Abraham, A., Zelinka, I., Baruque, B., Quintián, H., Calvo, J.L., Sedano, J., Corchado, E. (eds.) Int. Joint Conf. CISIS'12-ICEUTE'12-SOCO'12. AISC, vol. 189, pp. 477–485. Springer, Heidelberg (2013)
18. Prilepok, M., Platos, J., Snasel, V., Jahan, I.: Compression-based similarity in EEG signals. In: ISDA (2013)
19. Prilepok, M., Jahan, I., Snasel, V.: Electroencephalogram classification methods, Przeglad Elektrotechniczny (2013) ISSN 0033-2097, R. 89 NR 11/2013

Hierarchical Ensemble of Global and Local Classifier for Texture Classification

Ming Chen

China Three Gorges University,College of Computer and InformationTechnology,
No.8 Daxue Road, Yichang, Hubei provice, China
`ctguchen@qq.com`

Abstract. In this paper, we propose a novel hierarchical ensemble classifier for texture classification by combining global Fourier features and local Gabor features. Specifically, in our method, global features are extracted from images firstly by 2D Discrete Fourier Transform. Then, real and imaginary components of low frequency band are concatenated to form a single feature set for further processing. Gabor wavelet transform is exploited for local feature extraction. Firstly, Gabor wavelets are used to extract local features from the whole image. Then, these features are spatially partitioned into a number of feature sets, each corresponding to a local patch of the image. After the above processes, an image can be represented by one Global Fourier Feature Set (GFFS) and multiple Local Gabor Feature Sets (LGFSes). These feature sets contain different discriminative information: GGS contains global discriminative information and each LGFS contains different local discriminative information. In order to make full use of all these diverse discriminative information, we propose multiple component classifiers by applying Fisher Discriminant Analysis (FDA) on GFFS and each LGFS, respectively. At last, we combine them into one ensemble by weighted sum rule.

Keywords: component, wavelet transform, gabor, Discrete Fourier Transform, texture analysis.

1   Introduction

Feature extraction is the first stage of image texture analysis. Results obtained from this stage are used for texture discrimination, texture classification or object shape determination. Most common texture models will be shortly discussed as well. Texture classification of an image or sub-image is an important problem in texture analysis. Many procedures have been proposed. A global framework for texture classification based on random closed set theory is proposed in [1]. In this approach, a binary texture is considered as an outcome of a random closed set. Some distributional descriptors of this stochastic model are used as texture features in order to classify the binary texture, in particular spherical and linear contact distributions and K-functions.If a grayscale texture has to be classified, then the original texture is reduced to a multivariate random closed set where each component (a different

random set) corresponds with those pixels verifying a local property. Again, some functional descriptors of the multivariate random closed set defined from the texture can be used as texture features to describe and classify the grayscale texture. Marginal and cross spherical and linear contact distributions and K-functions have been used. Experimental validation is provided by using Brodatz's database and another standard texture database.

A texture representation should corroborate various functions of a texture. In [2], the authors present a novel approach that incorporates texture features for retrieval in an examplar-based texture compaction and synthesis algorithm. The original texture is compacted and compressed in the encoder to obtain a thumbnail texture, which the decoder then synthesizes to obtain a perceptually high quality texture.

The recent advanced representation for realistic real-world materials in virtual reality applications is the Bidirectional Texture Function (BTF), which describes rough texture appearance for varying illumination and viewing conditions. Such a function can be represented by thousands of measurements (images) per material sample. The resulting BTF size excludes its direct rendering in graphical applications and some compression of these huge BTF data spaces is obviously inevitable. In [3], the authors categorize, critically survey, and psychophysically compare such approaches, which were published in this newly arising and important computer vision and graphics area. In [4], the authors present a novel, fast probabilistic model-based algorithm for realistic BTF modeling allowing an extreme compression with the possibility of a fast hardware implementation. Its ultimate aim is to create a visual impression of the same material without a pixelwise correspondence to the original measurements. The analytical step of the algorithm starts with a BTF space segmentation and a range map estimation by photometric stereo of the BTF surface, followed by the spectral and spatial factorization of selected subspace color texture images. Single mono-spectral band-limited factors are independently modeled by their dedicated spatial probabilistic model. In [5], the authors use grating cell operators to obtain features and compare these operators in texture analysis tasks with commonly used feature extracting operators such as Gabor energy and co-occurrence matrix operators. For a quantitative comparison of the discrimination properties of the concerned operators a new method is proposed which is based on the Fisher linear discriminant and the Fisher criterion. The operators are also qualitatively compared with respect to their ability to separate texture from form information and their suitability for texture segmentation.

In vision and graphics, advanced object models require not only 3D shape, but also surface detail. While several scanning devices exist to capture the global shape of an object, few methods concentrate on capturing the fine-scale detail. Fine-scale surface geometry (relief texture), such as surface markings, roughness, and imprints, is essential in highly realistic rendering and accurate prediction. The authors present a novel approach for measuring the relief texture of specular or partially specular surfaces using a specialized imaging device with a concave parabolic mirror to view multiple angles in a single image [6]. Laser scanning typically fails for specular

surfaces because of light scattering, but our method is explicitly designed for specular surfaces. Also, the spatial resolution of the measured geometry is significantly higher than standard methods, so very small surface details are captured. Furthermore, spatially varying reflectance is measured simultaneously, i.e., both texture color and texture shape are retrieved.

Theoretical aspects of a technique for target detection and texture segmentation in synthetic aperture radar (SAR) imagery using a wavelet frame are presented. Texture measures consist of multi-scale local estimates of the following: 1) normalized second moment of the backscattered intensity and 2) variance of the wavelet-frame coefficients. In [7], the authors proposed an extension of a method to propose in the image-processing literature. Novel issues, which are considered in the passage to radar imagery, are the influence of speckle on texture measures afforded by the wavelet frame and their dependence on polarization states (polarimetric texture). In [8], the authors propose a texture analysis and classification approach with the linear regression model based on the wavelet transform. This method is motivated by the observation that there exists a distinctive correlation between the sample images, belonging to the same kind of texture, at different frequency regions obtained by 2-D wavelet packet transform.

The authors present an algorithm based on statistical learning for synthesizing static and time-varying textures matching the appearance of an input texture [9]. This algorithm is general and automatic and it works well on various types of textures, including 1D sound textures, 2D texture images, and 3D texture movies. The same method is also used to generate 2D texture mixtures that simultaneously capture the appearance of a number of different input textures. In our approach, input textures are treated as sample signals generated by a stochastic process.

In this paper, following the same belief to combine global and local features, we propose a novel hierarchical ensemble classifier for texture classification by combining global Fourier features and local Gabor features. Specifically, in our method, global features are extracted from whole image firstly by 2D Discrete Fourier Transform, which is strong tool to analyze images in frequency domain [10]. Then, real and imaginary components of low frequency band are concatenated to form a single feature set for further process. For local feature extraction, Gabor wavelet transform is exploited. Firstly, Gabor wavelets are used to extract local features from the whole image. Then, these features are spatially partitioned into a number of feature sets, each corresponding to a local patch of the image. After the above processes, an image can be represented by one Global Fourier Feature Set (GFFS) and multiple Local Gabor Feature Sets (LGFSes). These feature sets contain different discriminative information: GGS contains global discriminative information and each LGFS contains different local discriminative information. In order to make full use of all these diverse discriminative information, we propose to train multiple component classifiers by applying Fisher Discriminant Analysis (FDA) on GFFS and each LGFS respectively, and then combine them into one ensemble by the weighted sum rule.

The remaining part of the paper is organized as follows: in section 2, hierarchical ensemble classifier combining global and local features is introduced. In section 3, experiments and analysis are introduced. Section 4 gives references.

2   Hierarchical Ensemble Classifier Combining Global and Local Featrues

Texture is an important part of the visual world of animals and humans and their visual systems successfully detect, discriminate, and segment texture. Relatively recently progress was made concerning structures in the brain that are presumably responsible for texture processing. Neurophysiologists reported on the discovery of a new type of orientation selective neuron in areas V1 and V2 of the visual cortex of monkeys which they called grating cells. Such cells respond vigorously to a grating of bars of appropriate orientation, position and periodicity. In contrast to other orientation selective cells, grating cells respond very weakly or not at all to single bars which do not make part of a grating. Elsewhere we proposed a nonlinear model of this type of cell and demonstrated the advantages of grating cells with respect to the separation of texture and form information. In this section, we first illuminate the different roles of global and local features. Then, the detailed process of global and local feature extraction is introduced.

2.1   Different Roles of Global and Local Features

In this section, different roles of global and local features are introduced.

Global Fourier Features
2D Discrete Fourier Transform (DFT) is used to extract global facial features. An image can be transformed by 2D DFT into frequency domain as follow:

$$F(u,v) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) \exp[-j2\pi(\frac{ux}{M} + \frac{vy}{N})] \tag{1}$$

where $f(x,y)$ represents an 2D image of size M by N pixels, $0 \le u \le M-1$ and $0 \le v \le N-1$ are frequency variables. When the Fourier transform is applied to a real function, its output is complex, that is

$$F(u,v) = R(u,v) + jI(u,v) \tag{2}$$

where $R(u,v)$ and $I(u,v)$ are the real and imaginary components of $F(u,v)$ respectively. Hence, after Fourier transform, a face image is represented by the real and imaginary components of all the frequencies.

Though all the frequencies contain information about the input image, different bands of frequency play different roles. We know that generally low frequencies

reflect the holistic attributes of the input image. This can be illustrated intuitively by observing the effects of inverse transform with part of the frequency band.

Consequently, in our method, only the Fourier features in the low-frequency band are reserved as global features. Specifically, for a face image, we concatenate its real and imaginary components in the low-frequency band into a single feature set, named Global Fourier Feature Set (GFFS).

2.2   Local Gabor Features

In recent years, face descriptors based on Gabor wavelets have been recognized as one of the most successful face representation methods.Gabor wavelets are in many ways like Fourier transform but have a limited spatial scope. 2D Gabor wavelets are defined as follows:

$$\psi_{u,v}(z) = \frac{\left\|k_{u,v}\right\|^2}{\sigma^2} e^{(-\|k_{u,v}\|^2 \|z\|^2 / 2\sigma^2)} [e^{ik_{u,v}z} - e^{-\sigma^2/2}] \tag{3}$$

where $k_{u,v} = k_v e^{i\phi_u}$; $k_v = \dfrac{k_{max}}{f^v}$ give the frequency, $\phi_u = \dfrac{u\pi}{8}$, $\phi_u \in [0,\pi)$

gives the orientation. From the definition, we can see that Gabor wavelet consists of a planar sinusoid multiplied by a two dimensional Gaussian. The sinusoid wave is activated by frequency information in the image. The Gaussian insures that the convolution is dominated by the region of the image close to the center of the wavelet. That is, when a signal is convolved with the Gabor wavelet, the frequency information near the center of the Gaussian is captured and frequency information far away from the center of the Gaussian has a negligible effect. Therefore, compared with Fourier transform which extracts the frequency information in the whole image region, Gabor wavelets only focus on some local areas and extract information with multifrequency and multiorientation in these local areas.

Gabor wavelets can take a variety of different forms with different scales and orientations. Fig.1 shows 40 Gabor wavelets of 5scales and 8 orientations. It is obvious that Gabor wavelets with a certain orientation respond to the edges and bars in this orientation, and Gabor wavelets with a certain scale extract the corresponding frequency information. Hence, Gabor wavelets exhibit desirable characteristics of spatial locality and orientation selectivity.

As Gabor features are calculated b convolving Gabor wavelets with the whole face image, it covers all the positions of the face image. Thus, the local information provided by the spatial locations of Gabor features is lost when they are integrated to form one single feature vector. In order to reserve more location information, Gabor features are spatially partitioned into a number of feature sets named Local Gabor Feature Set (LGFS), each of which corresponds to a local patch of image. In addition, since each LGFS is relatively low dimensional, this can greatly facilitate the sequent feature extraction and pattern classification.

3   Hierarchical Ensemble Classifier Combining Global and Local Featrues

After feature extraction, we obtain $N+1$ feature sets, that is, one GFFS κ and ς_i $(i=1,\cdots,N)$. Then, $N+1$ classifiers can be trained by applying FDA to each feature set. As explained above, these feature sets contain different discriminant information for image recognition. Hence, the classifiers trained on these feature sets should have large diversity in error. Considering that the ensemble-based classifiers is generally superior to the single classifier when the predictions of the component classifiers have enough error diversity, we combine the classifiers trained on each feature set into a hierarchical ensemble to improve the system performance.

The hierarchical ensemble consists of two layers. In the first layer, N Local Component Classifiers (LCCs) C_{L_i} trained on ς_i $(i=1,\cdots,N)$ are combined to form a Local Ensemble Classifier (LEC) C_L, which is formulated as follow:

$$C_L = \sum_{i=1}^{N} \omega_{L_i} \cdot C_{L_i} \tag{4}$$

where ω_{L_i} is the weight of C_{L_i}. In the second layer, LEC C_L is combined with Global Classifier (GC) C_G trained on κ to form the Hierarchical Ensemble Classifier (HEC) C_H,

$$C_H = \omega_G C_G + (1 - \omega_G) C_L \tag{5}$$

where ω_G is the weight of C_G. As can be seen, in each step, sum rule, the most typical combination rule, is exploited to combine classifiers.

4   Experiments Results

In this section, we conduct series experiments on MIT, ORL, and AR databases to test our method. We conduct the first experiment on a small database which contains 40 subjects and a subject with only image. We divide each image into four sub-images, and then we can obtain 160 sub-images in all.Fig.1 shows some subjects of MIT database. Table 1 shows the classification rate on the small database.

Table 1. The Classification Rate of Different Methods on the MIT Database

Method	Gabor	LBP	Our method
Classification rate	50.1%	82.50%	**92.16%**
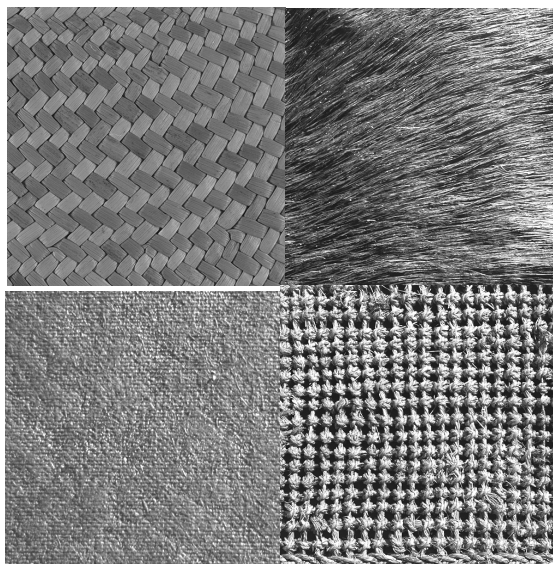
Fig. 1. Some Subjects of MIT Database

The ORL is one of the most popular face image databases.This database contains ten face images each for 40 different people. In order to provide suitable research material, the images of this database were taken at different times, and in various lighting. To model the faces in daily life, the faces had different expressions (open/closed eyes, smiling/not smiling and some of them were facilitated with details (glasses/no glasses). In our experiment, we choose the first 5 samples as training samples, and the rest as testing samples. Table 2 shows the result of some methods on ORL database. Fig. 2 depicts the ten face images of a subject in the ORL database.

Fig. 2. Samples From the ORL Database

Table 2. The Classification Rate of Different Methods on the ORL Database

Method	Gabor	LBP	Our method
Classification rate	88.24%	90.12%	**95.34%**

The AR face database contains over 4000 gray face images of 126 people. The images of each subject have different facial expressions, and were acquired under lighting conditions and with and without occlusions. Each subject provided 26 face images. We note that 12 face images of each subject are occluded with sunglasses or a scarf. The face images of 120 subjects were taken in two sessions. We used only the images of these 120 subjects in our experiment. We manually cropped the face portion of every image and then normalized them to 50 ×40 pixels [11].

Fig. 3 shows the normalized images of one subject. We used only the 14 non-occluded face images of each subject to test different solution schemes. The first and eighth images were used as training samples and the remaining images were used as testing samples. Table 3 shows experimental result of some methods on AR database.

Table 3. The Classification Rate of Different Methods on the AR Database

Method	Gabor	LBP	Our method
Classification rate	67.4%	66.12%	**76.22%**

Fig. 3. Normalized non-occluded face images with 50 □ 40 pixels of one subject. (a) neutral expression, (b) smile, (c) anger, (d) scream, (e) left light on, (f) right light on,(g) all sides light on. (h), (i), (j), (k), (l), (m) and (n) were taken in the second session under the same conditions as (a), (b), (c), (d), (e), (f) and (g), respectively.

5   Conclusions

In this paper, we propose a novel hierarchical ensemble classifier for texture classification by combining global Fourier features and local Gabor features. Specifically, in our method, global features are extracted from whole image firstly by 2D Discrete Fourier Transform. Then, real and imaginary components of low frequency band are concatenated to form a single feature set for further process. For local feature extraction, Gabor wavelet transform is exploited. Firstly, Gabor wavelets are used to extract local features from the whole image. Then, these features are spatially partitioned into a number of feature sets, each corresponding to a local patch of the image. After the above processes, an image can be represented by one Global Fourier Feature Set (GFFS) and multiple Local Gabor Feature Sets (LGFSes). These

feature sets contain different discriminative information: GGS contains global discriminative information and each LGFS contains different local discriminative information. In order to make full use of all these diverse discriminative information, we propose to train multiple component classifiers by applying Fisher Discriminant Analysis (FDA) on GFFS and each LGFS respectively, and then combine them into one ensemble by the weighted sum rule.

References

1. Irene, E., Guillermo, A.: A Random Set View of Texture Classification. IEEE Trans. Image Processing 11(8), 859–868 (2002)
2. Paruvelli, S., Liang, H.W., Wmin, L.S.: An Examplar-Based Approach for Texture Compaction Synthesis and Retrieval. IEEE Trans. on Image Processing 19(5), 1307–1319 (2010)
3. Jiri, F., Michal, H.: Bidirectional Texture Function Modeling: A State of the Art Survey. IEEE Trans. on PAMI 31(11), 1921–1941 (2009)
4. Michal, H., Jiri, F.: Extreme Compression and Modeling of Bidirectional Texture Function. IEEE Trans. on PAMI 29(10), 1859–1866 (2007)
5. Peter, K., Nilolay, P.: Nonlinear Operator for Oriented Texture. IEEE Trans. on Image Processing 8(10), 1395–1408 (1999)
6. Jing, W., Kristin, J.: Relief Texture from Specularities. IEEE Trans. on PAMI 28(3), 446–457 (2006)
7. Gianfranco, D.G., Sen, L.J., Dale, L.S.: Target Detection and Texture Segmentation in Polarimetric SAR Images Using a Wavelet Frame: Theoretical Aspects. IEEE Trans. on Geoscience and Remote Sensing 45(11), 3437–3454 (2007)
8. Zhong, W.Z., Hai, Y.J.: Texture Analysis and Classification With Linear Regression Model Based on Wavelet Transform. IEEE Trans. on Image Processing 17(8), 1421–1431 (2008)
9. Bar, J.Z., Ei, Y.R., Dani, L., Michael, W.: Texture Mixing and Texture Movie Synthesis Using Statistical Learning. IEEE Trans. on Visualization and Computer Graphics 7(2), 120–136 (2001)
10. Lai, J.H., Yuan, P.C., Feng, G.C.: Face Recogntion Using Holistic Fourier Invariant Features. Pattern Recognition 34(1), 95–109 (2001)
11. Yang, J., Zhang, D., Frangi, A.F., Yang, J.Y.: Two dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Trans. Pattern Anal. Machine Intell. 24(1), 131–137 (2004)

Pedestrian Detection Using HOG Dimension Reducing in Video Surveillance

Kebin Huang, Feng Wang, Xiaoshuang Xu, and Yun Cheng

Department of Digital Media Technology, Huanggang Normal University,
438000 Huangzhou, China
{huangkebin,mwfwf,xxshuang}@hgnu.edu.cn

Abstract. Pedestrian detection draws a mount of attention in these years. However, most of the classify-based pedestrian detection methods are facing huge training samples and high computation complexity. In this paper, it proposed a manifold learning based pedestrian detection method. First, modeling the video surveillance scene via mixed gaussian background model and collecting negative samples from the background images; Second, extract the positive and negative samples histogram of oriented gradients(HOG) features, using the local preserving projection(LPP) for dimensionality reduction; Finally, detecting the pedestrian from the input image under the framework of AdaBoost. Experiments show that the algorithm achieved good results both in speed and accuracy of pedestrian detection.

Keywords: Pedestrian detection, Manifold learning, HOG, AdaBoost.

1   Introduction

Pedestrian detection is an important research field of computer vision and pattern recognition, and has wide range applications in intelligent transportation, human-computer interaction, video search and video surveillance and other fields.By detecting, tracking, trajectory analysis and behavior recognition, real-time video surveillance system can detect abnormal events and alarms, and achieving the intelligent video surveillance system. The accuracy of detection and location will directly affect the performance of the entire system. However, in practical monitoring system, pedestrians are far away from the camera, images of pedestrians in video surveillance usually have low resolution and seriously affected by noise, light, background and other factors. These existing problems make the outdoors pedestrian detection becomes a challenging task.

Many researchers have been focused on pedestrian detection under visible light conditions. Pedestrian detection method based on the classification is the mainstream approach [3-8]. The sliding window-based classification method is proved to be the best method for pedestrian detection [4,6].

Such method first generates the sliding window of different scales and position in the testing image, extracting a particular feature of the window image, such as edges,

wavelet coefficients, and so on; then use offline trained classifier to classify the extracted feature, and judge whether the testing window contains the pedestrian. The proper feature and classifier are the two basic elements of such methods. The methods used for classifier training including neural networks, support vector machines [3] and Adaboost [14] and so on. The common features include target motion features and appearance features. The former mainly refers to the periodic characteristics of the target pedestrian movement [9,10]. The later includes the gray degree, shape [10,11], HOG[3,12,13], Harr wavelet feature [1], SIFT features[4]and other information.

From the view of machine learning, pedestrian detection system is a high dimensional feature space. Therefore, in order to reduce redundancy features, get the intrinsic characteristics, dimension reduction techniques have been used in recent years. From the global features reduction view, Munder and Gavrila [7] proposed to use linear PCA method for feature reduction. From the perspective of local image features, Kobayashi et al [12] firstly extract gradient histogram feature. Viola et al [10] first combined motion and appearance information together for training a strong cascade classifier using Adaboost method. It achieved a good low-resolution pedestrian detection under rain and snow conditions.

The dimensionality reduction methods mentioned above are linear methods. As the pedestrian feature space is likely to be non-linear in nature, so the nonlinear dimensionality reduction methods are worth trying. In this paper, LPP method is used for dimension reduction of pedestrian HOG feature. Also, the manifold method is expected to be effective for finding low-dimensional manifold structure embedded in high-dimensional data space.

Based on the ideal above, it proposed a manifold learning based pedestrian detection method. First, modeling the video surveillance scene via mixed gaussian background model and collecting negative samples from the background images; Second, extract the positive and negative samples HOG features, using the LPP for dimensionality reduction; Finally, detecting the pedestrian from the input image under the framework of AdaBoost. Experiments validate the proposed method.

2   Pedestrian Detection via Manifold Learning in Video Surveillance

2.1   Locality Preserving Projection

Locality preserving projection (LPP) is a local subspace learning algorithm. It aims to maintain the geometric characteristics and the local features of the observation data in dimensions reduction. Compared to other manifold learning methods, such as Isomap, LLE, or Laplacian Eigenmap, LPP can not only find projection on training data, but also can project to new test sample absent from the training set. Here give a brief introduction of LPP[14]. Suppose d-dimensional data set $X = \{x_1, x_2, ..., x_m\}$ distributed on submanifolds of a high-dimensional space, LPP aims to find an optimal transformation matrix A, mapping X into n-dimensional data set $Y = \{y_1, y_2, ..., y_m\}$,

that is, x→y = A^Tx, where d << n, m is the number of data samples. The objection function of LPP is

$$\arg\min_a \sum_{ij} (a^T x_i - a^T x_j)^2 W_{ij}$$ (1)

Where W_{ij} is the connection weights between points x_i and x_j, it is defined as

$$W_{ij} = W_{ji} = \begin{cases} \exp(-\|x_i - x_j\|^2 / t), \|x_i - x_j\|^2 < \varepsilon \\ 0, \text{ others} \end{cases}$$ (2)

In formula (2), ε>0 defined the range of local neighbors, t is constant. Minimize the objective function of LPP is to ensure that when the distance between points x_i and x_j is minimum, so the distance between y_i and y_j is minimum too. Formula (2) can be changed into the form below

$$a = \arg\min_{a^T XDX^T a = 1} \sum_{ij} (a^T x_i - a^T x_j)^2 W_{ij} = \arg\min_{a^T XDX^T a = 1} a^T XLX^T a$$ (3)

Where, X={$x_1, x_2, \ldots, x_m$}, D is diagonal matrix, $D_{ij} = \sum_j W_{ij}$ is the sum of the column in W, L=D-W is Laplace matrix.

By the Lagrange multiplier method, formula (3) transformed into solving matrix eigenvalue problem as

$$XDX^T a = \lambda XDX^T a$$ (4)

By using the local structure adjacent nature, locality preserving projection algorithm can get lower dimension feature data, which is suitable for fast sample searching in local similarity. Compared to the method searching by original pixel value or probability, the LPP based block matching method have faster block picking ability and better learning effects.

2.2 Feature Selection

Features of the image are very rich, including color, brightness, statistic features and so on. Since the luminance and color feature in pedestrian images are ever-changing, therefore the most suitable pedestrian detection feature is the gradient direction histogram feature, which is more robust to others and can reflect the body contour better. Here we adopt HOG for pedestrian detection. The specific algorithm of HOG is as follows [3]:

First, take the gamma adjustment and smooth filter to the input image. Gamma adjustment is to increase the intermediate level of the image while don't have much impact on the hierarchy of dark and bright portions. The use of Gaussian filter is:

$$L(x, y) = G(x, y, \sigma) \times I(x, y) \quad G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2 + y^2)} / 2\sigma^2$$ (5)

Where, G(x,y,σ) is a Gaussian function; (x,y) is coordinates; σ is the standard coordinates; L(x,y) is the smoothed image of the original image after convolution, σ determines the degree of smoothing.

Second, calculate the gradient magnitude and direction of each pixel point in the smoothed image. Suppose I(x,y) is the image, the horizontal gradient $G_x(x, y)$ and vertical gradient $G_y(x, y)$ in pixel (x,y) are calculated as follows:

$$G_x(x, y) = I(x+1, y) - I(x-1, y) \quad G_y(x, y) = I(x, y+1) - I(x, y-1) \tag{6}$$

The size of the gradient and the orientation in sample points (x, y) are defined as:

$$G(x, y) = \sqrt{G_x(x, y)^2 + G_y(x, y)^2} \quad \alpha(x, y) = \tan^{-1}\left(G_y(x, y)/G_x(x, y)\right) \tag{7}$$

Finally, calculate the HOG value by weighted the gradient modulus and direction of the image. HOG feature describes the distribution of gradient strength and gradient direction in local regions of an image. The distribution can well characterize the local appearance and shape. The HOG method has become the mainstream in pedestrian detection. Dalal et al[3] represents a detection window through 16x16 pixel HOG features and achieved good results, here we follow the same block division as Dalal. The gradient of HOG calculation method is similar to formula (7).

2.3   Classifier Training

The performance of pedestrian detection algorithm not only affects by the extracted features, but also affects by the detector performance. There are two categories classification mechanism used in the field, named single classifier and combined classifiers. Combined classifier mainly is cascade classifier. It is formed by a lot of class classifier, only passed the former one; it can get into the later one. The Adaboost algorithm can make a group of weak classifiers combined into a strong classifier.

As the support vector machine is the most common classification algorithm, with a strong theoretical support, and get a better validation in applications in other areas, we take the support vector machine as a weak classifier here. The objection function of the support vector machine [13] is

$$\min \frac{1}{2}\|\omega\|^2 + C\sum Z_i \quad \text{s. t.} \quad y_i(\omega \cdot x - b) \geq 1 - Z_i \tag{8}$$

Where ω is related vector used to separate hyperplane, $\|\omega\|^2$ is super flat border, Z_i is a slack variable.

In this paper, we adopt radial basis function (RBF) as the product functions:

$$K(x, x_i) = \exp\left\{-\|x - x_i\|^2 / 2\sigma^2\right\} \tag{9}$$

Selecting the support vector machine as a weak classifier can avoid local optima, and easy to use by setting less artificial parameter. In order to implement the coarse-to-fine strategy of pedestrian detection, Adaboost cascade classifier is adopted.

The detailed algorithm of Adaboost cascade classifier is described as follows:

Algorithm of Adaboost cascade classifier

1. Input: the training sample set $S = \{(x_1, y_1), \cdots, (x_N, y_N)\}$, $x_i \in X$ is a vector of samples, it represents the sample image; $y_i \in Y = \{0,1\}$, is the category labels; N is the total number of samples.

2. Initialization right of training sample by

$$w_{a,b} = \begin{cases} 1/2b, & \text{if } y_i = 1 \\ 1/2a, & \text{if } y_i = 0 \end{cases}$$

a, b are the number positive and negative samples separately.

3. For t = 1, ..., T (T is the number of weak classifiers to be selected):

   1) Normalized weights, $w_{t,i} \leftarrow w_{t,i} / \sum_{j=1}^{N} w_{t,j}$

   2) Train weak classifier $h_j(x)$, calculate the error by

   $$\varepsilon_j = \sum_{i=1}^{m} w_{t,i} |h_j(x_i) - y_i|$$

   3) Select a weak classifier h_t with the minimum weighted error ε_j;

   4) Update sample weights: $w_{t+1,i} = w_{t,i} \beta^{e_{t,i}}$

   $$e_{t,i} = \begin{cases} 1, & h_j(x_i) = y_i \\ 0, & h_j(x_i) \neq y_i \end{cases}, \quad \beta_t = \varepsilon_t /(1-\varepsilon_t)$$

4. Output: The final strong classifier.

---

3 Experimental Results

Based on the core algorithm, we developed a real-time pedestrian detection system using Matlab2009. The hardware system is: Intel Dual-Core CPU 3.2 GHz processor, 2G memory. Different video surveillance scenes are used as a sample source. By hand-cut way, we take1000 pedestrian images, then by reversal, translation and other operations, we get 4510 positive samples. All positive samples are normalized to 64 * 32 pixels.

Fig.1 (a) gives parts of the positive samples. Use GMM method to extract the video backgrounds of different scenarios, obtaining a total of 80 background images with size of 352 * 288 pixels. By sliding segmentation on the background, we established 4510

negative samples with the size of 64*32 pixels. In order to improve the refusal rate of negative samples, some background images with vehicles are used for negative samples learning. Fig. 2 (b) shows parts of the background images used for negative samples extracting.

(a)   Parts of the positive samples

(b)   Parts of the background images used for negative samples extracting

Fig. 1. Positive samples and background images

In order to verify the effectiveness of the proposed algorithm in this paper, we compared the results with methods proposed in literature [4] and [16], in terms of computational complexity and detection accuracy. Detection accuracy is measured by the true and false detection rate. True detection rate refers to the ratio of detecting pedestrian number to the total number of pedestrian. False detection rate refers to the ratio of the number of false detection to the total number of pedestrians.

3.1   Comparison to the Computational Complexity

In the experiment, libSVM is chosen as weak classifier. Classifier training iterated 100 times, choose the 21- hierarchical classifier. Table 1 shows the calculation time by

selecting a weak classifiers. It can be seen that by using pedestrians dimensionality reduction of the sample's HOG features, the time required to train classifiers greatly reduced. After the use of LPP dimensionality reduction, classifier training time is the least, only 1.64% of time required to directly entire feature space. Experimental results show the effectiveness of LPP dimensionality reduction method to improve pedestrian detection classifier training, creating the conditions for the real-time classifier training.

Table 1. Time of weak classifier choosing

Method of Weak classifier choosing	Computation time(s)	Relative time(%)
The whole feature space	About a week	100
PCA （400 dimension）	154098	25.48
LPP （70 dimension）	9891	1.64

3.2   Comparison to Detection Accuracy

We choose 4800 samples as training samples randomly, where positive and negative samples are 2400 separately. The test samples' number is 220, of which 110 positive and 110 negative test samples. For each sample, the original dimension is 3780. Table2 show the detection accuracy and time in different dimensions. Seen the table, we can found that compared with the direct traverse the entire feature space, the use of LPP not only reduce the detection time, but also can improve the detection accuracy rate.

Table 2. Detection accuracy and time in different dimensions

LPP dimensions	True detection rate	False detection rate	Detection time
10	100%	1.8182%	0.022487+ 0.015110
30	100%	0. 9091%	0.040522+ 0.027963
50	100%	0. 9091%	0.059236+0.026615
70	100%	0	0.061586+0.040336
90	100%	0	0.166117+0.106886
3780	94.5455%	10%	22.035859+22.288748

The results of pedestrian detection in practical video surveillance scenes can be seen in figure2. The experimental results validate the proposed method.

Fig. 2. Pedestrian detection in video surveillance scenes

4   Conclusions

This paper proposed a pedestrian detection using HOG dimension reducing in video surveillance. Two measures are adopted to improve pedestrian detection accuracy. One is using LPP method to reduce the feature dimension and extract the internal structure, which improve the recognition rate of positive samples. The other is using background modeling and negative samples, which can improve the rejection rate of negative samples. At the meanwhile, the proposed method has less computational complexity. Experimental results demonstrate the effectiveness of the algorithm.

References

1. Oren, M., Papageorgiou, C., Sinha, P., et al.: Pedestrian detection using wavelet templates. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 193–199. IEEE, Washington (1997)
2. Wu, Y., Yu, T., Hua, G.: A statistical field model for pedestrian detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 1023–1030. IEEE, Washington (2005)
3. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 886–893. IEEE, Washington (2005)
4. Xu, Y., Cao, X., Qiao, H.: An Efficient Tree Classifier Ensemble-Based Approach for Pedestrian Detection Systems. IEEE Transactions on Man, and Cybernetics, Part B: Cybernetics 41(1), 107–117 (2011)

5. Zhang, J.: An overview of fast pedestrian detection: Feature selection and cascade framework of boosted features. In: IEEE International Conference on Multimedia and Expo, pp. 1566–1567 (2009)
6. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian Detection: An Evaluation of the State of the Art. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI (2012)
7. Munder, S., Gavrila, D.M.: An experimental study on pedestrian classification. Transactions on Pattern Analysis and Machine Intelligence 28(11), 1863–1868 (2006)
8. Papageorgiou, C., Poggio, T.: A Trainable System for Object Detection. Int'l J. Computer Vision 38(1), 15–33 (2000)
9. Elzein, H., Lakshmanan, S., Watta, P.: A Motion and Shape-Based Pedestrian Detection Algorithm. In: Proc. IEEE Intelligent Vehicle Symp., pp. 500–504 (2003)
10. Viola, P., Jones, M., Snow, D.: Detecting Pedestrians Using Patterns of Motion and Appearance. In: Proc. Int'l Conf. Computer Vision, pp. 734–741 (2003)
11. Cutler, R., Davis, L.: Robust real-time periodic motion detection: Analysis and applications. In: IEEE Patt. Anal. Mach. Intell., vol. 22, pp. 781–796
12. Kobayashi, T., Hidaka, A., Kurita, T.: Selection of Histograms of Oriented Gradients Features for Pedestrian Detection. In: Ishikawa, M., Doya, K., Miyamoto, H., Yamakawa, T. (eds.) ICONIP 2007, Part II. LNCS, vol. 4985, pp. 598–607. Springer, Heidelberg (2008)
13. Qiang, Z., Shai, A., Chen, Y.: Fast human detection using a cascade of histograms of oriented gradients. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, New York, vol. 2, pp. 1491–1498 (2006)
14. He, X., Niyogi, O.: Locality preserving projections. In: Advances in Neural Information Processing Systems, vol. 16 (2003)

Reversible Watermarking Based on Position Determination and Three-Pixel Block Difference

Shaowei Weng[1], Jeng-Shyang Pan[2], and Tien-Szu Pan[3]

[1] School of Information Engineering,
Guangdong University of Technology, P.R. China
`wswweiwei@126.com`
[2] Harbin Institute of Technology
Shenzhen Graduate School, P.R. China
`jengshyangpan@gmail.com`
[3] Department of Electronic Engineering,
National Kaohsiung University of Applied Sciences

Abstract. Reversible watermarking based on position determination and three-pixel block difference is proposed in this paper. In the proposed method, for a three-pixel block, no modification is allowed to its center pixel (CP). This unchanged pixel along with all the neighbors surrounding this block constitute a set used for evaluating the intra-block correlation. The incorporation of CP in this set helps to largely enhance the estimation accuracy. According to the strength of correlation, we determine this block into a smooth or complex region. When the desired embedding rate is low, we only modify those blocks located in smooth regions while keeping the others unchanged. Therefore, the PSNR (peak signal to noise ratio) value is largely increased. Experimental result also demonstrate that the proposed method is effective.

1 Introduction

For some critical applications such as the fields of the law enforcement, medical and military image system, it is crucial to restore the original image without any distortions. The watermarking techniques satisfying these requirements are referred to as the reversible watermarking. The reversible watermarking is designed so that it can be removed to completely restore the original image.

The concept of reversible watermark firstly appeared in the patent owned by Eastman Kodak [1]. Several researchers had developed reversible watermarking [2–13]. RW based on difference expansion was first proposed by Tian [2]. Difference (between a pair of neighboring pixels) is shifted left by one unit to create a vacant least significant bit (LSB), and 1-bit watermark is appended to this LSB. This method is called Difference Expansion (DE). Alattar [3] generalized the DE technique by taking a set containing multiple pixels rather than a pair. Coltuc *et al.* proposed a threshold-controlled embedding scheme based on an integer transform for pairs of pixels [4]. In Thodi's work [5], histogram shifting was incorporated into Tian's method to produce a new algorithm called Alg.

D2 with a overflow map. Weng *et al.* proposed an integer transform based on invariability of the sum of pixel pairs [6]. Weng *et al* also proposed a new integer transform, and the embedding rate can approach to 1 bpp (bit per pixel) for a single embedding layer by overlapped pairing [7]. In Wang *et al.*'s method [8], a generalized integer transform and a payload-dependent location map were constructed to extend the DE technique to the pixel blocks of arbitrary length. Feng *et al.* presented a new reversible data hiding algorithm based on integer transform (proposed by [9]) and adaptive embedding [10]. Luo *et al.* first introduced an interpolation technique into RW to obtain the prediction-errors, which increased prediction-accuracy by using full-enclosing pixels to predict the current pixel [11]. In Li *et al.*'s method [12], an efficient reversible watermarking scheme was proposed by incorporating in prediction-error expansion two new strategies, namely, adaptive embedding and pixel selection. Wu *et al.* proposed reversible image watermarking on prediction errors by efficient histogram modification [13].

Besides these works, a new category of RW schemes has been proposed by Li *et al.* in [14–16]. Their advantage lies in that they can achieve high visual quality at some desired low ER, though their obtained maximum embedding capacity is only able to satisfy some practical application, e.g., the application of RW in medical image sharing (referring to the papers in [17]).

In Lin *et al*'s method [18], the images are divided into non-overlapping three-pixel blocks. Each block has two pairs, and each pair is composed of two neighboring pixels. The absolute difference between each pair is calculated. 1-bit watermark is embedded into a pair whose absolute difference has the highest occurrence rate. The main purpose of this paper is to further increase the visual quality at low embedding rates. The embedding process in Lin *et al*'s method is modified so as to keep the CP in some block unaltered. Therefore, CP can be used for evaluating the intra-block correlation. Specifically, it forms a set along with all the adjacent pixels surrounding this block. Based on the fact that adjacent pixels in natural images tend to have similar intensity levels, the correlation of this set is essentially consistent with the intra-block correlation. The incorporate of CP in this set is to make the correlation of this set closer to the intra-block correlation. Thus, we use the variance of this set to represent the intra-block correlation. Position determination is the process that we can determine this block in a smooth or complex region, according to the amplitude of the variance. When the desired embedding rate is low, we only modify those blocks located in smooth regions while keeping the others unchanged. Therefore, the PSNR value is largely increased. Experimental results also demonstrate the proposed method is effective.

The remains of the paper are organized as follows. In Section 1.1, the proposed method is introduced. Watermark embedding and data extracting and image restoration are presented in Section 1.2 and 1.3, respectively. The experimental results are shown in Section 2 and finally we conclude the paper in Section 3.

1.1   The Proposed Method

Let $p = (p_1, p_2, p_3)$ denote a pixel block containing three neighboring pixels. For any p, two absolute differences (denoted by d_1, d_2) between two neighboring pixels are calculated, i.e., $d_1 = |p_1 - p_2|$, $d_2 = |p_2 - p_3|$. Given a $\mathtt{W} \times \mathtt{H}$-size image, it is partitioned into non-overlapping three-pixel blocks. So, the number of differences occupys nearly two-thirds of all the pixels. The main purpose of this paper is to further increase the visual quality by evaluating each block whether it is located in the smooth or complex region. For any p, we do nothing with p_2 (namely CP), and therefore, p_2 can be used for estimating the intra-block correlation. Its corporation helps to further enhance the estimation accuracy. Specifically, the variance of p_2 and all the neighbors surrounding p is employed to evaluate the intra-block correlation. Thus, the smaller the variance, the stronger the intra-block correlation. We only modify those blocks located in smooth regions while keeping the others unaltered. The advantage of doing so is that we can decrease the embedding distortions at the low embedding rates.

Let $g(d)$ denote the number of differences whose absolute values equal d, where $0 \le d \le 253$. M_d and m_d are respectively used to represent the absolute values of differences having the largest and the small occurrence, i.e., $g(M_d) \ge g(M^{'})$ and $g(m_d) \le g(m^{'})$ for $0 \le M^{'}, m^{'} \le 253$. In Lin *et al.*'s method, 1-bit watermark is embedded into difference whose absolute value is equal to M_d. For a pixel block p, if $d_1 = M_d$ and $d_2 = M_d$, then this p is capable of carrying two-bit watermark information. In the proposed method, in order to more accurately determine this block in a smooth or complex region, we modify Lin *et al.*'s method to keep p_2 unaltered. Take p having $d_1 = M_d$ and $d_2 = m_d$ for example, there exists five types of relationships (i.e., $p_1 > p_2 > p_3$, $p_1 < p_2 < p_3$, $p_1 = p_2 = p_3$, $p_1 < p_2 > p_3$ and $p_1 > p_2 < p_3$) among p_1, p_2 and p_3. Suppose $p_1 > p_2 > p_3$. Since $d_1 = M_d$ and $d_2 = M_d$, after watermark embedding, $p_1^{'} = p_1 - w_1$, $p_2^{'} = p_2$ and $p_3^{'} = p_3 - w_2$, where w_1 and w_2 are respectively used to denote 1-bit watermark. Note that the relationship among $p_1^{'}$, $p_2^{'}$ and $p_3^{'}$ is identical to that of p_1, p_2 and p_3, i.e., $p_1^{'} > p_2^{'} > p_3^{'}$. $d_1^{'}$ and $d_2^{'}$ are respectively used to denote the watermarked differences of d_1 and d_2. After watermark embedding, $d_1^{'} = p_1^{'} - p_2^{'} = p_1 - p_2 + w_1 = d_1 + w_1$, $d_2^{'} = p_2^{'} - p_3^{'} = p_2 - p_3 + w_2 = d_2 + w_2$. Hence, in the extraction process, by comparing $d_1^{'}$ (or $d_2^{'}$) with M_d and m_d, 2-bit watermark can be correctly extracted. If $p_1 < p_2 > p_3$, and both w_1 and w_2 are 1, in order to ensure the invariability of p_2, p_1 and p_3 must be respectively modified as $p_1^{'} = p_1 - w_1$ and $p_3^{'} = p_3 - w_2$. Detailed modification is illustrated in Table 1.

1.2   Watermark Embedding

I is converted into a one-dimension pixel list respectively according to according to a predefined order. Specifically, we work the first row from left to right, the second row from right to left, then continue alternating row direction in this

manner, so that a one-dimensional pixel array I_{D1} is created. Three consecutive pixels p_1, p_2 and p_3 of I_{D1} is grouped into a pixel block $p = (p_1, p_2, p_3)$, where $0 \leq p_1 \leq 255$, $0 \leq p_2 \leq 255$ and $0 \leq p_3 \leq 255$.

Each p has two absolute differences d_1 and d_2. If d_1 (or d_2) equals M_d, then 1-bit watermark is embedded into d_1 (or d_2) in the watermark embedding process. d_1 is divided into three intervals: $d_1 = M_d$, $m_d \geq d_1 > M_d$ and $d_1 < M_d$ or $d_1 > m_d$ according to its value. Similarly, d_2 is also divided into three intervals: $d_2 = M_d$, $m_d \geq d_2 > M_d$ and $d_2 < M_d$ or $d_2 > m_d$. d_1 and d_2 are combined together to create nine combinations. Nine combinations respectively correspond to nine different embedding strategies, which are included in the flow chart of watermark embedding procedure (WEP) (see Fig. 1).

```
if d1 == Md
    if d2 == Md
        call embed_2_bits;
    else
        if Md < d2 ≤ md
            call embed_1_bit_and_increase_difference;
        else
            call embed_1_bit_and_leave_unchanged;
        end
    end
elseif Md < d1 ≤ md
    if d2 == Md
        call increase_difference_and_embed_1_bit;
    else
        if Md < d2 < md
            call increase_2_difference;
        else
            increase_difference_and_leave_unchanged;
        end
    end
else
    if d2 == Md
        call leave_unchanged_and_embed_1_bit;
    else
        if Md < d2 < md
            call leave_unchanged_and_increase_difference;
        else
            Do nothing;
        end
    end
end
```

Fig. 1 WEP

When $m_d < d_1$ or $d_1 < M_d$ and $m_d < d_2$ or $d_2 < M_d$, we do nothing with p. Besides, the remaining eight strategies achieve the WEP by calling their respective embedding functions. Eight functions are illustrated respectively in Tables 1, 3, 5, 7, 9, 11, 13, 15. Each of the eight tables tells us how to modify p in different types of relationships among p_1, p_2 and p_3. For any of eight tables, its corresponding one after watermark embedding is produced to tell us the range of $d_1^{'}$ (or $d_2^{'}$) and the relationship among the pixels of p. By means of Tables 2, 4, 6, 8, 10, 12, 14, 16, p_1 (or p_3) can be correctly retrieved in the extraction process by getting the range of $d_1^{'}$ (or $d_2^{'}$).

A location map L_M is generated and denoted by a bit sequence of size $W \times H$. For a pixel block p, if $0 \leq p_1^{'} \leq 255$, $0 \leq p_2^{'} \leq 255$ and $0 \leq p_3^{'} \leq 255$ after performing WEP, this p is marked by '1' in the map L_M. otherwise by '0'. The location map is compressed losslessly by an arithmetic encoder and the resulting bitstream is denoted by $\mathcal{L}$. L_S is the bit length of $\mathcal{L}$. If p is marked by '1' in L_M, we first select the corresponding WEP from Procedure 1 to Procedure 8 according to the eight combination of d_1 and d_2, and then, obtain $p_1^{'}$, $p_2^{'}$ and $p_3^{'}$ based on the relationship among p_1, p_2 and p_3. If p is marked by '0' in L_M, p is kept unchanged.

After the first L_S pixels have been processed, their LSBs are replaced by $\mathcal{L}$, and then the L_S LSBs to be replaced are appended to $\mathcal{P}$. Finally, the rest of $\mathcal{P}$ and the L_S appended bits are embedded into the remaining unprocessed blocks. After all the blocks are processed, a new watermarked image I_W is obtained.

Table 1. Procedure 1: Embedding_two_bits

Relationships among p_1, p_2 and p_3	Relationships among $p_1^{'}$, $p_2^{'}$ and $p_3^{'}$	Modifications to p_1 and p_3
$p_1 > p_2 > p_3$	$p_1^{'} > p_2^{'} > p_3^{'}$	$p_1^{'} = p_1 + w_1, p_3^{'} = p_3 - w_2$
$p_1 < p_2 < p_3$	$p_1^{'} > p_2^{'} > p_3^{'}$	$p_1^{'} = p_1 - w_1, p_3^{'} = p_3 + w_2$
$p_1 = p_2 = p_3$	$p_1^{'} \neq p_2^{'} \neq p_3^{'}$	$p_1^{'} = p_1 - w_1, p_3^{'} = p_3 - w_2$
$p_1 < p_2 > p_3$	$p_1^{'} < p_2^{'} > p_3^{'}$	$p_1^{'} = p_1 - w_1, p_3^{'} = p_3 - w_2$
$p_1 > p_2 < p_3$	$p_1^{'} > p_2^{'} < p_3^{'}$	$p_1^{'} = p_1 + w_1, p_3^{'} = p_3 + w_2$

1.3   Data Extraction and Image Restoration

The watermarked image I_W is converted into a one-dimensional pixel list in the same way as was done in embedding.

For I_W, LSBs of the first L_S watermarked pixels are collected into a bitstream $\mathcal{B}$. $\mathcal{B}$ are decompressed by an arithmetic decoder to retrieve the location map. By identifying the EOS symbol in $\mathcal{B}$, the bits from the start until EOS are decompressed by an arithmetic decoder to retrieve the location map. Data extraction and pixel restoration is carried out in inverse order as in embedding. M_d and m_d are transmitted to the receiving side. With help of M_d and m_d, the watermark

Table 2. Relationship among p_1', p_2' and p_3' and the range of d_1' (or d_2') after Procedure 1

Relationships among p_1', p_2' and p_3'	Range of d_1' (or d_2')		Restoration of p_1 and p_3
$p_1' > p_2' > p_3'$	$d_1' \in \{M_d, M_d + 1\}$, if $d_1' == M_d$	$d_2' \in \{M_d, M_d + 1\}$, if $d_2' == M_d$	$p_1 = p_1 - w_1, p_3 = p_3 + w_2$
$p_1' < p_2' < p_3'$	$w_1 = 0$ else	$w_2 = 0$ else	$p_1 = p_1 - w_1, p_3 = p_3 + w_2$
$p_1' \neq p_2' \neq p_3'$	$w_1 = 1$	$w_2 = 2$	$p_1 = p_1 - w_1, p_3 = p_3 + w_2$
$p_1' < p_2' > p_3'$			$p_1 = p_1 - w_1, p_3 = p_3 + w_2$
$p_1' > p_2' < p_3'$			$p_1 = p_1 - w_1, p_3 = p_3 + w_2$

Table 3. Procedure 2: embed_1_bit_and_increase_difference

Relationships among p_1, p_2 and p_3	Relationships among p_1', p_2' and p_3'	Modifications to p_1 and p_3
$p_1 > p_2 > p_3$	$p_1' \geq p_2' > p_3'$	$p_1' = p_1 + w_1, p_3' = p_3 - 1$
$p_1 < p_2 < p_3$	$p_1' \leq p_2' > p_3'$	$p_1' = p_1 - w_1, p_3' = p_3 + 1$
$p_1 < p_2 > p_3$	$p_1' < p_2' > p_3'$	if $w_1 == 1$ $\quad p_1' = p_1 - 1, p_3' = p_3 - 1$ else $\quad p_3' = p_3 - 1$
$p_1 > p_2 < p_3$	$p_1' > p_2' < p_3'$	if $w_1 == 1$ $\quad p_1' = p_1 + 1, p_3' = p_3 + 1$ else $\quad p_3' = p_3 + 1$

Table 4. Relationship among p_1', p_2' and p_3' and the range of d_1' (or d_2') after Procedure 2

Relationships among p_1', p_2' and p_3'	Range of d_1' (or d_2')	Restoration of p_1 and p_3
$p_1' \geq p_2' > p_3'$		$p_1 = p_1' - w_1, p_3 = p_3' + 1$
$p_1' \leq p_2' > p_3'$		$p_1 = p_1' + w_1, p_3 = p_3' - 1$
$p_1' < p_2' > p_3'$	$d_1' \in \{M_d, M_d + 1\}$, $M_d + 1 < d_2' < m_d + 1$ if $d_1' == M_d$ $\quad w_1 = 0$ else $\quad w_1 = 1$	if $w_1 == 1$ $\quad p_1 = p_1' + 1, p_3 = p_3' + 1$ else $\quad p_3' = p_3 + 1$
$p_1' > p_2' < p_3'$		if $w_1 == 1$ $\quad p_1 = p_1' - 1, p_3 = p_3' - 1$ else $\quad p_3 = p_3' - 1$

Table 5. Procedure 3: embed_1_bit_and_leave_unchanged

Relationships among p_1, p_2 and p_3	Relationships among p_1', p_2' and p_3'	Modifications to p_1 and p_3
$p_1 > p_2$	$p_1' > p_2'$	$p_1' = p_1 + w_1$
$p_1 \leq p_2$	$p_1' < p_2'$	$p_1' = p_1 - w_1$

Table 6. Relationship among p_1', p_2' and p_3' and the range of d_1' (or d_2') after Procedure 3

Relationships among p_1' and p_2'	Range of d_1' (or d_2')		Restoration of p_1 and p_3
$p_1' > p_2'$	$d_1' \in \{M_d, M_d + 1\}$	$d_2' < M_d$ or $d_2' > m_d$	$p_1 = p_1' - w_1$
$p_1' < p_2'$			$p_1 = p_1' + w_1$

Table 7. Procedure 4: increase_difference_and_embed_1_bit

Relationships among p_1, p_2 and p_3	Relationships among p_1', p_2' and p_3'	Modifications to p_1 and p_3
$p_1 > p_2 > p_3$	$p_1' > p_2' > p_3'$	$p_1' = p_1 + 1, p_3' = p_3 - w_2$
$p_1 < p_2 < p_3$	$p_1' > p_2' > p_3'$	$p_1' = p_1 - 1, p_3' = p_3 + w_2$

$p_1 < p_2 > p_3$	$p_1' < p_2' > p_3'$	if $w_2 == 1$
		    $p_1' = p_1 - 1, p_3' = p_3 - 1$
		else
		    $p_1' = p_1 - 1$

$p_1 > p_2 < p_3$	$p_1' > p_2' < p_3'$	if $w_2 == 1$
		    $p_1' = p_1 + 1, p_3' = p_3 + 1$
		else
		    $p_1' = p_1 + 1$

Table 8. Relationship among p_1', p_2' and p_3' and the range of d_1' (or d_2') after Procedure 4

Relationships among p_1', p_2' and p_3'	Range of d_1' (or d_2')	Restoration of p_1 and p_3
$p_1' > p_2' > p_3'$		$p_1 = p_1' - 1, p_3 = p_3' + w_2$

$p_1' > p_2' > p_3'$	$M_d + 1 < d_1' < m_d + 1$
	$d_2' \in \{M_d, M_d + 1\}$
	if $d_2' == M_d$
	    $w_2 = 0$
	else
	    $w_2 = 1$
		$p_1 = p_1' + 1, p_3 = p_3' - w_2$

$p_1' < p_2' > p_3'$		if $w_2 == 1$
		    $p_1 = p_1' + 1, p_3 = p_3' + 1$
		else
		    $p_1 = p_1' + 1$

$p_1' > p_2' < p_3'$		if $w_2 == 1$
		    $p_1 = p_1' - 1, p_3 = p_3' - 1$
		else
		    $p_1 = p_1' - 1$

Table 9. Procedure 5: increase_2_differences

Relationships among p_1, p_2 and p_3	Relationships among p_1', p_2' and p_3'	Modifications to p_1 and p_3
$p_1 > p_2 > p_3$	$p_1' > p_2' > p_3'$	$p_1' = p_1 + 1, p_3' = p_3 - 1$
$p_1 < p_2 < p_3$	$p_1' > p_2' > p_3'$	$p_1' = p_1 - 1, p_3' = p_3 + 1$
$p_1 < p_2 > p_3$	$p_1' < p_2' > p_3'$	$p_1' = p_1 - 1, p_3' = p_3 - 1$
$p_1 > p_2 < p_3$	$p_1' > p_2' < p_3'$	$p_1' = p_1 + 1, p_3' = p_3 + 1$

Table 10. Relationship among p_1', p_2' and p_3' and the range of d_1' (or d_2') after Procedure 5

Relationships among p_1', p_2' and p_3'	Range of d_1' (or d_2')	Restoration of p_1 and p_3
$p_1' > p_2' > p_3'$		$p_1 = p_1' - 1, p_3 = p_3' + 1$
$p_1' > p_2' > p_3'$	$M_d + 1 < d_1' < m_d + 1$	$p_1 = p_1' + 1, p_3 = p_3' - 1$
$p_1' < p_2' > p_3'$	$M_d + 1 < d_2' < m_d + 1$	$p_1 = p_1' + 1, p_3 = p_3' + 1$
$p_1' > p_2' < p_3'$		$p_1 = p_1' - 1, p_3 = p_3' - 1$

Table 11. Procedure 6: increase_difference_and_leave_unchanged

Relationships among p_1, p_2 and p_3	Relationships among p_1', p_2' and p_3'	Modifications to p_1 and p_3
$p_1 > p_2$	$p_1' > p_2'$	$p_1' = p_1 + 1$
$p_1 \leq p_2$	$p_1' < p_2'$	$p_1' = p_1 - 1$

Table 12. Relationship among p_1', p_2' and the range of d_1' (or d_2') after Procedure 6

Relationships among p_1', p_2' and p_3'	Range of d_1' (or d_2')		Restoration of p_1 and p_3
$p_1' > p_2'$	$M_d + 1 < d_1' < m_d + 1$	$d_2' < M_d$ or $d_2' > m_d$	$p_1 = p_1' - 1$
$p_1' < p_2'$			$p_1 = p_1' + 1$

Table 13. Procedure 7: leave_unchanged_and_embed_1_bit

Relationships among p_1, p_2 and p_3	Relationships among p_1', p_2' and p_3'	Modifications to p_1 and p_3
$p_2 > p_3$	$p_2' > p_3'$	$p_3' = p_3 + w_2$
$p_2 \leq p_3$	$p_2' < p_3'$	$p_3' = p_3 - w_2$

Table 14. Relationship among p_1', p_2' and p_3' and the range of d_1' (or d_2') after Procedure 7

Relationships among p_1', p_2' and p_3'	Range of d_1' (or d_2')		Restoration of p_1 and p_3
$p_2' > p_3'$	$d_1' < M_d$ or $d_1' > m_d$	$d_2' \in \{M_d, M_d + 1\}$	$p_3 = p_3' - w_2$
$p_2' < p_3'$			$p_3 = p_3' + w_2$

Table 15. Procedure 8: leave_unchanged_and_increase_difference

Relationships among p_1, p_2 and p_3	Relationships among p_1', p_2' and p_3'	Modifications to p_1 and p_3
$p_2 > p_3$	$p_2' > p_3'$	$p_3' = p_3 - 1$
$p_2 < p_3$	$p_2' < p_3'$	$p_3' = p_3 + 1$

Table 16. Relationship among p_2' and p_3' and the range of d_1' (or d_2') after Procedure 8

Relationships among p_1', p_2' and p_3'	Range of d_1' (or d_2')		Restoration of p_1 and p_3
$p_2' > p_3'$	$d_1' < M_d$ or $d_1' > m_d$	$M_d + 1 < d_2' < m_d + 1$	$p_3 = p_3 + 1$
$p_2' < p_3'$			$p_3 = p_3 - 1$

sequence can be correctly extracted. For each pair $p' = (p_1', p_2', p_3')$, if p''s location is associated with '0' in the location map, then it is ignored. Otherwise, p can be retrieved by means of Tables 2, 4, 6, 8, 10, 12, 14, 16.

2   Experimental Results

The proposed method is implemented and tested on various standard test images using MATLAB. The performance for two most frequently used 512×512 grayscale standard test images is presented in Fig. 2, where we use number of bits and PSNR values for measurement for capacity and distortion.

We modify Lin's method by keeping p_2 of any block p unchanged. By evaluating the correlation between p_2 and all the pixels surrounding p, we can get the intra-block correlation of p on the decoding side without the need of any additional information. We only embed watermark into blocks with high intra-block correlation, so that high PSNR value can be obtained at the low embedding rate. For 'Airplane (F-16)', it also can be seen from Fig. 2 that our method performs well, and significantly outperforms Li et al's when the embedding rates are lower than 0.2665 bpp. 'Baboon' is a typical image with large areas of complex texture, so the obtained bit-rate is slightly lower at the same PSNR. Fig. 2 shows that the proposed method achieves higher embedding capacity with lower embedding distortion than Li et al.'s, when the embedding capacity exceeds 10000 bits. At the current stage, we utilize three pixels to carry at most two-bit watermark and do not apply multi-layer embedding. In our future work, we will devote ourselves to the research that two pixels carry two-bit watermark.
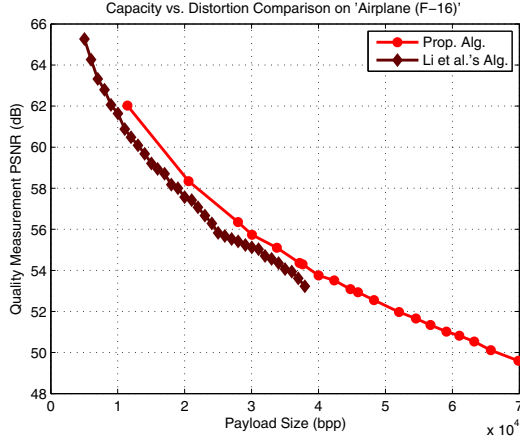
(a) 'Airplane (F-16)'
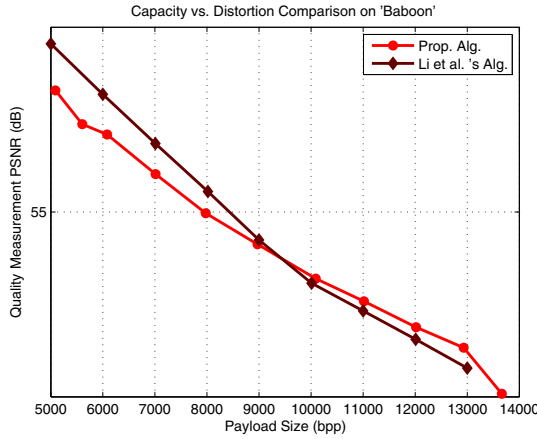
(b) 'Baboon'

Fig. 2. Performance comparisons between the proposed method Li *et al.* [15] for the test images: (a) 'Airplane (F-16)', (b) 'Baboon'.

3   Conclusions

Reversible watermarking based on position determination and three-pixel block difference is proposed in this paper. In the proposed method, for a three-pixel block, no modification is allowed to its center pixel (CP). This unchanged pixel along with all the neighbors surrounding this block constitute a set used for evaluating the intra-block correlation. The incorporation of CP in this set helps to largely enhance the estimation accuracy. According to the strength of correlation, we determine this block into a smooth or complex region. When the desired embedding rate is low, we only modify those blocks located in smooth

regions while keeping the others unchanged. Therefore, the PSNR (peak signal to noise ratio) value is largely increased. Experimental result also demonstrate that the proposed method is effective.

References

1. Honsinger, C.W., Jones, P., Rabbani, P.M., Stoffe, J.C.: Lossless recovery of an original image containing embedded data, US patent: 6278791W (2001)
2. Tian, J.: Reversible data embedding using a difference expansion. IEEE Trans. Circuits Syst. Video Technol. 13(8), 890–896 (2003)
3. Alattar, A.M.: Reversible watermark using the difference expansion of a generalized integer transform. IEEE Trans. Image Process. 13(8), 1147–1156 (2004)
4. Coltuc, D., Chassery, J.M.: Very fast watermarking by reversible contrast mapping. IEEE Signal Processing Letters 14(4), 255–258 (2007)
5. Thodi, D.M., Rodriguez, J.J.: Expansion embedding techniques for reversible watermarking. IEEE Trans. Image Process. 16(3), 721–730 (2007)
6. Weng, S.W., Zhao, Y., Pan, J.S., Ni, R.R.: Reversible watermarking based on invariability and adjustment on pixel pairs. IEEE Signal Processing Letters 45(20), 1022–1023 (2008)
7. Weng, S.W., Zhao, Y., Ni, R.R., Pan, J.S.: Parity-invariability-based reversible watermarking. IET Electronics Letters 1(2), 91–95 (2009)
8. Wang, X., Li, X.L., Yang, B., Guo, Z.M.: Efficient generalized integer transform for reversible watermarking. IEEE Signal Processing Letters 17(6), 567–570 (2010)
9. Wang, X., Li, X.L., Yang, B.: High capacity reversible image watermarking based on in transform. In: Proceedings of ICIP (2010)
10. Peng, F., Li, X., Yang, B.: Adaptive reversible data hiding scheme based on integer transform. Signal Processing 92(1), 54–62 (2012)
11. Luo, L., Chen, Z., Chen, M., Zeng, X., Xiong, Z.: Reversible image watermarking using interpolation technique. IEEE Transactions on Information Forensics and Security 5(1), 187–193 (2010)
12. Li, X.L., Yang, B., Zeng, T.Y.: Efficient reversible watermarking based on adaptive prediction-error expansion and pixel selection. IEEE Trans. on Image Process. 20(12), 3524–3533 (2011)
13. Wu, H.-T., Huang, J.W.: Reversible image watermarking on prediction errors by efficient histogram modification. Signal Processing 92(12), 3000–3009 (2012)
14. Li, X.L., Li, J., Li, B., Yang, B.: High-fidelity reversible data hiding scheme based on pixel-value-ordering and prediction-error expansion. Signal Process. 93(1), 198–205 (2013)
15. Li, X.L., Zhang, W.M., Gui, X.L., Yang, B.: A novel reversible data hiding scheme based on two-dimensional difference-histogram modification. IEEE Transactions on Information Forensics and Security 8(7), 1091–1100 (2013)
16. Peng, F., Li, X.L., Yang, B.: Improved pvo-based reversible data hiding. Digital Signal Processing 25, 255–265 (2014)
17. Coatrieux, G., Guillou, C.L., Cauvin, J.M., Roux, C.: Reversible watermarking for knowledge digest embedding and reliability control in medical images. IEEE Transactions on Information Technology in Biomedicine 13(2), 158–165 (2009)
18. Lin, C.-C., Hsueh, N.-L.: A lossless data hiding scheme based on three-pixel block differences. Pattern Recognition 41(4), 1415–1425 (2008)

A Novel Encryption Algorithm
for Quantum Images Based on Quantum Wavelet
Transform and Diffusion

Shen Wang⋆, Xianhua Song, and Xiamu Niu

School of Computer Science and Technology
Harbin Institute of Technology, Harbin, China
Shen.Wang@hit.edu.cn

Abstract. In this paper, a novel quantum encryption scheme for quantum images based on quantum wavelet transform (QWT) and double diffusions is proposed. Firstly, diffusion operation applied on the input quantum image, and then QWT worked on the new quantum image to transform this image to the frequency domain. and following the diffusion operation is implemented on the QWT transformed quantum image. finally ,inverse QWT are used.The encryption keys are generated by a sensitive chaotic logistic map, which guarantee the security of the scheme. at the same time,we designed the corresponding quantum circuits to demonstrates that the reasonable of the proposed scheme.

Keywords: Quantum computation, Quantum image encryption, QWT, diffusion operation, Chaotic system.

1 Introduction

Since quantum computation is proposed, is has been raised great interested during the quantum field. Any quantum algorithm or unitary operation can be considered feasible depending on whether can design the corresponding equivalent quantum circuit [1,2]. Currently, quantum computation have been applied in many branches of science and technology such as image processing and computational geometry[3,4,5], pattern recognition[6,7,8].

Quantum image is the corresponding form of the classical image in quantum computer. And many quantum image representations are proposed such as Qubit Lattice [9], Real Ket[10], flexible representation of quantum image(FRQI)[11], an enhanced quantum representation(NEQR) for digital image is proposed in [12]. Moreover, quantum image representation for log-polar image (QUALPI) was proposed for the storage and processing quantum image sampled in log-polar coordinates [13].

With the deep research about quantum image, quantum image encryption has been a hot topic. In recent years, qubits encryption with hybrid keys has been proposed [14], but these schemes have been invented to qubits encryption due to some intrinsic features of images such as high correlation pixels. And the encryption algorithms utilizing quantum image geometric transformations has been proposed in [15]. However, the encrypted images are not noise-like and the sketches can be identified. Recently, quantum image encryption method based on Quantum Fourier transform and double phase encoding is designed [16]. Furthermore, Yang applied this method to the quantum color image encryption and obtain some good results [17].

In this paper, we proposed a novel quantum image encryption method, which is based on the QWT and diffusion operation. QWT can transform the image to its frequency domain, and it is conveniently for us to research the image properties. And in article we mainly use the QWT wavelet. Diffusion operation and chaotic mapping can make the quantum image color information distributed more uniform.

The framework of this paper is as follows. In section 2, we will give the background on quantum image, QWT and the chaotic logistic map. And the concrete quantum image encryption and decryption algorithm will be given in section 3. And the corresponding experiments and analysis are showed in part 4. Finally in part 5, we conclude this paper.

2   Related Works

In this paper, we will use the FRQI quantum model to realize our scheme. The concrete form of FRQI is described as following, which the image size is

$$2^n \times 2^n$$

in Eq. 1.

$$|I(\theta)\rangle = \frac{1}{2^n} \sum_{i=0}^{2^{2n}-1} (\cos\theta_i |0\rangle + \sin\theta_i |1\rangle) \otimes |i\rangle \tag{1}$$

$$\theta_i \in \left[0, \frac{\pi}{2}\right], i = 0, 1, \cdots 2^{2n} - 1$$

Wherein, $\cos\theta_i |0\rangle + \sin\theta_i |1\rangle$ encodes the color information and $|i\rangle$ encodes the corresponding position information of the quantum images. The position information includes two parts: the vertical and horizontal coordinates. Considering a quantum image in $2n$-qubits system,

$$|i\rangle = |y\rangle |x\rangle = |y_{n-1}y_{n-2}\cdots y_0\rangle |x_{n-1}x_{n-2}\cdots x_0\rangle$$

$$x, y \in \{0, 1, \cdots 2^n - 1\}$$

$$|y_j\rangle, |x_j\rangle \in \{|0\rangle, |1\rangle\}, j = 0, 1, \cdots, n-1$$

here
$$|y\rangle = |y_{n-1} y_{n-2} \cdots y_0\rangle$$
encodes the first n -qubits along the vertical location and
$$|x\rangle = |x_{n-1} x_{n-2} \cdots x_0\rangle$$
encodes the second n -qubits along the horizontal axis. Obviously ,the FRQI is a normalized state.

$$\||I(\theta)\rangle\| = \frac{1}{2^n} \sqrt{\sum_{i=0}^{2^{2n}-1} \left(\cos^2 \theta_i + \sin^2 \theta_i\right)} = 1 \tag{2}$$

Furthermore, the author proposed a polynomial preparation theorem that proves the existence of a unitary transformation which can turn a empty quantum image to the FRQI state. Mostly existing quantum image encryption is based on FRQI, so we select this quantum image model for our encryption scheme.

Wavelet transformation researches the multi-scales structure of signal and is important in image compression and processing. Concrete quantum circuits for Harr and Daubechies wavelet are designed in[16]. The quantum circuit of D^4 wavelet are designed in[18].

Chaos theory has been applied in different field of science and technology, such as mathematics, physics, engineering [18] since it was proposed in 1970s. And owing to many properties of chaos system, it has raised the interested of many researches. At the same time, its properties have the corresponding counterparts in traditional cryptosystems, such as confusion and diffusion. A general chaotic logistic map was defined as follows[19].

$$x_{\delta+1} = \frac{4\eta'^2 x_\delta (1 - x_\delta)}{1 + 4\left(\eta'^2 - 1\right) x_\delta (1 - x_\delta)} \tag{3}$$

Where $\delta = 0, 1, \cdots, n$, x_0 is the initial value, $x_0 \in (0,1)$, $-4 \le \eta' \le 4$. If given the initial values x_0, η' , iterating equation 3, a chaotic sequence is generated whose elements are distributed in $(0,1)$.Analysis has been proved that the generalized logistic map is more uniform than logistic map. we will use it to generate key in the diffuse phase of this encryption scheme. Diffusion operation contains many color transformations to recode the color angle in FRQI. And in this paper we use twice diffusion operation and corresponding twice logistic mapping, so it makes key space enough big for our experiment and to guarantee our schemes feasibility.

In some cases, it is the Contact Volume Editor that checks all the final pdfs. In such cases, the authors are not involved in the checking phase.

3   Quantum Image Encryption and Decryption Scheme

In our proposed quantum encryption scheme, firstly, we use diffusion operations to act on the input image ,and then apply QWT to the new quantum image.

following, the diffusion operation is used on the encrypted image, finally, inverse QWT was applied on the quantum image.Next the concrete encryption and decryption process are described.

Encryption Scheme. Algorithm 1. Encryption scheme Input: quantum image $|I(\theta)\rangle$ size of $2n+1$ qubits, initial parameters η', η'' and x_0', x_0'' and $-4 \leq \eta' \leq 4$, $-4 \leq \eta'' \leq 4$,$x_0, x_0' \in (0,1)$ Output: encrypted quantum image

$$D(|I(\theta)\rangle)$$

Diffusion Phase:

Step 1: Quantum image preparation Given a empty quantum image sized $2^n \times 2^n$, using the polynomial preparation theorem can obtain the input quantum image whose FRQI is 1:

Step 2: Generating 2^{2n} reals γ_k using the initial parameters η' and x_0' through iterate equation (4) about 2^{2n} times.

Step 3: Designing the rotation gates $R_y(2\varphi_k)$, $\varphi_k \in [0, \pi/2]$, $k = 0, 1, \cdots, 2^{2n}-1$ using the reals γ_k ,where,$\varphi_k = \arccos(\gamma_k)$,arcos is the inverse function of cosine function. Thus, a sequence of unitary transforms is constructed using equation 4 and 5

$$C_k = \left(I \otimes \sum_{j=0, j \neq k}^{2^{2n}-1} |j\rangle \langle j| \right) + R_y(2\varphi_k) \otimes |k\rangle \langle k| \tag{4}$$

$$R_y(2\varphi_k) = \begin{pmatrix} \cos\varphi_k & -\sin\varphi_k \\ \sin\varphi_k & \cos\varphi_k \end{pmatrix} \tag{5}$$

$R_y(2\varphi_k)$ is a rotation matrix, obviously, the controlled rotation matrix C_k is a unitary matrix, applying C_k on quantum image $|I(\theta)\rangle$ gives us that:

$$\begin{aligned} &C_k(|I(\theta)\rangle) \\ &= C_k \left(\frac{1}{2^n} \sum_{i=0}^{2^{2n}-1} (\cos\theta_i |0\rangle + \sin\theta_i |1\rangle) \otimes |i\rangle \right) \\ &= \frac{1}{2^n} \left[\sum_{i=0, i \neq k}^{2^{2n}-1} (\cos\theta_i |0\rangle + \sin\theta_i |1\rangle) \otimes |i\rangle \right. \\ &\left. + (\cos(\theta_k + \varphi_k) |0\rangle + \sin(\theta_k + \varphi_k) |1\rangle) \otimes |k\rangle \right] \end{aligned} \tag{6}$$

Step 4: Then using operation

$$C_k \left(k = 0, 1, \cdots 2^{2n} - 1 \right)$$

2^{2n} times, that is

$$\prod_{k=0}^{2^{2n}-1} C_k$$

transform, and we will get the firstly encryption image.

$$
\prod_{k=0}^{2^{2n}-1} C_k \left(|I(\theta)\rangle \right)
$$
$$
= \frac{1}{2^n} \sum_{i=0}^{2^{2n}-1} \left(\cos(\theta_i + \varphi_i)|0\rangle + \sin(\theta_i + \varphi_i)|1\rangle \right) \otimes |i\rangle
\tag{7}
$$

QWT phase:

Step1: *Quantum wavelet transform* execute QWT on the firstly encryption quantum image $\prod_{k=0}^{2^{2n}-1} C_k \left(|I(\theta)\rangle \right)$, obtaining its wavelet form of encryption quantum image $QWT \left(\prod_{k=0}^{2^{2n}-1} C_k \left(|I(\theta)\rangle \right) \right)$ shown as follows

$$
QWT \left(\prod_{k=0}^{2^{2n}-1} C_k \left(|I(\theta)\rangle \right) \right)
$$
$$
= \frac{1}{2^n} \sum_{i=0}^{2^{2n}-1} QWT \left((\cos\theta_i' |0\rangle + \sin\theta_i' |1\rangle) \otimes |i\rangle \right)
\tag{8}
$$
$$
= \sum_{i=0}^{2^{2n}-1} |wc_i\rangle \otimes |i\rangle
$$

Where $\theta_i' = \theta_i + \varphi_i$, and this step represents the QWT of the firstly encryption image. Noticing that in our paper we use D^4 wavelet.

Second diffusion Phase:

After QWT, we will again use the diffusion operation. The method just like the above diffusion operation, however, the different point is that use another initial values η'' and x_0'' randomly produce the chaotic logistic sequence using equation (3). At the same time, design the transformation $C_k' \left(k = 0, 1, \cdots 2^{2n} - 1 \right)$. Then apply these transform on the QWT encryption image.

Inverse QWT Phase:

The last step of the algorithm is the inverse QWT. We just produce an inverse D^4 wavelet as the transform.

Furthermore, we give the framework of this quantum image encryption algorithm.

Quantum image is the representation of classical image in quantum computer, so the whole implementation process must satisfy the principal of quantum mechanics. Obviously the QWT satisfies quantum mechanics.

Next we will give the concrete quantum circuit of the quantum image encryption scheme.

It is obvious that all the phases are inverse, next we will describe the decryption algorithm.

Decryption Scheme. Algorithm 2. Decryption scheme

Input: encrypted quantum image $D \left(|I(\theta)\rangle \right)$, the same parameters η' and x_0', η'' and x_0''.

Output: decrypted quantum image $|I(\theta)\rangle$.
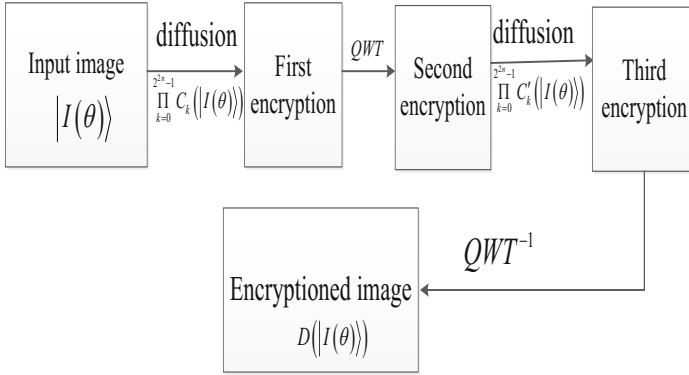
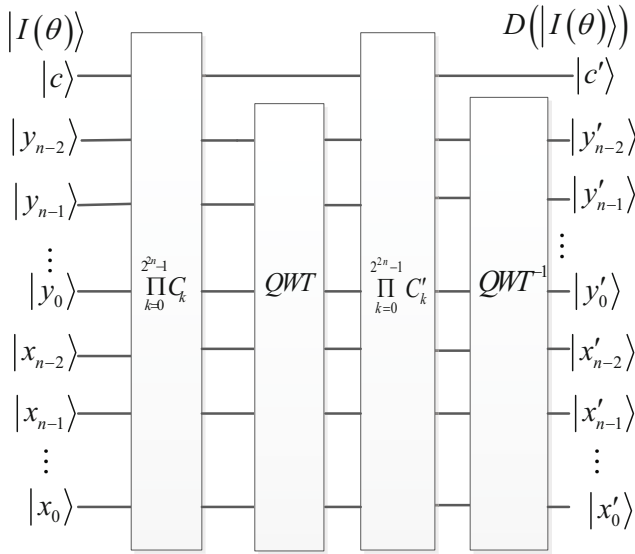**Fig. 1.** the framework of the quantum encryption algorithm

Fig. 2. The quantum circuit of quantum image encryption

QWT Phase: Firstly, we applied QWT to the encrypted image $D\left(\left|I\left(\theta\right)\right\rangle\right)$.

Inverse diffusion phase:

step1: using the same keys of η'' and x_0'' to generate 2^{2n} reals γ_k', which is the same as the γ_k' used in step of the again diffusion phase of algorithm 1, by iterating logistic map equation (3) about 2^{2n} times.

Step 2: Designing the rotation gates $R_y\left(2\varphi_k'\right)$, $\varphi_k' \in [0, \pi/2]$, using the reals γ_k' produced in step 1. Then a sequence of unitary transforms C_k' is constructed by equation (5) and (6).

Step 3: Rotate the color codes of $QWT\left(D\left(\left|I\left(\theta\right)\right\rangle\right)\right)$ using the Hermitian conjugate of the matrix C_k' for all $k=0,1,\cdots 2^{2n}-1$, then we get the inverse diffusion result $\prod\limits_{k=0}^{2^{2n}-1}C_k^{*'}\left(QWT\left(\left|I\left(\theta\right)\right\rangle\right)\right)$, where $\prod\limits_{k=0}^{2^{2n}-1}C_k^{*'}$ is the Hermitian conjugate of the matrix C_k'.

QWT^{-1}phase:

Step 1: Execute *Quantum inverse wavelet transform* QWT^{-1} on the middle encryption quantum image $\prod\limits_{k=0}^{2^{2n}-1}C_k^{*'}\left(QWT\left(\left|I\left(\theta\right)\right\rangle\right)\right)$.

Another diffusion phase:

Using the same keys of η' and x_0' to generate 2^{2n} reals γ_k, and produce the same rotate transformation C_k. Then we can use the conjugate of $C_k, k=0,1,\cdots 2^{2n}-1$ to the above quantum image and last we get quantum image$\left|I\left(\theta\right)\right\rangle$.

From all the steps of above, we can get the conclusion that our quantum image encryption scheme is feasible in theory. Then we will analysis concrete experiment to show our scheme is in fact reasonable.

4   Conclusion

In this paper, we proposed a novel quantum image encryption method based on QWT and diffusion operation. And theory analysis show that our algorithm is reasonable in quantum theory and can be realized especially reflected in the designed quantum circuit. At the same time , all the transformation used in the article is unitary, so we can easily get the inverse quantum circuit and decryption algorithm.

References

1. Le, P.Q., Iliyasu, A.M., Dong, F., Hirota, K.: Efficient color transformations on quantum images. Journal of Advanced Computational Intelligence and Intelligent Informatics 15(6), 698–706 (2011)
2. Deutsch, D.: Quantum theory, the Church-Turing principle and the universal quantum computer. Proc. R. Soc. London 400, 97–117 (1985)
3. Lanzagorta, M., Uhlmann, J.: Quantum algorithmic methods for computational geometry. Mathematical Structures in Computer Science 20(6), 1117–1125 (2010)
4. Venegas-Andraca, S.E., Bose, S.: Storing, processing and retrieving and image using quantum mechanics. In: Proc. SPIE Conf. Quantum Inf. Comput., vol. 5105, pp. 137–147 (2003)
5. Venegas-Andraca, S.E., Bose, S.: Quantum computation and image processing: New trends in artificial intelligence. In: Proceedings of the International Conference on Artifficial Intelligence, IJCAI 2003, pp. 1563–1564 (2003)
6. Trugenberger, C.: Quantum pattern regonition. Quantum Information Processing 1(6), 471–493 (2002)
7. Trugenberger, C.: Probabilistic quantum memories. Phys. Rev. Lett. 87, 067901 (2001)
8. Trugenberger, C.: Phase transitions in quantum pattern recoginition. Phys. Rev. Lett. 89, 277903 (2002)

9. Venegas-Andraca, S.E., Bose, S.: Storing, processing and retrieving an image using quantum mechanics. In: Proc. SPIE Conf. Quantum Inf. Comput., vol. 5105, pp. 137–147 (2003)

10. Latorre, J.I.: Image compression and entanglement. arXiv: quant-ph/0510031 (2005)

11. Le, P.Q., Dong, F., Hirota, K.: A flexible representation of quantum images for polynomial preparation, image compression and processing operations. Quantum Inf. Process. 10(1), 63–84 (2010)

12. Zhang, Y., Lu, K., Gao, Y.H., Wang, M.: NEQR: a novel enhanced quantum representation of digital images. Quantum Inf. Process., 1–28 (2013)

13. Zhang, Y., Lu, K., Gao, Y.H., Xu, K.: A novel quantum representation for log-polar images. Quantum Inf. Process., 1–24 (2013)

14. Zhou, N., Liu, Y., Zeng, G., Zhang, J.: Novel qubit block encryption algorithm with hybrid keys. Physica A 375, 693–698 (2007)

15. Zhou, R.G., Wu, Q., Zhang, M.Q., Shen, C.Y.: Quantum image Encryption and Decryption Algogorithm Based on Quantum Images Geometric Transformations. International Journal of Theoretical Physics, 1–16 (2012)

16. Yang, Y.G., Xia, J., Jia, X., Zhang, H.: Novel Image Encryption/Decryption Based on Quantum Fourier Transform and Double Phase Encoding. Quantum Inf. Process., 1–17 (2013)

17. Yang, Y.G., Jia, X., Sun, S.J., Pan, Q.X.: Quantum cryptographic algorithm for color images using quantum Fourier transform and double random-phase encoding. Information Science. Elsevier (2014)

18. Barenco, A., Ekert, A., Suominen, K.-A., Torma, P.: Approximate quantum Fourier transform and decoherence. Phys. Rev. A 54, 139 (1996)

19. Abd EI-Latif, A.A., Niu, X.M., Amin, M.: A new image cipher in time and frequency domains. Optics Communications 285, 4241–4251 (2012)

Interleaving and Sparse Random Coded Aperture for Lens-Free Visible Imaging

Zhenglin Wang and Ivan Lee

School of Information Technology and Mathematical Sciences,
University of South Australia, Mawson Lakes, SA 5095, Australia

Abstract. Coded aperture has been applied to short wavelength imaging (e.g., gamma-ray), and it suffers from diffraction and interference for taking longer wavelength images. This paper investigates an interleaving and sparse random (ISR) coded aperture to reduce the impact of diffraction and interference for visible imaging. The interleaving technique treats coded aperture as a combination of many small replicas to reduce the diffraction effects and to increase the angular resolution. The sparse random coded aperture reduces the interference effects by increasing the separations between adjacent open elements. These techniques facilitate the analysis of the imaging model based only on geometric optics. Compressed sensing is applied to recover the coded image by coded aperture, and a physical prototype is developed to examine the proposed techniques.

Keywords: Computational imaging, Coded aperture imaging, Image reconstruction techniques.

1 Introduction

Coded aperture was introduced in 1960s [1], and it plays an important role in astronomical and medical imaging for short wavelength sources, such as χ and γ rays. These sources are considered as non-diffracting and non-refracting radiation, and they are difficult to be imaged by a lens-based camera. Coded aperture consists of a mask with many pinholes distributed with a pre-defined pattern. The mask with bigger openings substitutes the single hole in a pinhole camera, allowing more lights in. The coded aperture technique can improve the signal-to-noise ratio (SNR), reduce the exposure time, widen the field-of-view, and simultaneously maintain several benefits of a pinhole camera, such as low cost, negligible image distortion and infinite depth-of-field [2], [3]. Hence, there has been increasing interest in applying the coded aperture technique to visible or infrared imaging [4], [5], [6].

When coded aperture is applied for capturing visible wavelength images at high resolution without incorporating lens, diffraction and interference effects become major challenges [5]. A fine-scale mask was adopted in [5] to attain high resolution images, and the diffraction and interference effects had to be taken into account in their imaging model. As a result, its point spread function (PSF) had

to be obtained by extra physical tests instead of pure geometric optics analysis. Their imaging model may have the following problems: 1) it makes configurable coded aperture unfeasible because each mask has to attain its PSF by extra physical tests; 2) the reconstruction accuracy is affected in that the measured PSF inevitably introduces some errors. In addition, the fine-scale mask may increase the manufacturing difficulties.

A pinhole camera is probably the simplest lens-free camera. It usually chooses a proper pinhole size to make the diffraction circle (for a point source) fall within the geometric spot, so that the diffraction effects are negligible. For a pinhole camera, the image resolution is mainly determined by the resolution of sensor array. However, the image resolution for a coded aperture based camera depends on both the mask and the sensors. The element size in the mask is desired to be small enough to achieve high resolution, but a tiny hole produces severe diffraction to affect the image quality. To address this conflict, an interleaving technique is investigated. We still follow the pinhole principle to design the element size to reduce the diffraction effects. The interleaving technique is employed at decoding to treat the mask as a combination of many fine-scale replicas, assisting to recover a high resolution image.

Most current coded apertures are designed for astronomical and medical imaging where the intensities of the radiation sources are very low. The mask with big openings can increase the flux onto the detector. Meanwhile, the open elements in the mask can be arranged densely without considering the interference effects. Furthermore, the elements are usually required to be distributed in a specific pattern to facilitate the decoding process [3], [7]. On the other hand, in many visible applications, the intensities of the light sources are high or a big exposure time can be used for compensation. Instead, the interference effects become significant to the reconstruction accuracy. Therefore, a sparse random coded aperture is proposed in this article to make a tradeoff between the dense coded aperture and the single pinhole. The open elements sparsely spread on the mask to reduce the interference effects. Additionally, the sparse masks may facilitate the fabrication.

The resulting picture by a coded aperture is a superposition of many individual pinhole images. Each sensor records an intensity summation of all light sources (note that a light source will contribute zero to the sum if it corresponds to a closed element). So, a decoding process is necessary to extract a meaningful image from the recorded picture. Several linear decoding algorithms have been developed for different coded apertures [3], [7], [8], [9], [10]. These algorithms usually impose specific restrictions on the coded aperture patterns for the sake of seeking a proper decoding function (e.g., balanced correlation [3]). They cannot recover the image properly for the proposed sparse random coded aperture. Some researchers have applied the emerging theory of compressed sensing (CS) [11], [12] to coded aperture imaging [13], [14], [15]. Their researches targeted at lens-based imaging systems and the mask was placed in an intermediate image plane (e.g., the coded aperture compressive temporal imaging (CACTI) system [15]). So, they are different from our studies of lens-free imaging system. CS

has been used for diverse image applications (e.g.,[24]), and it is studied here to provide a robust nonlinear reconstruction way for the proposed coded aperture.

2   Coded Aperture Design

In coded aperture imaging applications, the object distance is assumed to be much bigger than the image distance, so the incident lights from a point source to the camera can be considered parallel and uniform. When a sensor array is used to capture the image, the imaging system will be digitalized. If only geometric optics is taken into account, the imaging system can be modeled as

$$P(k,l) = \sum_i \sum_j O(i,j)A(i+k,j+l) + N(k,l) \tag{1}$$

where $P(k,l)$ denotes a sample on the sensor (k,l), $O(i,j)$ denotes the intensity of the point source at (i,j), A denotes the discrete PSF of the coded aperture, and N denotes the noise [3]. Furthermore, (1) can be represented by

$$P = O * A + N \tag{2}$$

where $*$ denotes the correlation operator.

A coded aperture consists of many open or closed equal-size elements. Each element is desired to have the same dimension as a sensor. It is also required to be aligned with the sensor. Then, A can be given by a zero-one matrix, the open elements meaning '1's and the closed elements meaning '0's. In addition, when P, A and O are finite, (1) can be expressed by a system of linear equations:

$$\boldsymbol{p} = \Phi\boldsymbol{x} + \boldsymbol{n} \tag{3}$$

where the vectors $\boldsymbol{p}$ and $\boldsymbol{x}$ are obtained by lexicographic reordering of the matrices P and O, and Φ is a Toeplitz matrix with $\{0, 1\}$ elements generated by packing A in lexical order [16].

The above imaging model bases on geometric optics, which does not take into account the diffraction and interference effects. However, the diffraction effects are significant when visible light passes through an individual hole whose size is comparable to the wavelength; as well, the interference effects arise when visible light passes through different holes whose separations are small. On the other hand, a tiny hole is desired to achieve high angular resolution, and densely distributed holes are required to favor conventional decoding algorithms. So, the above imaging model is inapplicable when conventional coded apertures are employed for visible imaging. For example, researchers in [5] adopted a fine-scale mask with element side length of $41\mu m$ to capture high resolution images (the angular resolution is around $1mrad$), and they had to measure the PSF by extra physical tests. We propose an interleaving and sparse random coded aperture to reduce the diffraction and interference effects so that the above imaging model can be applied directly without spending extra efforts on the PSF.

Fig. 1. Interleaving coded aperture (a grayscale indicates a sub-mask)

2.1 Interleaving Coded Aperture

For a pinhole camera, the pinhole diameter (for a circular hole) is suggested to be $d = 1.56\sqrt{\lambda f}$ where λ is the wavelength and f is the focal length [17]. Such a pinhole size is usually considered optimum because it can make the best balance between geometric imaging and diffraction effects. Although the elements in our coded aperture are equal-size squares, this formula is still used to calculate the side length. For example, the focal length is about $40mm$ in our lens-free camera and light with wavelength of $550nm$ (yellow-green) is dominant in a black-and-white picture, so the chosen side length is around $0.23mm$. Such element dimensions produce an angular resolution of about $6mrad$, which is worse than that in [5]. On the other hand, the Canon sensor array has a pixel pitch of $4.3\mu m$, which can provide a finer resolution. Thus, an interleaving coded aperture technique is studied to improve the angular resolution.

In the interleaving coded aperture, each element in the mask is regarded as an array of $k \times k$ sub-elements. All these sub-elements have the same value with their parent. An example is shown in Fig. 1. Each element in the mask is divided into an array of 2×2 sub-elements. The mask is seen as a combination of 4 sub-masks interleaving together. Each sub-mask has the same coded pattern as the parent mask. As a result, the angular resolution is doubled. The larger k, the finer the resolution is. The interleaving coded aperture technique is based on two principles: 1) when the light rays are from a small far field of view (e.g., $6mrad$), they are parallel to pass through a small open hole (e.g. a $0.23mm$ square) to reach the sensors; 2) each sensor and sub-element are identical in size and aligned in position. Then, the imaging system can be regarded as k^2 independent sub-masks working together. At decoding, each sub-mask and its corresponding samples are used to reconstruct a low resolution image independently. These reconstructed images are finally reassembled into a high resolution image. If the interleaving coded aperture technique is employed with $k = 8$, the angular resolution in previous example will achieve around $0.75mrad$, which is better than that in [5].

2.2 Sparse Random Mask

In 1968, Dicke proposed a random coded aperture consisting of many randomly spaced pinholes having an overall transmission $\approx 50\%$, to improve the SNR whilst retaining the good angular resolution commensurate with a single hole size
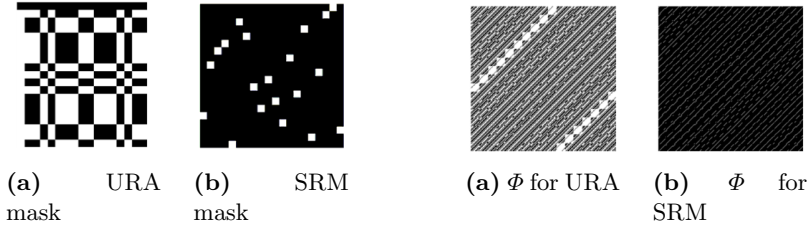
(a)   URA   **(b)**   SRM
mask   mask

(a) Φ for URA   **(b)** Φ for
SRM

Fig. 2. Different basic coded aperture
patterns

Fig. 3. Lexical-order matrices for the
masks used in this paper

[1]. The random coded aperture can adopt autocorrelation or generalized inverse
for decoding, but the former method produces an inherent noise in reconstruction
while the latter is sensitive to noise [3]. Then, Fenimore and Cannon [3] proposed
uniformly redundant arrays (URA) for coded aperture (Fig. 2(a)). The URA
mask exhibits better reconstruction accuracy because it favors to construct a
decoding function G such that the system point-spread function (SPSF) $A * G$
is exactly a delta function. The decoding process is as

$$\tilde{O} = P * G = (O * A) * G + N * G = O + N * G. \tag{4}$$

The URA mask and its derivation MURA [10] have been widely applied to
astronomical imaging [16].

However, the URA mask was designed to improve the SNR and to facilitate
the decoding for short wavelength imaging. It has two limitations. First, the
URA mask must have dimensions r-by-s where r and s are prime numbers, and
$(r-s)$ equals 2; second, the open elements need to occupy about $50\%(rs/2)$ of the
URA mask for obtaining a SPSF of being an exact delta function (see Fig. 4(a)).
These limitations restrict the flexibilities of the URA mask, and simultaneously
result in open elements densely spreading on the mask. Specifically, when it is
adopted for visible imaging, the interference effects will be outstanding.

So, a sparse random mask (SRM) is examined in this paper to reduce the
interference effects. In a sense, the SRM mask makes a tradeoff between the
dense coded aperture and the single pinhole. On the other hand, the SRM mask
leads to produce a sparse random PSF matrix. Sparse random matrices have been
intensively investigated for signal acquisition in the CS context. They can relieve
the computational complexity for large-scale signals. In addition, many readily
available CS-based recovery algorithms are potential for decoding. Therefore,
a sparse random matrix with $\{0, 1\}$ elements proposed in [18] are adopted to
design our mask. To reduce the interference effects as much as possible, only one
open element is selected in each column or row. A resulting 20-by-20 SRM mask
is shown as Fig. 2(b).

The famous double-slit experiment model is used to explain the advantage of
SRM. A point light source passes through two adjacent pinholes to interference,
producing many bright and dark fringes rather than two bright spots on the
film. The gaps of neighboring bright fringes are approximately $f\lambda/a$ where f is

the image distance, λ is the wavelength and a is the separation between the two pinholes [19]. As the separation becomes larger, the interferences will diminish and the fringes will be formed into two bright spots. The separations between adjacent open elements in the SRM mask are likely to be bigger than those in the URA mask, so the SRM mask can lessen the interference effects. In addition, since only one open element in each column or row needs to be aligned with the sensor, the SRM mask is advantageous to reduce the alignment error.

3   CS for Sparse Random Coded Aperture Decoding

The recorded picture by coded aperture imaging is an unclear superposition of multiple copies of the target scene. Decoding must be performed to obtain the ground truth. Conventional decoding algorithms are mostly based on linear deconvolution. For the URA mask, the efficient decoding method is balanced correlation as (4). G is constructed by

$$G(i,j) = \begin{cases} 1, & \text{if} \quad A(i,j) = 1; \\ -1, & \text{if} \quad A(i,j) = 0. \end{cases} \tag{5}$$

The SPSF of $A * G$ is a perfect delta function (Fig. 4(a)). The target image can be exactly recovered if no noise exists. Moreover, when Φ in (3) is non-singular, another intuitive decoding method is generalized inverse, formulated by

$$\tilde{x} = \Phi^{-1}p = x + \Phi^{-1}n. \tag{6}$$

But generalized inverse is sensitive to noise [16]. On the other hand, The SPSF for a SRM mask is an unsatisfactory delta function (see Fig. 4(b)), resulting
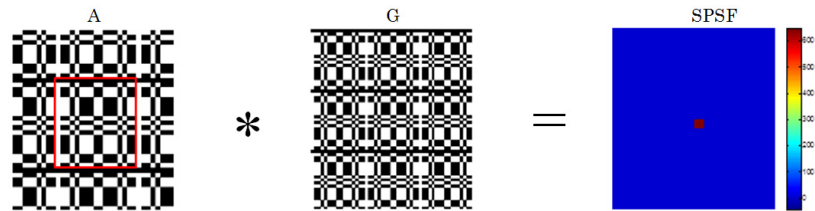
(a) The SPSF for URA is an ideal δ function
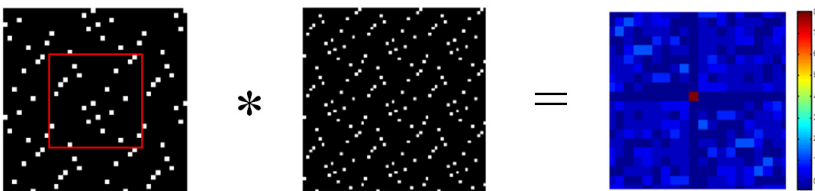
(b) The SPSF for SRM is an unsatisfactory δ function

Fig. 4. The SPSFs for URA and SRM masks

that the balanced correlation method cannot obtain a good reconstruction. At the same time, Φ for the SRM mask is singular so that generalized inverse is inapplicable as well. So, we have to explore another decoding approach for the SRM mask.

The new theory of CS has been applied to image sampling. A well-known CS-based camera is the single-pixel camera [20], which employs a digital micromirror device (DMD), controlled by a random number generator (RNG), to measure the target image in a random sense. The CS sampling is formulated by

$$y = \Phi x + n \tag{7}$$

where y is a vector of the observations, x is the vector form of the target image, Φ denotes the measurement matrix modeling the measuring procedure, and n denotes the noise. Then, x can be well reconstructed by many CS-based recovery algorithms even if the samples are less than the real pixels. (7) is identical to (3). Indeed, the single-pixel camera acquires linear measurements of the target image rather than real pixels, and so does coded aperture imaging. So, the decoding algorithms for CS are potential to be adopted for coded aperture imaging.

Another reason for selecting sparse random coded aperture is that the lexical-order form Φ of the masks (Fig. 3) must obey the restricted isometry property (RIP) [21]. RIP is the premise that the CS-based recovery algorithms can be applied. When Φ for the SRM mask is singular, a pre-process as below is necessary. The maximum linearly independent rows are abstracted from Φ, and then reorganized into a new rectangle matrix $\bar{\Phi}$. The remaining rows and their corresponding entries in y are discarded. Given Φ is an N-by-N ($N = r * s$) matrix and $M = rank(\Phi)$, $\bar{\Phi}$ denotes the remaining M-by-N matrix and $\bar{y}$ denotes the remaining M-length vector. Then, (7) is rewritten as

$$\bar{y} = \bar{\Phi} x + n \tag{8}$$

Although $M < N$ and (8) is ill-posed, CS can solve it efficiently [22]. The gradient projection for sparse reconstruction (GPSR) [23] is one of CS-based reconstruction algorithms. GPSR exhibits good performance both in terms of reconstruction accuracy and computation time so it is adopted in this paper.

4   Experimental Results

Our lens-free coded aperture camera is implemented as below. The masks are printed on transparencies, and substitute the original optical lens set in the Holga lens kit. Then, the Holga lens kit is joined to a Canon EOS 600D camera base (maximum 5184×3456 pixels) pixels to form a lens-free coded aperture camera. Due to the limitation of the aperture of the Holga lens kit, the dimensions of the masks must be smaller than $10 \times 10mm$. So, two types of coded aperture patterns are tested: the 19-by-17 URA mask (Fig. 5(a,b)) and the 20-by-20 SRM mask (Fig. 5(c,d)). The Cannon sensor array has a minimum pixel pitch of $0.0043mm$. The laser printer has a precision of $1200dpi$. But, if taking into account the
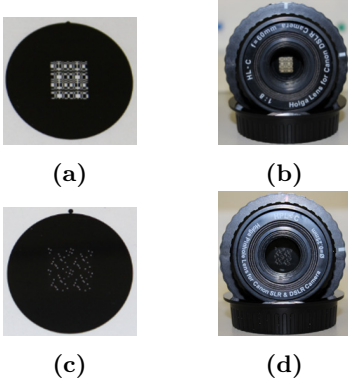
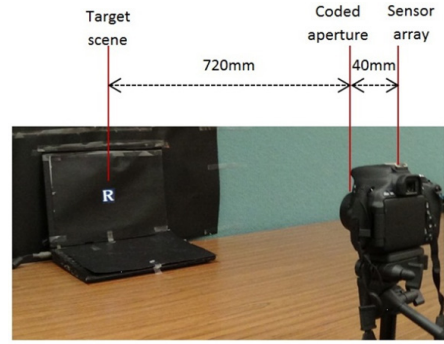**Fig. 5.** (a, b): URA mask; (c, d): SRM mask

Fig. 6. The test site

ink diffusion, the masks probably obtain a precision of about $0.03mm$, which is approximately equal to 7 pixel width. Therefore, the element side length is chosen to be $0.24mm$ because 1) it is close to the optimum pinhole size of $0.23mm$; 2) it is advantageous to align the sensors and the mask elements. Each element can be regarded as a combination of 8×8 sub-elements. The final masks (cyclic version) are approximately $9.12 \times 8.16mm$ for URA and $9.6 \times 9.6mm$ for SRM. The desired sensor array sizes are respectively $4.56 \times 4.08mm$ and $4.8 \times 4.8mm$. The Canon camera base provides a sufficient sensor array of size $22.3 \times 14.9mm$. The remaining sensors are useful for calibration (e.g., alignment).

Our test site is shown as Fig. 6. The target picture (Fig. 7(a)) displays on a laptop. In order to reduce the noise from the light sources outside the fully coded field of view (FCFV) [16], black papers are used to cover the surroundings of the target picture. In addition, we turn off all lights and make the room as dark as possible. We test the experiments with different shutter speeds and ISO speeds, and the best results present in Fig 7. We compare two test cases: 1) with interleaving and without interleaving, 2) the RUA mask with balanced correlation or generalized inverse and the SRM mask with GPSR. All the PSFs only take into account geometric optics. The experiments are tested under nearly the same conditions. Except the diffraction and interference effects, the noises and physical limitations are assumed the same for both types of masks.

The empirical results show that a reduction in diffraction and interference effects can be clearly observed for the SRM mask. Specifically, each transparent element in the SRM mask just acts as an independent pinhole camera to form a clear image on the sensor array (e.g., Fig. 7(c)), and the recorded picture is a superposition of multiple pinhole images with few diffraction and interference effects. By contrast, the interference effects severely impact the imaging quality for the URA mask due to its densely distributed transparent elements. The interleaving coded aperture technique apparently improves the image resolutions and qualities for both types of masks. The improvement for the URA and SRM masks are respectively from 19×17 to 152×136 pixels and from 20×20 to

160×160 pixels. However, the SRM mask can better preserve the image texture than the URA mask, such as the curve part of R in Fig. 7(d). Currently, the coded aperture camera cannot achieve good quality for capturing complex images. The major limitations include that the transparencies are not ideally transparent (open) or opaque (closed) and that the mask and sensor array suffer alignment error.
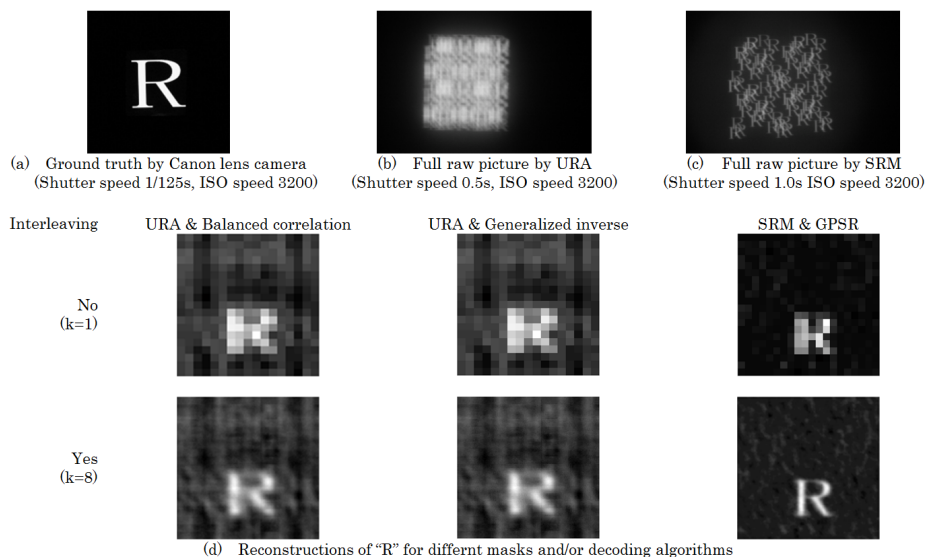

(a) Ground truth by Canon lens camera   (b) Full raw picture by URA   (c) Full raw picture by SRM
(Shutter speed 1/125s, ISO speed 3200)   (Shutter speed 0.5s, ISO speed 3200)   (Shutter speed 1.0s ISO speed 3200)

(d) Reconstructions of "R" for differnt masks and/or decoding algorithms

Fig. 7. Physical test results

5   Conclusion

This paper proposes an interleaving and sparse coded aperture technique for lens-free visible imaging. Compared with the well-known URA mask, the proposed SRM mask significantly reduces the diffraction and interference effects and improves the image quality. Moreover, the PSF in the proposed imaging model can be analyzed only on the basis of geometric optics, rather than being measured physically. Besides, the SRM mask facilitates the fabrication. Conventional linear decoding algorithms are inapplicable to the SRM mask decoding, so a CS-based nonlinear reconstruction approach is adopted. Finally, a prototype of lens-free camera is implemented to demonstrate the proposal. In contrast to the pinhole camera, the proposed camera is promising to widen the FOV, reduce the exposure time and increase the SNR whilst retaining the same image quality. In contrast to the lens-based camera, the proposed camera is inferior in terms of image sharpness, but it possesses several distinct merits, such as wide FOV, infinite depth of field and negligible image distortions.

References

1. Dicke, R.H.: Scatter-Hole Cameras for X-Rays and Gamma Rays. The Astrophysical Journal 153, L101–L106 (1968)
2. Gottesman, S.R.: Coded apertures: past, present, and future application and design. In: Proc. SPIE, vol. 6714, pp. 1–11 (2007)
3. Fenimore, E.E., Cannon, T.M.: Coded Aperture Imaging with Uniformly Redundant Arrays. Appl. Opt. 17, 337–347 (1978)
4. Slinger, C., Eismann, M., Gordon, N., Lewis, K., McDonald, G., McNie, M., Payne, D., Ridley, K., Strens, M., De Villiers, G., Wilson, R.: An investigation of the potential for the use of a high resolution adaptive coded aperture system in the mid-wave infrared. In: Proc. SPIE, vol. 6714, pp. 1–12 (2007)
5. Ridley, D., Villiers, D., Payne, A., Wilson, A., Slinger, W.: Visible band lens-free imaging using coded aperture technique. In: Proc. SPIE, vol. 7468, pp. 1–10 (2009)
6. Gottesman, S.R., Isser, A., Gigioli, J.G.W.: Adaptive coded aperture imaging: progress and potential future applications. In: Proc. SPIE, vol. 8165, pp. 1–9 (2011)
7. Byard, K.: Index class apertures-a class of flexible coded aperture. Appl. Opt. 51, 3453–3460 (2012)
8. Gottesman, S.R., Schneid, E.J.: PNP - A New Class of Coded Aperture Arrays. IEEE Trans. Nucl. Sci. 33, 745–749 (1986)
9. Gourlay, A.R., Stephen, J.B.: Geometric coded aperture masks. Appl. Opt. 22, 4042–4047 (1983)
10. Gottesman, S.R., Fenimore, E.: New family of binary arrays for coded aperture imaging. Appl. Opt. 28, 4344–4352 (1989)
11. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory 52, 1289–1306 (2006)
12. Candès, E.J.: Compressive sampling. In: Proceedings of the International Congress of Mathematicians, Madrid, Spain, pp. 1433–1452 (2006)
13. Wagadarikar, A., John, R., Willett, R., Brady, D.: Single disperser design for coded aperture snapshot spectral imaging. Appl. Opt. 47, 44–51 (2008)
14. Marcia, R.F., Harmany, Z.T., Willett, R.M.: Compressive coded aperture imaging. In: Proc. SPIE, vol. 7246, pp. 1–13 (2009)
15. Llull, P., Liao, X., Yuan, X., Yang, J., Kittle, D., Carin, L., Sapiro, G., Brady, D.: Coded aperture compressive temporal imaging. Opt. Express 21, 526–545 (2013)
16. Caroli, E., Stephen, J.B., Dicocco, G., Natalucci, L., Spizzichino, A.: Coded Aperture Imaging in X-Ray and Gamma-Ray Astronomy. Space Sci. Rev. 45, 349–403 (1987)
17. Young, M.: Pinhole Optics. Appl. Opt. 10, 2763–2767 (1971)
18. Berinde, R., Indyk, P.: Sparse recovery using sparse random matrices. MIT-CSAIL Technical Report (2008)
19. Born, M., Wolf, E.: Principles of optics: electromagnetic theory of propagation, interference and diffraction of light. Cambridge U. Press (1999)
20. Duarte, M.F., Davenport, M.A., Takhar, D., Laska, J.N., Sun, T., Kelly, K.F., Baraniuk, R.G.: Single-pixel imaging via compressive sampling. IEEE Signal Process. Mag. 25, 83–91 (2008)
21. Candes, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Communications on Pure and Applied Mathematics 59, 1207–1223 (2006)

22. Bajwa, W.U., Haupt, J.D., Raz, G.M., Wright, S.J., Nowak, R.D.: Toeplitz-Structured Compressed Sensing Matrices. In: Proceedings of IEEE/SP 14th Workshop on Statistical Signal Processing, pp. 294–298 (2007)
23. Figueiredo, M., Nowak, R., Wright, S.: Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems. IEEE J. Sel. Topics Signal Process. 1, 586–597 (2007)
24. Patsakis, C., Aroukatos, N.: LSB and DCT steganographic detection using compressive sensing. Journal of Information Hiding and Multimedia Signal Processing 5(1), 20–32 (2014)

Computational Intelligence Approaches for Digital Media Analysis and Description

Alexandros Iosifidis, Anastasios Tefas, and Ioannis Pitas

Department of Informatics,
Aristotle University of Thessaloniki
54124 Thessaloniki, Greece
{aiosif,tefas,pits}@aiia.csd.auth.gr

Abstract. This paper provides an overview of recent research efforts for digital media analysis and description. It focuses on the specific problem of human centered video analysis for activity and identity recognition in unconstrained environments. For this problem, some of the state-of-the-art approaches for video representation and classification are described. The presented approaches are generic and can be easily adapted for the description and analysis of other semantic concepts, especially those that involve human presence in digital media content.

Keywords: Digital Media Analysis, Digital Media Description, Human Action Recognition, Video Representation, Video Classification.

1 Introduction

Recent advances in technological equipment, like digital cameras, smart-phones, etc., have led to an increase of the available digital media, e.g., videos, captured every day. Moreover, the amount of data captured for professional media production (e.g., movies, special effects, etc) has dramatically increased and diversified using multiple sensors (e.g., 3D scanners, multi-view cameras, very high quality images, motion capture, etc), justifying the digital media analysis as a big data analysis problem. As expected, most of these data are acquired in order to describe human presence and activity and are exploited either for monitoring (visual surveillance and security) or for personal use and entertainment. Basic problems in human centered media analysis are face recognition, facial expression recognition and human activity recognition. According to YouTube statistics[1], 100 hours of video are uploaded by the users every minute. Such a data growth, as well as the importance of visual information in many applications, has necessitated the creation of methods capable of automatic processing and decision making when necessary. This is why a large amount of research has been devoted in the analysis and description of digital media in the last two decades.

In this paper a short overview on recent research efforts for digital media analysis and description using computational intelligence methods is given. Computational intelligence methods are very powerful in analyzing, representing and

---

[1] http://www.youtube.com/yt/press/statistics.html

classifying digital media content through various architectures and learning algorithms. Supervised, unsupervised and semi-supervised algorithms can be used for digital media feature extraction, representation and characterization. The specific problem that will be used as a case study for digital media analysis is the human-centered video analysis for activity and identity recognition. These two problems have received considerable research study in the last two decades and numerous methods have been proposed in the literature, each taking into account several aspects of the problem, relating to the application scenario under consideration. In this paper, we focus on the recognition of human activities in unconstrained environments, a problem which is usually referred to as *human action recognition in the wild*.

A pipeline, that most of the methods proposed in the literature follow, consists of two processing steps, as illustrated in Figure 1. In the first step, a process aiming at the determination of a video representation that, hopefully, preserves information facilitating action discrimination, is performed. In the second step, the previously calculated video representations are employed for action discrimination. These two processing steps are, usually, applied to a set of (annotated) videos, forming the so-called training video database. After training, a new (unknown) video can be introduced to the method and classified to one of the known classes appearing in the training video database. In the following sections, we describe some of the most successful and effective approaches that have been proposed for the two aforementioned processing steps.
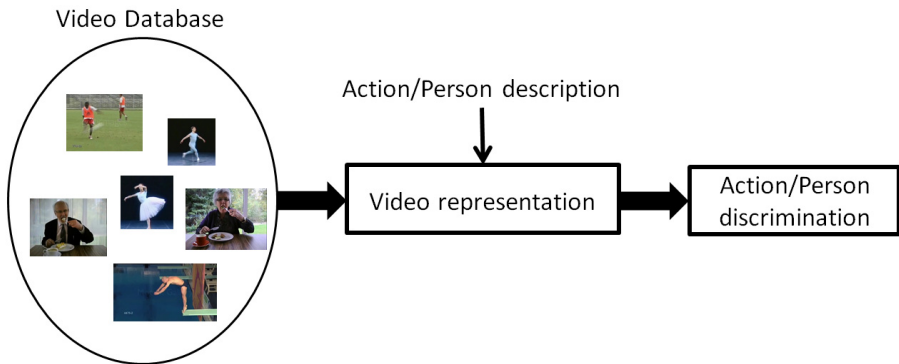

Fig. 1. Action recognition and person identification pipelineL

2   Problem Statement

Before describing the various approaches proposed for video representation and classification, we provide an overview of the problem. Let us assume that the training video database $\mathcal{U}$ consists of N_T videos depicting persons performing actions. Such videos will be noted as action videos hereafter. We employ the

different actions appearing in $\mathcal{U}$ in order to form an action class set $\mathcal{A}$. Similarly, the persons appearing in $\mathcal{U}$ are employed in order to form a person ID class set $\mathcal{P}$. Let us assume that the N_T action videos have been manually annotated, i.e., they have been characterized according to the performed action and/or the ID of the persons appearing in them. Thus, they are followed by an action class and a person ID label, α_i and h_i, $i = 1, \ldots, N_T$, respectively. We would like to employ the videos in $\mathcal{U}$, and the corresponding labels α_i, h_i in order to train an algorithm that will be able to automatically perform action recognition and/or person identification, i.e., to classify a new (unknown) video to an action and/or person ID class appearing in the action class set $\mathcal{A}$ and/or the person ID class set $\mathcal{P}$, respectively.

(a)   (b)

Fig. 2. Local video locations of interest: a) STIPs and b) video frame interest points tracked in consecutive video frames

3   Action Video Representation

Video representations proposed for action recognition and person identification problems exploit either global body information, e.g. binary silhouettes corresponding to the human body video locations [1–3], or shape and motion information appearing in local video locations of interest [4–6]. In the first case, videos are usually described by sets of binary images depicting the human body silhouettes during action execution. Such silhouettes are obtained by applying video frame segmentation techniques, like background subtraction or chroma keying. Due to this preprocessing step, such representations set several assumptions, like a relatively simple background and static cameras. However, such requirements are unrealistic in most cases. For example, most of the videos uploaded in video sharing websites (like Youtube) have been recorded by non-experts (users) in scenes containing cluttered backgrounds by using moving cameras. Another example can be given for movie productions, where the leading actor may perform an action in a scene containing several extras performing the same or different action.

Video representations belonging to the second category are able to operate in the above mentioned cases, since they are evaluated directly on the color

(grayscale) video frames and do not require video segmentation. Perhaps the most successful and effective local video representations have been designed around the Bag of Features model. According to this model, a video is represented by one or multiple vectors denoting the distribution of local shape and/or motion descriptors. These descriptors are calculated on local video locations corresponding either to Space Time Interest Points (STIPs) [7], or to video frame interest points that are tracked in successive video frames [5, 6], or to video frame pixels belonging to a pre-defined grid [8]. Example video frame locations of interest belonging to the first two categories are illustrated in Figure 2. STIPs determination is usually performed by applying extended versions of interest point detectors, like the Harris and Hessian ones.

The adopted descriptor types may be either handcrafted, or learned directly from data. Popular handcrafted descriptors include the Histogram of Oriented Gradients (HOG), the Histogram of Optical Flow (HOF) [9], the Motion Boundary Histogram (MBH) [5] and the Relative Motion Descriptor (RMD) [10]. Regarding data derived video representations exploiting local video information, a popular choice is to use overlapping 3D blocks, where the third dimension refers to time, in order to learn representative 3D blocks (filters) describing local shape and motion information. This is achieved by applying Deep Learning techniques, like the Independent Subspace Analysis (ISA) algorithm which has been proposed in the context of human action recognition [8]. Example filters learned by applying ISA on video frames depicting traditional Greek dances are illustrated in Figure 3.

Since the above mentioned descriptors contain complementary information, multiple descriptor types are usually employed for video representation. Let us denote by $\mathbf{x}_{ij}^d$, $j = 1, \ldots, N_i$, $d = 1, \ldots, D$ the descriptors (of type d) calculated for the i-th video in $\mathcal{U}$. D is the number of adopted descriptor types. We employ $\mathbf{x}_{ij}^d$, $i = 1, \ldots, N_T$, $j = 1, \ldots, N_i$ in order to determine a set of K_d descriptor prototypes forming the so-called codebook. This is achieved by clustering $\mathbf{x}_{ij}^d$, usually applying the K-Means algorithm [11], in K_d clusters and using the cluster mean vectors $\mathbf{v}_k^d$, $k = 1, \ldots, K_d$ as codebook vectors. After the determination of $\mathbf{v}_k^d$, each video is represented by D vectors obtained by quantizing $\mathbf{x}_{ij}^d$ according to $\mathbf{v}_k^d$. We denote by $\mathbf{s}_i^d \in \mathbb{R}^{K_d}$ the D vectors representing action video i. We would like to employ the action vectors $\mathbf{s}_i^d$ and the class labels α_i, h_i in order to train a classifier that will be able to automatically classify action videos to one of the classes appearing in $\mathcal{A}$ and/or $\mathcal{P}$.

Fig. 3. Filters learned by the ISA algorithm when trained on video frames depicting traditional Greek dances

4   Action Video Classification

After applying the above described process, each action video in $\mathcal{U}$ is represented by D vectors $\mathbf{s}_i^d$. By employing $\mathbf{s}_i^d$, $i = 1, \ldots, l$ and the corresponding class labels α_i (h_i), supervised learning techniques can be employed in order to discriminate the classes appearing in $\mathcal{A}$ ($\mathcal{P}$). This is usually achieved by training N_A (N_P) nonlinear Support Vector Machine (SVM) classifiers in an one-versus-rest scheme. In order to fuse the information captured by different descriptor types d, a multi-channel RBF-χ^2 kernel function is used, which has been shown to outperform other kernel function choices in BoF-based classification [12]:

$$[\mathbf{K}]_{i,j} = exp\left(-\frac{1}{A^d} \sum_{k=1}^{K_d} \frac{(s_{ik}^d - s_{jk}^d)^2}{s_{ik}^d + s_{jk}^d}\right).\tag{1}$$

A_d is a parameter scaling the χ^2 distances between the d-th action video representations and is set equal to the mean χ^2 distance between the training vectors $\mathbf{s}_i^d$.

Except SVM classifiers, Neural Networks (NNs) have been proven effective for the classification of action videos. Single-hidden Layer Feedforward Neural (SLFN) networks have been adopted for action recognition and person identification in [13–16]. A SLFN network consists of K_d input (equal to the dimensionality of $\mathbf{s}_i^d$), L hidden and N_A (N_P) (equal to the number of classes forming the classification problem) output neurons. In order to perform fast and efficient network training, the Extreme Learning Machine (ELM) algorithm has been employed in [14]. Typically, D NNs are trained, each for a different action video representation d, and network output combination is subsequently performed.

In ELM, the network's input weights $\mathbf{W}_{in}^d$ and the hidden layer bias values $\mathbf{b}$ are randomly assigned, while the output weights $\mathbf{W}_{out}^d$ are analytically calculated. Let us denote by $\mathbf{v}_j$ the j-th column of $\mathbf{W}_{in}^d$ and by $\mathbf{w}_k$ the k-th column of $\mathbf{W}_{out}^d$. For a given activation function $\Phi(\cdot)$, the output $\mathbf{o}_i^d = [o_1^d, \ldots, o_{N_A}^d]^T$ of the ELM network corresponding to training action vector $\mathbf{s}_i$ is calculated by:

$$o_{ik}^d = \sum_{j=1}^{L} \mathbf{w}_k^T \, \Phi(\mathbf{v}_j, b_j, \mathbf{s}_i^d), \ k = 1, \ldots, N_A.\tag{2}$$

By storing the hidden layer neurons outputs in a matrix $\boldsymbol{\Phi}_d$, i.e.:

$$\boldsymbol{\Phi} = \begin{bmatrix} \Phi(\mathbf{v}_1, b_1, \mathbf{s}_1^d) & \cdots & \Phi(\mathbf{v}_1, b_1, \mathbf{s}_{N_T}^d) \\ \cdots & \ddots & \cdots \\ \Phi(\mathbf{v}_L, b_L, \mathbf{s}_1^d) & \cdots & \Phi(\mathbf{v}_L, b_L, \mathbf{s}_{N_T}^d) \end{bmatrix},\tag{3}$$

Equation (2) can be written in a matrix form as $\mathbf{O}_d = \mathbf{W}_{out}^{d\,T}\boldsymbol{\Phi}_d$. Finally, by assuming that the network's predicted outputs $\mathbf{O}_d$ are equal to the network's desired outputs, i.e., $\mathbf{o}_i^d = \mathbf{t}_i$, and using linear activation function for the output neurons, $\mathbf{W}_{out}^d$ can be analytically calculated by $\mathbf{W}_{out}^d = \boldsymbol{\Phi}_d^{\dagger} \mathbf{T}^T$, where $\boldsymbol{\Phi}_d^{\dagger} =$

$\left(\mathbf{\Phi}_d \mathbf{\Phi}_d^T\right)^{-1} \mathbf{\Phi}_d$ is the Moore-Penrose generalized pseudo-inverse of $\mathbf{\Phi}_d^T$ and $\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_{N_T}]$ is a matrix containing the network's target vectors.

A regularized version of the ELM algorithm has, also, been used in [13, 15]. According to this, the network output weights $\mathbf{W}_{out}$ are calculated by solving the following optimization problem:

$$\textbf{Minimize:} \quad L_P = \frac{1}{2}\|\mathbf{W}_{out}^{d\,T}\|_F + \frac{c}{2}\sum_{i=1}^{N_T}\|\boldsymbol{\xi}_i\|_2^2 \tag{4}$$

$$\textbf{Subject to:} \quad \boldsymbol{\phi}_i^{d\,T}\mathbf{W}_{out}^d = \mathbf{t}_i^T - \boldsymbol{\xi}_i^T, \;\; i = 1, \dots, N_T, \tag{5}$$

where $\boldsymbol{\xi}_i$ is the training error vector corresponding to action vector $\mathbf{s}_i^d$, $\boldsymbol{\phi}_i^d$ denotes the i-th column of $\mathbf{\Phi}_d$, i.e., the $\mathbf{s}_i^d$ representation in the ELM space, and c is a parameter denoting the importance of the training error in the optimization problem. By substituting the condition (5) in (4) and solving for $\frac{\vartheta L_P}{\vartheta \mathbf{W}_{out}^d} = 0$, $\mathbf{W}_{out}^d$ can be obtained by:

$$\mathbf{W}_{out}^d = \left(\mathbf{\Phi}_d\mathbf{\Phi}_d^T + \frac{1}{c}\mathbf{I}\right)^{-1}\mathbf{\Phi}_d\mathbf{T}^T, \tag{6}$$

or

$$\mathbf{W}_{out}^d = \mathbf{\Phi}_d\left(\mathbf{\Phi}_d^T\mathbf{\Phi}_d + \frac{1}{c}\mathbf{I}\right)^{-1}\mathbf{T}^T. \tag{7}$$

where $\mathbf{I}$ is the identity matrix.

Exploiting the fact that the ELM algorithm can be considered to be a non-linear data mapping process to a high dimensional feature space followed by linear projection and classification, the Minimum Variance ELM (MVELM) and the Minimum Class Variance ELM (MCVELM) algorithms have been proposed in [16, 17] for action recognition. These two algorithms aim at simultaneously minimizing the network output weights norm and (within-class) variance of the network outputs. The network output weights $\mathbf{W}_{out}^d$ are calculated by solving the following optimization problem:

$$\textbf{Minimize:} \quad L_P = \frac{1}{2}\|\mathbf{S}_d^{1/2}\mathbf{W}_{out}^{d\,T}\|_F + \frac{c}{2}\sum_{i=1}^{N_V}\|\boldsymbol{\xi}_i\|_2^2 \tag{8}$$

$$\textbf{Subject to:} \quad \boldsymbol{\phi}_i^{d\,T}\mathbf{W}_{out}^d = \mathbf{t}_i^T - \boldsymbol{\xi}_i^T, \;\; i = 1, \dots, N_T, \tag{9}$$

and the network output weights are given by:

$$\mathbf{W}_{out}^d = \left(\mathbf{\Phi}_d\mathbf{\Phi}_d^T + \frac{1}{c}\mathbf{S}_d\right)^{-1}\mathbf{\Phi}_d\mathbf{T}^T. \tag{10}$$

$\mathbf{S}_d$ in (8), (10) is either the within-class scatter matrix $\mathbf{S}_w^d$ of the network hidden layer outputs, i.e., the representation of $\mathbf{s}_i^d$ in the so-called ELM space, or the total scatter matrix $\mathbf{S}_T^d$ in the ELM space. In the case of unimodal action classes

in the ELM space, $\mathbf{S}_w^d$ is of the form:

$$\mathbf{S}_w^d = \sum_{j=1}^{N_A} \sum_{i=1}^{N_T} \frac{\beta_{ij}}{N_j} (\boldsymbol{\phi}_i^d - \boldsymbol{\mu}_j^d)(\boldsymbol{\phi}_i^d - \boldsymbol{\mu}_j^d)^T. \tag{11}$$

In (11), β_{ij} is an index denoting if training action vector $\mathbf{s}_i^d$ belongs to action class j, i.e., $\beta_{ij} = 1$, if $c_i = j$ and $\beta_{ij} = 0$ otherwise, and $N_j = \sum_{i=1}^{N_T} \beta_{ij}$ is the number of training action vectors belonging to action class j. $\boldsymbol{\mu}_j^d = \frac{1}{N_j} \sum_{i=1}^{N_T} \beta_{ij} \boldsymbol{\phi}_i^d$ is the mean vector of class j in the ELM space.

In the case of multi-modal action classes, $\mathbf{S}_w$ is of the form:

$$\mathbf{S}_{w,CDA}^d = \sum_{j=1}^{N_A} \sum_{k=1}^{b_j} \sum_{i=1}^{N_T} \frac{\beta_{ijk}(\boldsymbol{\phi}_i^d - \boldsymbol{\mu}_{jk}^d)(\boldsymbol{\phi}_i^d - \boldsymbol{\mu}_{jk}^d)^T}{N_{jk}}. \tag{12}$$

Here, it is assumed that class j consists of b_j clusters, containing N_{jk}, $j = 1, ..., N_A$, $k = 1, ..., b_j$ action vectors each. β_{ijk} is an index denoting if action vector $\mathbf{s}_i^d$ belongs to the k-th cluster of action class j and $\boldsymbol{\mu}_{jk}^d = \frac{1}{N_{jk}} \sum_{i=1}^{N_T} \beta_{ijk} \boldsymbol{\phi}_i^d$ denotes the mean vector of the k-th cluster of class j in the ELM space.

Finally, $\mathbf{S}_T$ is given by:

$$\mathbf{S}_T^d = \sum_{i=1}^{N_T} (\boldsymbol{\phi}_i^d - \boldsymbol{\mu}^d)(\boldsymbol{\phi}_i^d - \boldsymbol{\mu}^d)^T, \tag{13}$$

where $\boldsymbol{\mu}^d = \frac{1}{N_T} \sum_{i=1}^{N_T} \boldsymbol{\phi}_i^d$ is the vector of the entire training set in the ELM space.

The information captured by different descriptor types is fused either by combining the network outputs corresponding to different descriptor types, e.g., by calculating the mean network output, or by optimally weighting the contribution of each NN according to combination weights $\boldsymbol{\gamma} \in \mathbb{R}^D$ by solving the following optimization problem:

$$\textbf{Minimize: } \mathcal{J} = \frac{1}{2} \sum_{d=1}^{D} \|\mathbf{W}_{out}^d\|_F^2 + \frac{c}{2} \sum_{i=1}^{N} \|\boldsymbol{\xi}_i\|_2^2 \tag{14}$$

$$\textbf{Subject to: } \left(\sum_{d=1}^{D} \gamma_d \mathbf{W}_{out}^{d\,T} \boldsymbol{\phi}_i^d \right) - \mathbf{t}_i = \boldsymbol{\xi}_i, \;\; i = 1, ..., N, \tag{15}$$

$$\|\boldsymbol{\gamma}\|_2^2 = 1, \tag{16}$$

An iterative optimization process consisting of two convex optimization problems has been proposed in [18] to this end.

By exploiting the fast and efficient ELM algorithm for SLFN network training, a dynamic classification scheme has been proposed for human action recognition in [19]. It consists of two iteratively repeated steps. In the first step, a non-linear mapping process for both the training action vectors and the test sample under

consideration is determined by training a SLFN network. In the second step, test sample-specific training action vectors selection is performed by exploiting the obtained network outputs corresponding to both the training action vectors and the test sample under consideration. These two steps are performed in multiple levels. At each level, by exploiting only the more similar to the test sample training action vectors, the dynamic classification scheme focuses the classification problem on the classes that should be able to discriminate. Considering the fact that after performing multiple data selections for a level $l > 1$ the cardinality of the training action vector set that will be used for SLFN network training will be very small compared to the dimensionality of the ELM space, the regularized version of ELM algorithm (6) has been employed in [19]. By using (7), the network output vector corresponding to $\mathbf{s}_i^d$ is obtained by:

$$\mathbf{o}_i^d = \mathbf{W}_{out}^{d\,T} \phi_i^d = \mathbf{T} \left(\mathbf{\Omega}_d + \frac{1}{c}\mathbf{I} \right)^{-1} \mathbf{K}_i^d, \tag{17}$$

where $\mathbf{K}_i^d = \mathbf{\Phi}_d^T \phi_i^d$, $\mathbf{\Omega}_d = \mathbf{\Phi}_d^T \mathbf{\Phi}_d$ are the kernel matrices corresponding to $\mathbf{s}_i^d$ and the entire SLFN training set, respectively. Thus, in this case the ELM space dimensionality is inherently determined by exploiting the kernel trick [24] and needs not be defined in advance. This ELM network training formulation also has the advantage that combined kernel matrices of the form (1) can be exploited.

The semi-supervised ELM (SELM) algorithms [23] has also been proposed for dynamic action classification in [22]. SELM solves the following optimization problem:

$$\textbf{Minimize: } \mathcal{J} = \|\mathbf{W}_{out}^{d\,T}\mathbf{\Phi}_d - \mathbf{T}\|_F \tag{18}$$

$$\textbf{Subject to: } \sum_{i=1}^{N_T} \sum_{j=1}^{N_T} w_{ij} \left(\mathbf{W}_{out}^{d\,T}\phi_i^d - \mathbf{W}_{out}^{d\,T}\phi_j^d \right)^2 = 0, \tag{19}$$

where w_{ij} is a value denoting the similarity between ϕ_i^d and ϕ_j^d. $\mathbf{W}_{out}^d$ is given by:

$$\mathbf{W}_{out}^d = \left(\left(\mathbf{J} + \lambda \mathbf{L}_d^T \right) \mathbf{\Phi} \right)^\dagger \mathbf{J}\mathbf{T}^T, \tag{20}$$

where $\mathbf{J} = diag(1,1,\ldots,0,0)$ with the first l diagonal entries as 1 and the rest 0 and $\mathbf{L}_d$ is the Graph Laplacian matrix [20] encoding the similarity between the training vectors ϕ_i^d.

In the dynamic classification scheme proposed in [22], test sample-specific training action vectors selection is performed by calculating the Euclidean distances between a given test sample and the training action vectors. The l training action vectors closest to the test sample are employed in order to form the labeled set of the SELM algorithm, while the remaining ones are used as unlabeled. Finally, the test sample under consideration is classified to the class corresponding to the highest SELM output.

Experimental results in real video data using all the previously presented methods can be found in the corresponding references. The results indicate that computational intelligence techniques can be used for solving difficult tasks, such as video analysis and semantic information extraction in digital media.

5   Conclusion

In this paper a survey on recent research efforts for digital media analysis and description based on computational intelligence methods has been presented. The specific problem that has been used as a case study is the human centered video analysis for activity and identity recognition. The presented approaches are generic and can be easily adapted for the description and analysis of other semantic concepts, especially those that involve human presence in digital media content.

References

1. Ji, X., Liu, H.: Advances in view-invariant human motion analysis: a review. IEEE Transactions on Systems Man and Cybernetics, Part-C 40(1), 13–24 (2010)
2. Iosifidis, A., Tefas, A., Nikolaidis, N., Pitas, A.: Multi-view Human Movement Recognition based on Fuzzy Distances and Linear Discriminant Analysis. Computer Vision and Image Understanding 116, 347–360 (2012)
3. Iosifidis, A., Tefas, A., Pitas, A.: View-invariant action recognition based on Artificial Neural Networks. IEEE Transactions on Neural Networks 23(3), 412–424 (2012)
4. Wang, H., Ullah, M., Klaser, A., Laptev, I.: Evaluation of local spatio-temporal features for action recognition. In: British Machine Vision Conference (2009)
5. Wang, H., Klaser, A., Schmid, C., Laptev, I.: Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision 103(1), 60–79 (2013)
6. Guha, T., Ward, R.: Learning Sparse Representations for Human Action Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 34, 1576–1588 (2011)
7. Laptev, I.: On space-time interest points. International Journal of Computer Vision 64(2), 107–123 (2005)
8. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.Y.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: Computer Vision and Pattern Recognition (2011)
9. Laptev, I., MarszalSchmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition (2008)
10. Oshin, O., Gilbert, A., Bowden, R.: Capturing the relative distribution of features for action recognition. In: Face and Gesture Workshop (2011)

11. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. Academic Press (2008)
12. Zhang, J., Marszalek, M., Lazebnik, M., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision 73(2), 213–238 (2007)
13. Iosifidis, A., Tefas, A., Pitas, I.: Person Identification from Actions based on Artificial Neural Networks. In: Symposium Series on Computational Intelligence (2013)
14. Minhas, R., Baradarani, S., Seifzadeh, S., Wu, Q.J.: Human action recognition using extreme learning machine based on visual vocabularies. Neurocomputing 73(10-12), 1906–1917 (2010)
15. Iosifidis, A., Tefas, A., Pitas, I.: Multi-view Human Action Recognition under Occlusion based on Fuzzy Distances and Neural Networks. In: European Signal Processing Conference (2012)
16. Iosifidis, A., Tefas, A., Pitas, I.: Minimum Class Variance Extreme Learning Machine for Human Action Recognition. IEEE Transactions on Circuits and Systems for Video Technology 23(11), 1968–1979 (2013)
17. Iosifidis, A., Tefas, A., Pitas, I.: Minimum Variance Extreme Learning Machine for Human Action Recognition. In: International Conference on Accoustics, Speech and Signal Processing (2014)
18. Iosifidis, A., Tefas, A., Pitas, I.: Multi-view Regularized Extreme Learning Machine for Human Action Recognition. In: Hellenic Conference on Artificial Intelligence (2014)
19. Iosifidis, A., Tefas, A., Pitas, I.: Dynamic action recognition based on Dynemes and Extreme Learning Machine. Pattern Recognition Letters 34, 1890–1898 (2013)
20. Belkin, M., Niyogi, P., Sindhawani, V.: Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. Journal of Machine Learning Research 7, 2399–2434 (2006)
21. Iosifidis, A., Tefas, A., Pitas, I.: Dynamic Action Classification based on Iterative Data Selection and Feedforward Neural Networks. In: European Signal Processing Conference (2013)
22. Iosifidis, A., Tefas, A., Pitas, I.: Active Classification for Human Action Recognition. In: IEEE International Conference on Image Processing (2013)
23. Liu, J., Cheng, Y., Liu, M., Zhao, Z.: Semi-supervised ELM with application in sparse calibrated locations estimation. Neurocomputing 74, 2566–2572 (2011)
24. Scholkpf, B., Smola, A.: Learning with kernels. MIT Press (2001)

Part V

Computational Systems and Its Applications

Register Allocation Based on Boolean Satisfiability

Yang Mengmeng and Liu Jie

Software College, Shenyang Normal University, 110034 Shenyang, China
`mengmengmelody.yang@gmail.com, nan127@sohu.com`

Abstract. Graph Coloring is an effective method which is used to solve the register allocation problem, it is also an NP-complete problem, heuristic algorithms and various evolutionary algorithms have been proposed in order to improve the performance of register allocation, in this paper, we propose to solve this problem by converting the graph coloring problem into Boolean Satisfiability problem (SAT), the experiments show that our algorithm can use fewer number of registers, which can improve the execution efficiency of the generated codes.

Keywords: Register Allocation, Graph Coloring, SAT Problem.

1   Introduction

Register allocation is an important stage before the compiler outputs code. It is the process of assigning the intermediate variable to the available registers. It decides the register that each variable should be put into, it is also related to the performance of the generated code, so many compilers using a variety of register allocation algorithms, hoping to improve the efficiency of the compiler, the traditional register algorithms are mostly based on graph coloring algorithm.

Given an undirected graph G (V, E), V represents the set of vertices, E represents the set of all edges. Color all the vertices of a graph G, and any two connected vertices cannot be colored by the same color, which is known as graph coloring. If the number of colors in a coloring scheme does not exceed k, it is called k coloring.

The idea of using graph coloring to solve register allocation problem was first proposed by Chaitin [1] et al. Graph G represents limitations and conflicts in the process of allocating, named register interference graph, all vertices of the graph represents the liveness range, and the edge between vertices represents the conflict among liveness range. Two conflicting lifetime cannot be active at the same time, that is, two vertices of each edge cannot use the same color. The problem that we need to solve is how to find a smallest possible value k to color graph G.

Various solutions have been proposed for graph coloring problem, Gend Lal Prajapati [2] proposed to divide the neighbor nodes of each vertex into two sets, set A contain vertices which are connected to the vertex w, and set B includes vertices which are not connected to the vertex w. Find the color which is different from all the colors that have been used in the connected vertices from the first element of the color

set to reduce the number of colors; WDChen [3] proposed to select the color value of the maximum potential node v first, and to use the smallest number of available color. Many heuristic algorithms [7-11] also have been used to solve the graph coloring problem , Carla Negri Lintzmaye [4] and others get a good solution with Ant Colony Optimization, the method ensures the quality of solution, but it costs a high running time. Later on they used evolutionary algorithms [5] to solve the same problem, on the base of maintaining the original performance the proposed algorithm reduced the running time and improve the efficiency of the algorithm. Raja Marappan [6] and others proposed to solve this problem using genetic algorithm, which applies fitness-proportionate selection method, Single Parent Conflict Gene crossover (SPCGX) and conflict-edge mutation operators to reduce the search space, but it requires several initial populations which affects the efficiency and quality. There are also some other methods to solve this problem [12-14].

In this paper, we propose to solve this problem by converting graph coloring problem into SAT problem and compare this method with the Kempe's algorithm, the experiments demonstrate that we can get a smaller k value through the proposed algorithm.

The second part describes the Kempe's algorithm; our algorithm is proposed in the third part; In part 4, we report our experimental results; the last part gives the conclusion.

2   Kempe's Algorithm

Kempe's algorithm is a classic algorithm which is used to solve the problem of k graph coloring, the specific steps are:

Step1: find a vertex whose degree is less than k, push this vertex onto the stack s, at the same time, move this vertex and its all vertices from the graph G.
Step2: Repeat step1 until all the nodes are piled into stack s and graph G becomes empty FIG.
Step3: Pop the vertex w from the stack s and give w a color that is different from the color of its neighbors.
Step4: Repeat step3 until all vertices in the stack are colored.
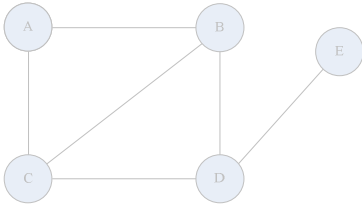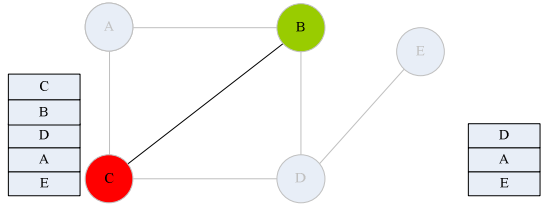Simple example:
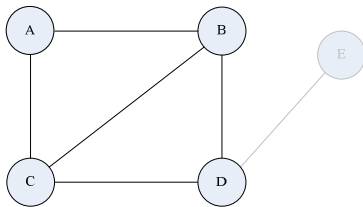
Fig. 1.
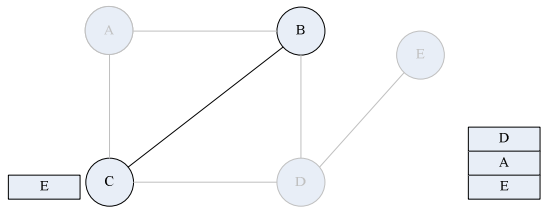
Fig. 2.
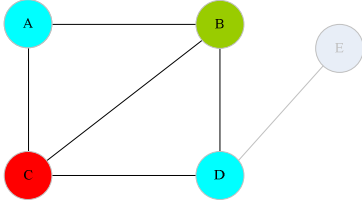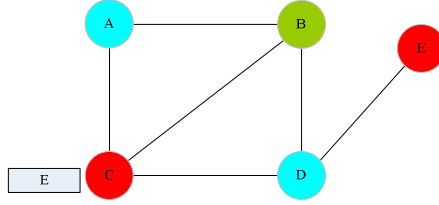
Fig. 3.

Fig. 4.

Fig. 5.

Fig. 6.

Theorem. If we can color a graph G=(V, E) using k colors based on Kempe's algorithm, then G is colorable using k colors.

Proof. We use induction on |V|.

(Base Case): Suppose |V| = 0. This theorem is trivially true.

(Induction Step): Suppose this theorem is true for all |V| = n-1, we want to show that this it is also true for |V| = n. According to our induction hypothesis, the first n-1 vertices can be colored successfully, assume that the vertex at the bottom of the stack is v_i. According to Kempe's algorithm, the degree of v_i is less than k. Let us also assume that (in the worst case) each of v_i's adjacent vertices is assigned a different color, we can still color vi using any of the remaining colors. Hence all of the vertices in V can be colored using k colors.

3   Our Algorithm

3.1   SAT Program

SAT problem is the satisfiability problem in propositional logic , given a CNF:

$$A = B_1 \wedge B_2 \wedge ... \wedge B_n \tag{1}$$

The form of clause :

$$Bi = b_1 \vee b_2 \vee ... \vee b_m \tag{2}$$

Where bi is a Boolean variable or the negation of the Boolean variable.

SAT problem is to find some assignment to variables so that the conjunctive normal form A is true. That is, if each clause of CNF A has at least one Boolean variable being true, A is true, otherwise A is false.

3.2   Specific Operation

Here are the specific steps of the proposed algorithm :
Step1: Convert register allocation problem into a register interference graph G (V, E);
Step2: According to the graph coloring constraints convert the graph into logical formulas and output in the form of a text file;

Where the constraints are:

1. Each vertex must select a color from the set {1,2 ... k};
2. Any two colors cannot be assigned to the same vertex, that is, each vertex can choose a unique color;
3.Two adjacent vertices cannot be assigned the same color.
Converted into a logical language:

$$1.\begin{cases} V(1,1) \lor V(1,2) \lor ... \lor V(1, \ K) \\ ... \\ V(n,1) \lor V(n,2) \lor ... \lor V(n, \ K) \end{cases} \tag{3}$$

$$2.\begin{cases} \neg V(1,1) \lor \neg V(1,2) \\ \neg V(1,1) \lor \neg V(1,3) \\ ... \\ \neg V(n, \ k-1) \lor \neg V(n,k) \end{cases} \tag{4}$$

If a and b are two connected vertices, then:

$$3.\begin{cases} \neg V(a,1) \lor \neg V(b,1) \\ ... \\ \neg V(a,k) \lor \neg V(b,k) \\ ... \end{cases} \tag{5}$$

Where V(i,j) represents the i-th vertex having the j-th color and n is the number of vertices.

Step3: Use minisat to solve the satisfiability problem.
The pseudo code of the algorithm:
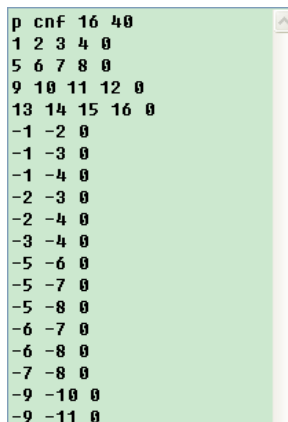
```
Given an undirected graph G = (V, E), k-coloring
if (There is an edge between V1 and V2)
    S[V1,V2]=1;
```

```
else
    S[Vi,Vj]=0;
FILE *fpt;
fpt = fopen("SAT.txt","w");
fprintf(fpt,"p cnf %d %d\n",N,M);
for i=0 to n-1 do                    //i represents vetex
  for c=1 to k do            //c represents color
    fprintf(fpt,"%d ",i*k+c);    fprintf(fpt,"0\n");
// Constraint 1
for i=1 to n-1 do
  for j=1 to k-1 do
      for c=j+1 to k do
        fprintf(fpt,"-%d -%d 0\n",i*k+j,i*k+c);
//Constraint 2
for i=0 to n-2 do       //n represents the number of vetices
  for j=i+1 to n-1 do
    if S[i][j]==1  then
      for c=1to k do           //c represents color
        fprintf(fpt,"-%d -%d 0\n",i*k+c,j*k+c);
//Constraint 3
fclose(fpt);
  Text document shown in figure 7:
```


Fig. 7. Text document

4   Experimental Results and Analysis

This paper compared the proposed algorithm with Kempe's algorithm, which is a classic graph coloring algorithm. We use C to implement these two algorithms and test the performance of the algorithms, the experimental results are shown in Table 1:

Table 1. Experimental results

vertex	edge	Maximum degree	Kempe's algorithm	Proposed algorithm
10	25	8	5	5
20	93	14	6	5
30	218	20	9	7
40	400	26	11	8
50	605	32	13	9
100	2478	64	19	16
150	5616	91	28	23

As can be seen from the table, compared to Kempe's algorithm the proposed algorithm can obtain a smaller k value, that is to say our algorithm can use fewer colors to color graph. With the increase of vertices, edges, and the degree, the proposed algorithm show more obvious advantages, that is, using our algorithm, the same number of registers can be used to store more variables to save storage space.

In addition to getting a smaller k value and using fewer colors to do graph coloring, our method is very simple, we just use minisat to solve the problem.

5   Conclusion

Register allocation is an important step in the process of compiling codes, the quality of the register allocation algorithm determines the quality of the compiled code and the efficiency of the compiler, graph coloring is widely used to solve the problem of register allocation, the traditional algorithms view graph coloring algorithm as a combinatorial optimization problems and use heuristic to solve it, while this paper converts constraints in graph coloring to Boolean formulas and solve the problem by existing sat solver. This method uses a smaller number of registers, which can effectively improve the performance of register allocation, improve the quality of the compiler output code, experimental results demonstrate the effectiveness of our algorithm.

References

1. Chaitin, G.J., Auslander, M.A., Chandra, A.K., Cocke, J., Hopkins, M.E., Markstein, P.W.: Register Allocation via Coloring. Computer Languages 6(1), 47–57 (1981)
2. Gend, L.P., Amit, M., Neha, B.: An Efficient Colouring of Graphs Using Less Number of Colours. In: World Congress on Information and Communication Technologies, pp. 666–669 (2012)
3. Chen, W.D.: New Algorithms for Graph Coloring Problem. Microcomputer Applications 25(4), 391–395 (2004)
4. Carla, N.L., Mauro, H.M., Anderson, F.: Register Allocation with Graph Coloring by Ant Colony Optimization. In: International Conference of the Chilean Computer Science Society, pp. 247–255 (2011)
5. Carla, N.L., Mauro, H.M., Anderson, F.: Register Allocation by Evolutionary Algorithm. In: International Conference of the Chilean Computer Science Society, pp. 207–215 (2012)
6. Raja, M., Gopalakrishnan, S.: A New Genetic Algorithm for Graph Coloring. In: International Conference on Computational Intelligence, Modelling and Simulation, pp. 49–54 (2013)
7. Dorrigiv, M., Markib, H.Y.: Algorithms for the graph coloring problem based on swarm intelligence. In: CSI International Symposium on Artificial Intelligence and Signal Processing (AISP), pp. 473–478 (2012)
8. Ge, F.Z., Wei, Z., Tian, Y.M., Huang, Z.J.: Chaotic ant swarm for graph coloring. In: IEEE International Conference on Intelligent Computing and Intelligent Systems (ICIS), pp. 512–516 (2010)
9. Yesil, C., Yilmaz, B., Korkmaz, E.E.: Hybrid local search algorithms on Graph Coloring Problem. In: International Conference on Hybrid Intelligent Systems (HIS), pp. 468–473 (2011)
10. Consoli, P., Collera, A., Pavone, M.: Swarm Intelligence heuristics for Graph Coloring Problem. In: IEEE Congress on Evolutionary Computation (CEC), pp. 1909–1916 (2013)
11. Fister, I., Brest, J.: Using differential evolution for the graph coloring. In: IEEE Symposium on Differential Evolution (SDE), pp. 1–7 (2011)
12. Falk, H.: WCET-aware register allocation based on graph coloring. In: ACM/IEEE Conference on Design Automation, pp. 726–731 (2009)
13. Mittal, A., Jain, P., Mathur, S., Bhatt, P.: Graph Coloring with Minimum Colors: An Easy Approach. In: International Conference on Communication Systems and Network Technologies (CSNT), pp. 638–641 (2011)
14. Wang, X.H., Qiao, Q.G.: Solving Graph Coloring Problems Based on a Chaos Nueral Network with Non-monotonous Activation Function. In: Fifth International Conference on Natural Computation, ICNC, vol. 1, pp. 414–417 (2009)

Robust Medical Image Watermarking Scheme with Rotation Correction

Lin Gao, Tiegang Gao, Guorui Sheng, and Shun Zhang

College of Software, Nankai University,Tianjin, China
{gao2689,gaotiegang}@gmail.com,
shengguorui@hotmail.com, zhangshun198612@163.com,

Abstract. In order to protect the image contents, many reversible medical image watermarking schemes have been proposed, although reversibility guaranteed the lossless of the cover image, but it also has some shortcomings such as it is vulnerable to geometrical attacks. So a novel robust watermarking scheme for medical image based on the combination Redundancy Discrete Wavelet Transform (RDWT) and Singular Value Decomposition (SVD) is proposed in this letter. Different from the reversibility, which guarantees the perceptional lossless, the proposed scheme achieves satisfied visual quality by exploiting the visually masking property of RDWT, in the meantime, Speeded-Up Robust Features (SURF) and Random sample consensus (RANSAC) based rotation correction scheme is put forward, which can be used to restore the attacked image to the original state. The experimental results show that the proposed scheme has the large amounts of embedding capacity; it is also robust against rotation attacks; and the perceptional quality of watermarked image meets the need of usage in medical images**.**

Keywords: Medical image watermarking, robust, rotation correction.

1   Introduction

Digital watermark has been widely applied to copyright protection of multimedia, content authentication, and transaction tracking. With the rapid growth of many standards of image storage (such like JPEG[1] and DICOM[2]), application of digital image used in medical became more commonly than before. As the digital images can be easily modified with ordinary software, which may lead to wrong diagnosis and treatment, the security of application for digital medical image has become a critical issue. In order to solve this problem, reversible image watermarking schemes have been proposed[3-6]. Reversible embedding has been introduced to these schemes to ensure that the cover image could be restored to the original state after the extraction of the watermark. However, these schemes are vulnerable to attacks, especially when the watermarked image is attacked with geometric attacks.

To resist geometric attacks against medical image watermark, some schemes applied to natural images can be used as reference[7-9]. However, these schemes could

not be used in medical image watermarking either because of the low embedding capacity or the unacceptable perceptual quality.

Recently, some researchers have suggested some watermarking schemes based on the combination of two or three transforms, which are named hybrid watermarking scheme[10-12]. Hybrid watermarking scheme supports high embedding capacity while keep the minimum distortion to the cover image. For example, S. Rastegar et al. applied radon transform to the host image, and then decomposed it into three levels using DWT[10]. The singular values of the third-level DWT sub-bands are modified by the singular values of the binary watermark. This scheme performed well against geometric attacks, but has a security weakness because SVD is applied on the watermark during embedding process[13]. C.-c. Lai et al. overcame this security problem by divide the watermark into two halves and embed the two halves directly into the singular values of LH and HL sub-bands of DWT transform of the host image[11]. Although this scheme fulfills the requirement of the security, the embedding capacity is still insufficient, and because of the down sampling operation are applied in every level of DWT, these two schemes also have weakness on imperceptibly, especially in medical image watermarking usage. Q. Li et al. tried to reduce the visual distortion by applying the HVS in the embedding scheme[12], this method still can't meet the requirement in medical image watermarking. To overcome the down sampling distortion in DWT, RDWT[14] is proposed and applied in the digital image watermarking[15]. The RDWT avoid the distortion of DWT introduced and has a shift-invariant property that DWT not has.

In this paper, the RDWT and SVD are introduced to implement hybrid watermarking in medical image. In the scheme, the host image is used as watermark. After extraction of watermark, extracted watermark is used to carry out the rotation correction for the possibly attacked image, thus, the original image can be restored to the original state.

The proposed scheme combined the hybrid watermarking scheme based on RDWT and SVD and the SURF[16] based image registration scheme. First the 1-level RDWT is performed on the host image to decompose the image in to four sub-bands. Unlike DWT, each of the RDWT sub-bands has the same size as the original image. Then SVD is performed on each sub-band and the host image itself is used as watermark to embed in each sub-band by directly modify the singular values. This embedding method meets the capacity, security and imperceptibly requirements. By embedding the host image as watermark, the rotate correction can be performed after the extraction. Because the proposed scheme has robustness against common geometric attacks, the rotation correction can be easily implemented between the watermarked image and the extracted watermark image.

The rest of this paper is organized as follows. The related works are reviewed in Section 2, the proposed scheme is presented in Section 3, and experimental results are given and discussed in Section 4. Finally the conclusion is drawn in Section 5.

2   Related Works

2.1   RDWT

Redundant Discrete Wavelet Transform (RDWT) is proposed and introduced into robust watermarking schemes for its shift variant property which DWT not has. RDWT can be implemented in many methods; the core idea of it is to avoid the down sampling process of DWT while keeping other advantages of it. The size of the each sub-band of the RDWT coefficients maintains the same size as that of the original image. This made the capture of the local texture of the RDWT domain can be done more precisely than in DWT domain. And more important, this feature made RDWT shift-invariant.

2.2   SVD

As an optimal matrix decomposition technique, SVD packs maximum signal energy into a few coefficients and has the ability to adapt to local variations of a given image. It has been widely used in image processing such as image compression, noise reduction and image watermarking. Several watermarking approaches have been proposed based on the SVD.

The general idea of SVD usage on image is shown in following. Suppose A is an image matrix with values in real number domain, then the SVD of A is defined as follows:

$$A = USV^T \tag{1}$$

where U and V are orthogonal matrices and S represents the diagonal matrix shown below:

$$S = \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix}$$

The singular values are diagonal elements $\left(\sigma_1, \ldots, \sigma_n \right)$ of the matrix S. And satisfy $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$

2.3   SURF

SURF is a novel scale and rotation invariant detector and descriptor. It uses an integer approximation to the determinant of Hessian blob detector, which can be computed extremely quickly with an integral image (3 integer operations). For features, it uses

the sum of the Haar wavelet response around the point of interest. Again, these can be computed with the aid of the integral image. SURF also has enough robustness against the attacks the watermark may encounter in this paper. An ideal descriptor and detector introducing in this paper's proposed scheme should have robustness against possible attacks and also should have fast calculation speed. As the name suggested, SURF has been design to offer robust at lowest time consuming. This makes the SURF a suitable detector and descriptor to be applied in the proposed scheme.

2.4   RANSAC

The RANSAC[17] is initially designed to interpret images in terms of a predefined set of models. Usually, these models contain gross errors. Instead of conventional smoothing techniques using as much of the data as possible to obtain an initial solution and then attempting to eliminate the invalid data points, RANSAC uses as small an initial data set as feasible and enlarges this set with consistent data when possible. In this paper, the feature points provided by the SURF matching process also have the same property as the RANSAC paradigm proposed with. The geometric attacks also make the set of matched SURF feature points contain gross errors. With properly selected thresholds, the RANSAC algorithm could estimate the free parameters used by attacker with strong confidence at acceptable resource and speed.

3   The Proposed Scheme

3.1   Watermark Embedding Algorithm

The detailed algorithm of embedding is given as the following:

(1) Perform 1-level RDWT on the cover image to decompose it into four sub-bands: LL, LH, HL and HH.

(2) Apply SVD to the four sub-bands (LL, HL, LH, HH), respectively. The formula is shown as follows:

$$A^i = U^i S^i V^{iT} \tag{2}$$

where, i indicates the sub-bands (LL, LH, HL, HH).

(3) Embed the watermark in each of the sub-bands by modify the singular values Si of them, and then apply SVD to each of the sub-bands as follows:

$$S^i + \alpha W = U_W^i S_W^i V_W^{iT} \tag{3}$$

where i indicate the sub-bands and α is the embedding factor, here, the α is set to 0.02 in the LL sub-band, and set to 0.005 in other sub-bands.

(4) Rebuild the new RDWT coefficients for each sub-band:

$$A^{inew} = U^i S_W^i V^{iT} \tag{4}$$

where, i indicate the sub-bands.

(5) Apply the inverse RDWT with the new RDWT coefficients to obtain the watermarked image Aw:

$$A_W = RDWT^{-1} \tag{5}$$

The embedding algorithm embedded the original image itself as watermark. The Uw, Vw and S corresponds to the respective RDWT sub-bands should be stored for the extraction of the watermark.

3.2 Watermark Extraction Algorithm

The detailed algorithm is given as follows:

(1) Apply RDWT on the possibly distorted watermarked image A_w^* to decompose it into four sub-bands LL, LH, HL, HH.

(2) Apply SVD to each sub-band:

$$A_W^* = U^{*i} S^{*i} V^{*iT} \tag{6}$$

where i indicate the sub-bands.

(3) Calculate

$$D^{*i} = U_W^i S^{*i} V_W^{iT} \tag{7}$$

where i indicate the sub-bands.

(4) Extract the watermark as following:

$$W^{*i} = \frac{D^{*i} - S^i}{\alpha} \tag{8}$$

where W^{*i} is the extracted watermark from level i.

The extraction algorithm extracts the original image embedded in it. By using the Uw, Vw and S corresponds to the respective RDWT sub-bands stored by the embedding algorithm, the watermark could be restored correctly. This guaranteed that the rotation correction scheme can be performed by the extracted watermark.

3.3 Rotation Correction Scheme

After the extraction of the watermark, the watermark image is used as the template for rotation correction of the possibly distorted image. The implementation of the proposed image correction scheme is given as follow:

1. Use the extracted watermark image as template and the watermarked image as distorted image, respectively.

2. Extract and match the SURF features from the template and distorted image, the matched feature points are formed into pairs for further processing.

3. Perform RANSAC on the matching point pairs; remove the outliers while computing the transform matrix. In the proposed scheme, the maximum number of random samplings is limited to 1000 and the desired confidence is set to 99.8.

4. Estimate the parameter the attacker possibly used by using the transform matrix.

5. Recover the distorted image using the transform matrix.

After the correction phase, the distorted image would be restored to the original state.

4   Experimental Results and Discussion

The proposed scheme is implemented by using MATLAB 2012a and tested with six medical images. All the images are gray scale images of size 512 by 512. To evaluate the visual quality of the watermarked image, the SSIM (Structural SIMilarity) [18]is used to evaluate the watermarked image.

To evaluate the robustness of the scheme, the NC is used for calculating the similarity between the original watermark and extracted watermark image. The NC between A and B can be defined as:

$$NC(A,B) = \frac{\sum_m \sum_n \left(A_{mn} - \overline{A}\right)\left(B_{mn} - \overline{B}\right)}{\sqrt{\left(\sum_m \sum_n \left(A_{mn} - \overline{A}\right)^2\right)\left(\sum_m \sum_n \left(B_{mn} - \overline{B}\right)^2\right)}} \qquad (9)$$

where m and n are pixel coordinates of image A and B, $\overline{A}$ and $\overline{B}$ are the mean value of image A and B, respectively. Generally, the NC is considered acceptable if it is 0.75 or above.

IM-0001-0022   IM-0001-0019   IM-0001-0018

IM-0001-0012   IM-0001-0006   IM-0001-0001
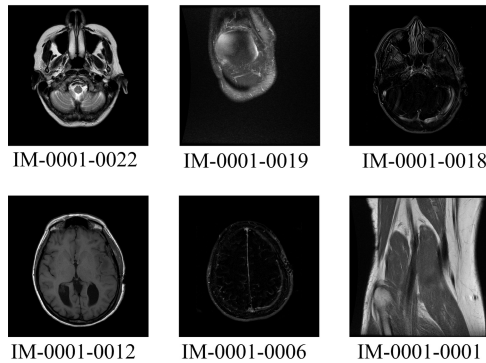
Fig. 1. Images used for test

4.1   Evaluation of the Watermarking Scheme

The proposed scheme is compared with two existing watermark scheme, one is the algorithm proposed by C.-c. Lai et al. , the other one is given by Q. Li et al..

Evaluation of Visual Quality

The SSIM is used for evaluation of three methods' visual quality. All of the three methods have the SSIM equal 1.0000 on the six test images. However, the NC values calculated in Fig.2 suggests that the proposed scheme has the best quality of the watermark while maintain the same visual quality as the other two schemes. It should be

note that the proposed scheme has double the embedding capacity than the other two schemes since the RDWT is shift-variant. This made the size of every RDWT sub-band has the same size as the original image. Comparing to the DWT sub-bands has half size as the original image; the proposed RDWT based scheme has doubled the embedding capacity than former DWT based schemes.
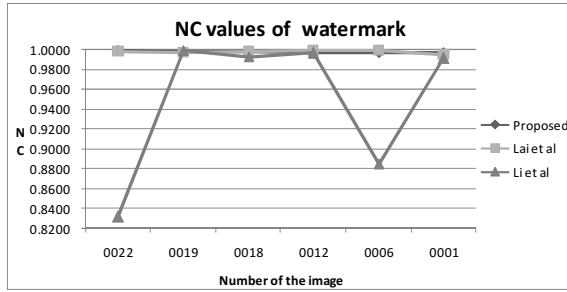

Fig. 2. NC values of watermarks extracted from unattacked image

Evaluation of Robustness

To evaluate the performance of the three methods, the watermarked images are rotated at specific counterclockwise angle and then the watermark is extracted and NC is calculated. Fig.3 shows the statistics of the NC value corresponding to the six test images.
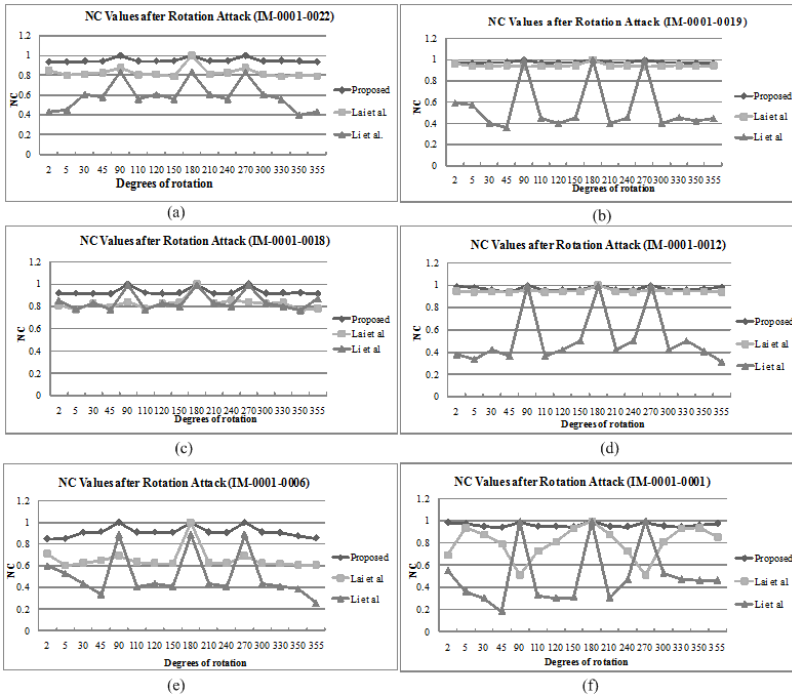
Fig. 3. NC values of watermarks extracted from attacked image

Fig.3 suggests that the proposed scheme is more robust against rotation attack than C.-c. Lai et al.'s and Q. Li et al.'s scheme. All the NC value of the mean watermark image extracted by the proposed scheme is above 0.8. This suggests that the proposed scheme is robust against rotation attack.

4.2   Evaluation of the Rotation Correction Scheme

Because the proposed watermarking scheme has ideal robustness against rotation and scaling attacks, the rotation correction can be implemented in the proposed scheme.

To evaluate the rotation correction scheme, the proposed scheme is implemented using MATLAB 2012a. The experimental results are given in Fig.4 and Fig.5.The rotate angle is the angle used in rotation attack, SSIM is induced to evaluate the visual similarity between the recovered image and the unattacked watermarked image. To evaluate the accuracy of the estimated correction angle, the AR (Accuracy Rate) is introduced. The AR is given by:

$$AR = 1 - \frac{|CA - RA|}{RA} \tag{10}$$

where CA is the estimate correction angle, RA is the rotate angle used in rotation attack.
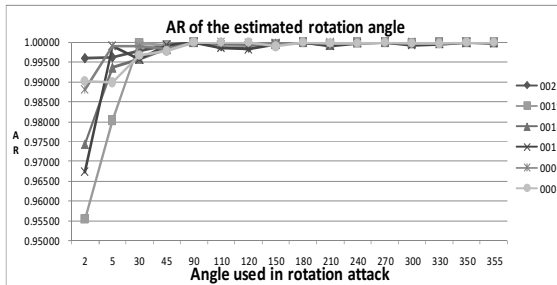
Fig. 4. AR of the estimated angle
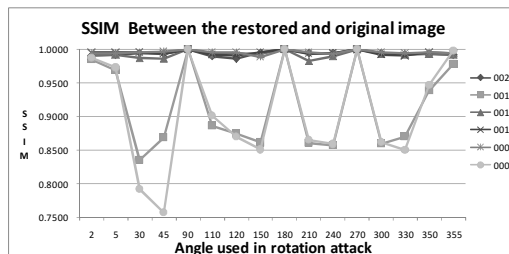
Fig. 5. SSIM between the restored and original image

Fig.4 suggests that the proposed scheme can restore the rotated image precisely. All of the AR values in Fig.4 are above 0.9 and most of them exceeded 0.98. This means that the estimated correction angle is precisely and efficient for using in rotation correction.

Fig.5 suggests that the recovered image has good visual similarity as the original unattacked watermarked image.

5 Conclusion

In this paper, a RDWT-SVD based hybrid watermarking scheme is proposed. The proposed scheme has large amounts of embedding capacity compared with some existing algorithm, and it also meets the need of robustness and high perceptual quality. Moreover, the proposed watermarking scheme has the ability of rotation correction function. Experimental results show that the proposed scheme can estimate the parameter of the attack rapidly and accurately, and also can restore the attacked image to the original state as good as possible.

References

1. Rabbani, M., Joshi, R.: An overview of the JPEG 2000 still image compression standard. Signal Processing:Image Communication 17, 3–48 (2002)
2. Rosslyn, S.: Digital Imaging and Communications in Medicine (DICOM) Part 5: Data Structures and Encoding (2011)
3. Fallahpour, M., Megias, D., Ghanbari, M.: Reversible and high-capacity data hiding in medical images. IET Image Processing 5, 190 (2011)
4. Weng, S., Pan, J.-S., Gao, X.: Reversible Watermark Combining Pre-processing Operation and Histogram Shifting. Journal of Information Hiding and Multimedia Signal Processing 3(4), 320–326 (2012)
5. Lin, S.-L., Huang, C.-F., Liou, M.-H., Chen, C.-Y.: Improving Histogram-based Reversible Information Hiding by an Optimal Weight-based Prediction Scheme. Journal of Information Hiding and Multimedia Signal Processing 4(1), 19–33 (2013)
6. Horng, G., Huang, Y.-H., Chang, C.-C., Liu, Y.: (k, n)-Image Reversible Data Hiding. Journal of Information Hiding and Multimedia Signal Processing 5(2), 152–164 (2014)
7. Lee, H., Lee, J., Kim, N., Kim, S.J., Shin, Y.G.: Robust feature-based registration using a Gaussian-weighted distance map and brain feature points for brain PET/CT images. Computers in Biology and Medicine 38, 945–961 (2008)
8. Na, W., Yamaguchi, K., Cedillo-Hernandez, M., Miyatake, M.N., Perez-Meana, H.: Robust Image Watermarking Using Feature Points and Image Normalization. In: 2010 IEEE Electronics, Robotics and Automotive Mechanics Conference, pp. 313–318 (2010)
9. Tsai, J.-S., Huang, W.-B., Kuo, Y.-H., Horng, M.-F.: Joint robustness and security enhancement for feature-based image watermarking using invariant feature regions. Signal Processing 92, 1431–1445 (2012)

10. Rastegar, S., Namazi, F., Yaghmaie, K., Aliabadian, A.: Hybrid watermarking algorithm based on Singular Value Decomposition and Radon transform. AEU - International Journal of Electronics and Communications 65, 658–663 (2011)
11. Lai, C.-C., Tsai, C.-C.: Digital Image Watermarking Using Discrete Wavelet Transform and Singular Value Decomposition. IEEE Transactions on Instrumentation and Measurement 59, 3060–3063 (2010)
12. Li, Q., Yuan, C.: Adaptive DWT-SVD domain image watermarking using human visual model. The Communication Technology, 1947–1951 (2007)
13. Ling, H.-C., Phan, R.C.-W., Heng, S.-H.: On the security of a hybrid watermarking algorithm based on singular value decomposition and Radon transform. AEU - International Journal of Electronics and Communications 65, 958–960 (2011)
14. Fowler, J.: The redundant discrete wavelet transform and additive noise. IEEE Signal Processing Letters 12, 629–632 (2005)
15. Lagzian, S., Soryani, M., Fathy, M.: Robust watermarking scheme based on RDWT-SVD: Embedding data in all subbands. In: 2011 International Symposium on Artificial Intelligence and Signal Processing (AISP), pp. 48–52 (2011)
16. Bay, H., Ess, A., Tuytelaars, T., Gool, L.V.: Speeded-up robust features (SURF). Computer Vision and Image Understanding 110(3), 346–359 (2008)
17. Fischler, M., Bolles, R.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM 24 (1981)
18. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing: A Publication of the IEEE Signal Processing Society 13, 600–612 (2004)

Design of Data Encryption Transmission System Based on FPGA⋆

Yan Yu, Bingbing Song, Xiaozhen Liu, Qun Ding, and Ziheng Yang

Electronic Engineering College, Heilongjiang University,
Harbin, China
{yuyanclever,songbingbing1988,15046658153}@163.com,
qunding@aliyun.com, Yzheda@126.com

Abstract. In this paper, the hardware system of data encryption and security transmission is studied. A kind of encryption nuclear is designed using chaotic sequences and stream cipher. In addition, a scheme of data encryption system based on FPGA is proposed, thus a new type of encryption system is obtained. Meanwhile, the simulation of the key generation circuit is accomplished. The experiment shows that this hardware encryption system can achieve the encryption and decryption. It ensures the secure transmission of data further.

Keywords: data encryption system, FPGA, secure transmission.

1 Introduction

In the early twentieth century, the famous scholar Ch. mo Enrico Zanoni proposed a new method to realize the chaotic encryption algorithm using FPGA [1]. In 1990s, chaos theory emerged in cryptography. Because of its prominent inherent features, it became a hot point in cryptography studies. Since Pecora and Carroll put forward secure communication based on chaos, chaos secure communication has become a hot topic in the field of electronic communication [2]. With the development of network and communication technology, the loopholes of network are more and more obvious. During the process of information transmission, the messages maybe intercepted and attacked by others. As a result, peoples privacy will be threatened. Nowadays, the security of information transmission has been perilous, so it is necessary to encrypt them [3]. In this paper, a data encryption transmission system is designed. Its core part of encryption is combining chaotic sequences and stream cipher. The plaintext is encrypted before sending it to the receiving terminal through the serial port. After receiving the ciphertext, the receiving terminal will decrypt it. So the reliability of data transmission is achieved.

The detail of the paper is organized as follow: firstly, the overall design concept of this data encryption transmission system is introduced; secondly, the design principle of this system is described, and after building the block diagram of Logistic chaotic module and A5/1 module, it is also simulated; thirdly, apply this encryption method to FPGA hardware device. Eventually, we draw a conclusion that this system can realize the performance of encryption and decryption, and it has good prospects.

2   Overall Scheme Design of Data Encryption System

2.1   Encryption System Design Principles

The data encryption transmission system based on FPGA consists of the transmitting terminal and the receiving terminal. The transmitting terminal includes the main control chip, the encryption nuclear and the serial communication module. Correspondingly, the receiving terminal includes the main control chip, the decryption nuclear and the serial communication module. In this scheme, the main control chip is Cyclone II EP2C8Q208C8N. Both sides communicate with each other by using the simplex communication. Firstly, before the transmitting terminal sends the ciphertext to the receiving terminal through the serial port, the plaintext is encrypted. And then the receiving terminal will decrypt the ciphertext received through the serial port. Thus, the correct plaintext is obtained. Eventually, the data encryption transmission based on FPGA is completed accomplished. It ensures the security and reliability of the information transmission. The principle diagram of the overall scheme is shown in Fig. 1.

Data transmission: firstly, sending the initial key to the transmitting terminal through the serial port module by the computer. Thus the key sequence generator module is started to generate the key sequence used for encryption. Then input
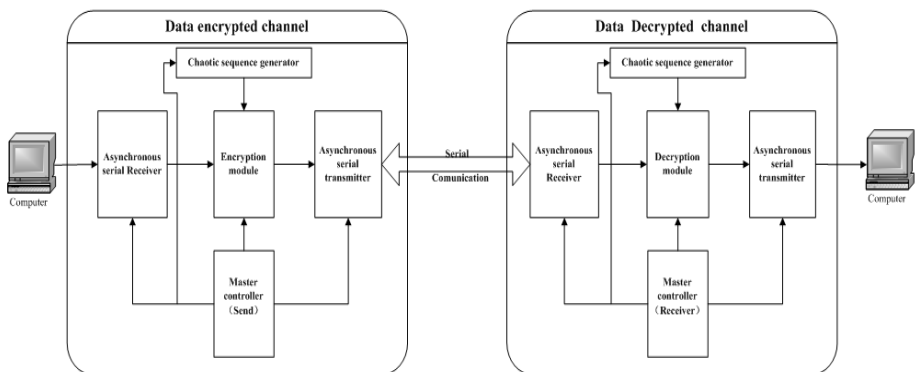

Fig. 1. Principle diagram of the overall scheme

the plaintext, and the process of encrypting the plaintext begins. The ciphertext is transmitted through the serial port.

Data acceptance: Firstly, the ciphertext is received through the serial port by the receiving terminal. Then input the initial key by the computer. Thus the key sequence genetrator module is started and the ciphertext is decrypted at the same time. Eventually, the information decrypted is obtained.

2.2   Key Sequence Generation

Overall Design. Encryption algorithms rely on the core theory of cryptography, security architecture, and security communication protocols of network. And the the selection and design of the key is very important. The encryption module and decryption module in this system combine the chaos theory with stream cipher. And the typical Logistic chaotic sequence and A5/1 are adopted [4,5,6,7]. A5/1 includes three linear feedback shift registers. The hardware design of A5/1 is simple, and its technology is relatively mature [8]. Because of the basic features of chaotic system, such as initial value sensitivity, pseudo-randomness, the chaotic sequence is chosen [9]. However, because the key bits and frame bits are linear filled when A5/1 is initialized, it is easily attacked [10]. When applying the chaotic sequence to digital communication after quantization, it will lose the chaotic characteristics, resulting in reducing its security [11,12]. So the key generation combines the chaotic sequence with A5/1. The block diagram of key generation is shown in Fig. 2.
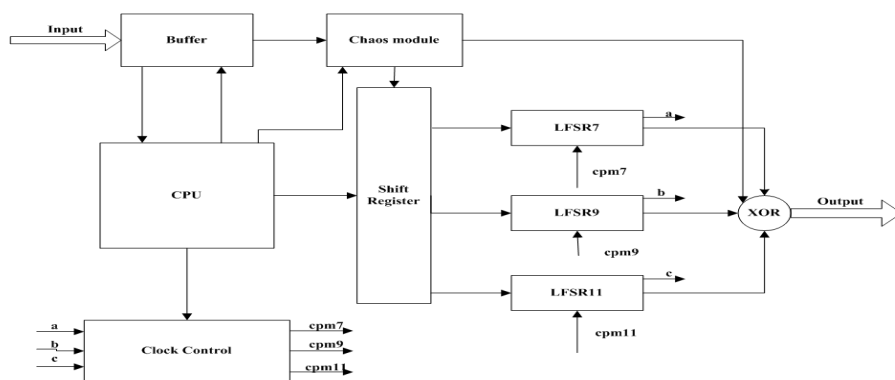

Fig. 2. Block diagram of key generation

As shown in Fig.2, the key generation consists of CPU module, buffer module, clock module, Logistic chaotic module, shift module, A5/1 module, and XOR module. The implementation of the whole process is that: firstly, input the initial key to the buffer module, then the CPU controls the buffer module to send the

key to Logistic module. Meanwhile, shift the Logistic outputs and they are sent to the three linear feedback shift registers. Finally, calculate the outputs of Logistic and linear feedback shift registers by XOR, thus the key sequence is obtained.

Design of Graphical Modular. The chaotic sequence is combined with A5/1, and its process is similar to A5/1. Because A5/1 is easy to control on hardware systems, and it also can generate pseudo-random sequences, so the reliability of information transmission can be better improved by combining these two methods [13]. According to block diagram of key generation, each sub-module is designed using VHDL in Quartus II environment, then use the modules to design the project. Logistic chaotic sequence module and A5/1 module is shown in Fig. 3 and Fig.4.
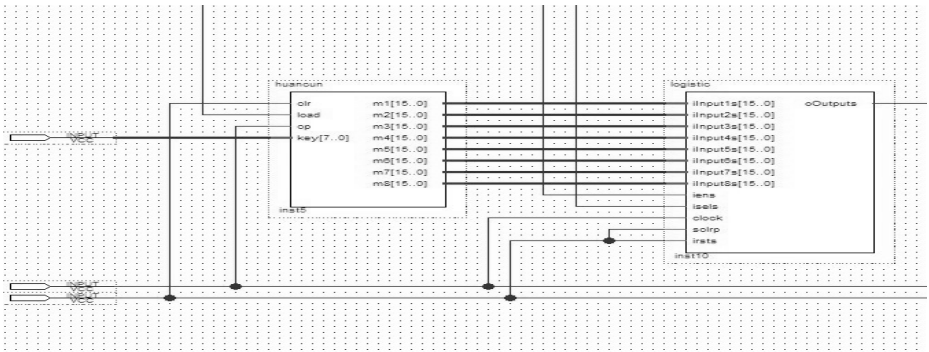
Fig. 3. Logistic chaotic sequence module
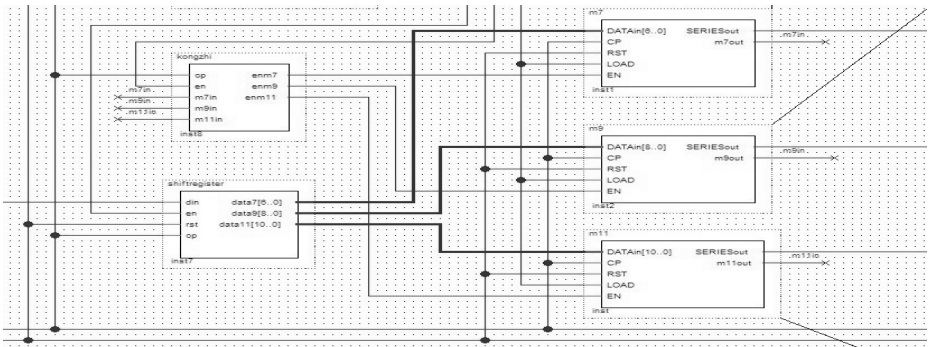
Fig. 4. A5/1 module

Waveform Simulation. After designing the project of the key generation in Quartus II environment, the waveform simulation can begin. The simulation result is shown in Fig.5.
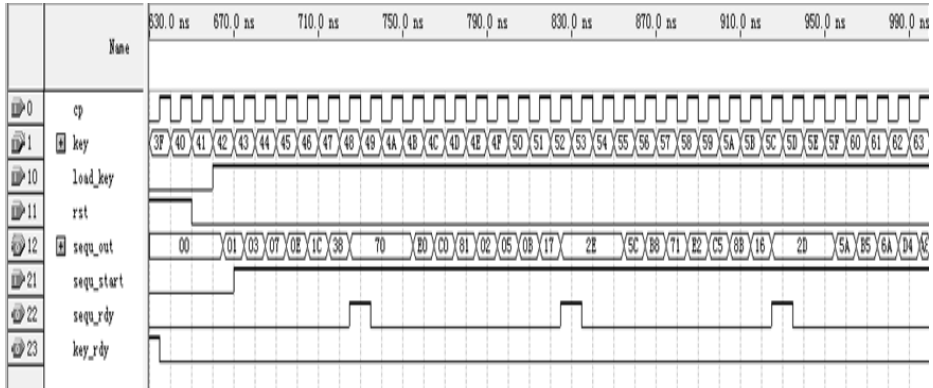

Fig. 5. Simulation result of the key generation

In Fig. 5, the input key is the initial key, load_key is enable signal, and sequ_out is the output sequence. From Fig. 5 we can find that different initial keys generate different output sequences. And it seems to be non-periodic. So using this method to encrypt information is reliable.

3   Design of Hardware Circuit

Since the software encryption is easily intercepted and deciphered during the transmission, so the information security is further threatened [14]. But this data encryption transmission system uses the combination of chaotic sequence and A5/1 to generate key sequences. What is more, this scheme is applied to FPGA hardware device. Therefore, the information transmission depends on the hardware completely. And this system has higher security and reliability. This system uses Cyclone II EP2C8Q208C8N [15]. It has lots of advantages, such as love power consumption, high density and low costs [16]. The hardware circuit is shown in Fig. 6.

Download the project file designed in Quartus II environment to FPGA, and start the whole system. Then we can observe plaintext and ciphertext. The plaintext and ciphertext sent by the transmitting terminal is shown in Fig. 7. Correspondingly, the ciphertext received and information after decrypting by the receiving terminal is shown in Fig. 8.

From the four figures we can find that the message the receiving terminal receives is as the same as the ciphertext the transmitting terminal sends. And the information the receiving terminal decrypts is also as the same as the plaintext the trans-
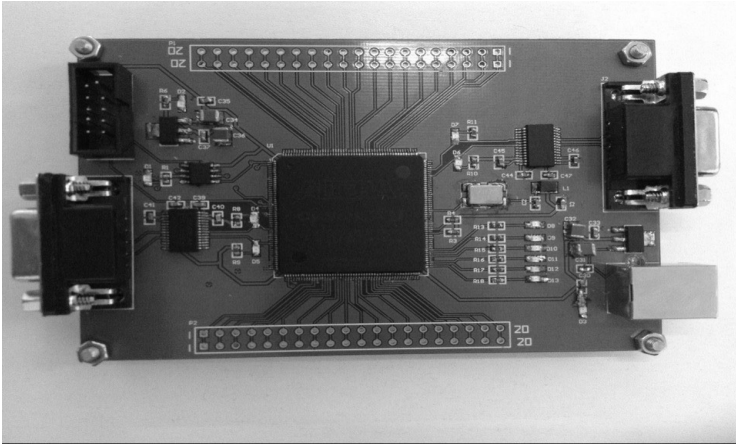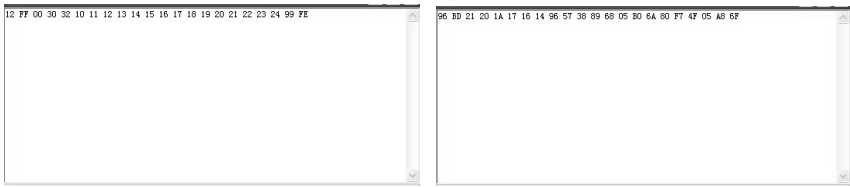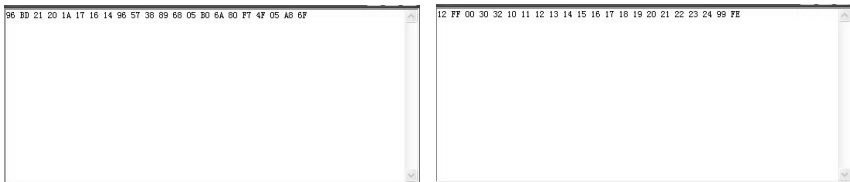
Fig. 6. Hardware circuit


```
12 FF 00 30 32 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 99 FE
```



```
96 BD 21 20 1A 17 16 14 96 57 38 89 68 05 B0 6A 80 F7 4F 05 A8 6F
```

(a) Plaintext sent by the transmitting terminal

(b) Ciphertext sent by the transmitting terminal

Fig. 7. Information sent by the transmitting terminal


```
96 BD 21 20 1A 17 16 14 96 57 38 89 68 05 B0 6A 80 F7 4F 05 A8 6F
```



```
12 FF 00 30 32 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 99 FE
```

(a) Plaintext sent by the receiving terminal

(b) Ciphertext sent by the receiving terminal

Fig. 8. Information sent by the receiving terminal

mitting terminal transmits. So we can draw a conclusion that this data encryption transmission system can achieve the performance of encryption and decryption.

4   Conclusion

This paper presents a data encryption system based on FPGA. It applies a method to generate key sequences. That is combining Logistic sequence and

A5/1. Logistic sequence is used to send the initial key to A5/1. This makes up the limit of the initial key of stream cipher. What is most important is that this method is applied to FPGA hardware device. It is not limited to upgrade the algorithm. And convenient conditions are provided to study other algorithms. The experimental result shows that this scheme has high reliability and security. The system has been repeatedly tested to ensure the security of data transmission. The transmission speed is also optimized. So it has better future.

References

1. Huang, Z.E., Yu, S.M., Zhou, W.J.: Chaotic Digital Image Encryption and Its Hardware Implementation based on FPGA Technology. Communication Technology 12, 343–346 (2008)
2. Pu, C.L., Li, W.X., Lin, J.G.: A Kind of Chaotic Synchronization System and Its Application for Secure Communications. Bulletin of Science and Technology 22, 842–845 (2006)
3. Wang, L., Sun, H.J., Zhao, J.: Application of the Data Encryption Technology in the Field of Computer Network Security. Telecom Power Technology 30, 54–57 (2013)
4. Zhang, A.H., Jiang, Z.Q.: Improving for Chaotic Image Encryption Algorithm Based on Logistic Mapping. Nanjing University of Posts and Telecommunications (Natural Science) 29, 69–73 (2009)
5. Fan, J.L., Zhang, X.F.: Piecewise Logistic Chaotic Map and Its Performance Analysis. Acta Electrorica Dinica 37, 720–724 (2009)
6. Liu, P., Yan, C., Huang, X.G.: Optimized Method of Generating the Spread-Spectrum Sequences Based on Logistic-Map. Communications 28, 134–138 (2007)
7. Li, L., Han, W.B., Wang, Z.: The Research of Parameters Selection on the Model of Time-Memory-Data Trade-Off Attack to A5/1. Electronics & Information Technology 34, 1911–1915 (2012)
8. Ding, Q., Peng, X.Y., Yang, Z.H.: The Cipher Chip of Combining Stream Based on the Neural Network Algorithm. Acta Electronica Sinica 34, 409–412 (2006)
9. Guan, C.Y., Gao, F.: An Encryption Algorithm Based on Chaotic Sequences. Beijing Institute of Technology 23, 363–366 (2003)
10. Chen, W., Hu, Y., Yang, Y.X.: Application of Correlation Attack in Algorithm Identify. Electronics & Information Technology 28, 828–831 (2006)
11. Liu, X.Y., Qiu, S.S., Huang, G.Z., Fan, Y.: An Overview of Chaos-based Digital Communication Systems. Telecommunication Engineering, 1–9 (2005)
12. Wu, H., Ding, Q., Zhou, P.: Logistic Chaotic Sequence Design and Application in Data Encryption. Scientific Instrument 30, 372–375 (2009)
13. Chen, W., Yang, Y.X., Niu, X.X.: Improvement of A5/1 Algorithm Against Correlation Attack. Beijing University of Posts and Telecommunication 29, 120–123 (2006)
14. Li, Z.Q., Sun, X.X., Du, C.B.: Hardware Design and Implementation of Wi-Fi Technology Based Encryption Systerm. In: 2013 International Conference on SNS&PCS, vol. 5, pp. 144–147 (2013)
15. Song, J.Y., Mo, X.K., Niu, Q.J.: Design of JTIDS Message Restorer Based on FPGA. Fire Control&Command Control 37, 153–156 (2012)
16. Niu, Q.J., Mo, X.K., Zhao, W.J., Song, J.Y.: Design of JTIDS Transmission Code Generator Based on FPGA. Telecommunication Engineering 50, 43–46 (2010)

An Efficient Image Encryption Scheme Based on ZUC Stream Cipher and Chaotic Logistic Map

Hai Cheng[1], Chunguang Huang[1], Qun Ding[*], and Shu-Chuan Chu[2]

[1] Heilongjiang University,Key Laboratory of Electronics Engineering, China
[2] School of Computer Science,
Engineering and Mathematics Flinders University, Australia
chenghaihd@163.com,dahuangr@163.com,qunding@aliyun.com

Abstract. Digital color image encryption is different from text encryption because of some inherent features of image such as huge data capacity and high correlation among the neighboring pixels. Because of the desirable cryptographic properties of the chaotic maps such as sensitivity to initial conditions and random-like behave, more and more researches use these properties for encryption. This paper proposed an efficient image encryption scheme. Logistic chaos-based stream cipher is utilized to permute the color image. The MD5 hash function and the ZUC stream cipher algorithm are combined to diffusion the color image. Theoretical and experimental analyses both confirm the security and the validity of the proposed algorithm.

Keywords: ZUC stream cipher, logistic chaotic map, color image encryption.

1 Introduction

With the great development of network and information technology, multimedia technology and its applications especial digital images is wildly used in the computer network. It has become important and necessary to protect the information security issues against illegal copying when the color image transmitted through over the internet and wireless networks. To meet the challenge, a variety of encryption schemes have been proposed. DES, RSA, AES, RC5[1] and other popular encryption method can be used to encrypt the digital color images. Instead of using traditional block cipher for image encryption, chaotic logistic map becomes popular new days.

Many researchers have used chaotic algorithm in image encryption schemes[2-5]. Pareek[6] et al. presented a new method to encrypt the image based on chaotic logistic map. Kwok and Tang[4] presented a fast chaos-based image encryption

system with a stream cipher structure based on a new pseudo-random. Gao[7] proposed a new image total shuffling matrix to shuffle the position of image pixels. Then relationship between plain image and cipher image is confused. Behnia[8] et al. proposed an implementation of digital image encryption scheme based on the mixture of chaotic system.

In this paper, a digital color image encryption using then combination of ZUC stream cipher [9] and chaotic logistic map function. The logistic chaotic map is used to permute the digital color image because of sensitivity to the initial value. The MD5 hash function [10] is also used to generate the 128-bit initial vector of ZUC stream cipher because of the sensitivity to the tiny change of the image. And the ZUC stream cipher is used to diffuse the permuted image.

This paper will be arranged as follows. In Section 2 the ZUC stream cipher is proposed. The proposed image cryptosystem is mentioned in section 3. In Section 4, performance of proposed encryption method is evaluated. Finally, some concluding remarks are drawn is Section 5.

2   ZUC Stream Cipher

ZUC is a new stream cipher due for possible inclusion in the Long Term Evolution standards for mobile devices. The ZUC algorithms are the new cryptographic algorithms recommended by CCSA to be used in 3GPP LTE (Long Term Evolution). And they have been made the work item by 3GPP SA3.The ZUC algorithms have been evaluated by the algorithm standardization group ETSI SAGE, and also by two other teams of eminent experts, and are believed to be strong and suitable for LTE.

ZUC is a word-oriented stream cipher. A 128-bit initial key and a 128-bit initial vector is used as input, and a keystream of 32-bit words is generated. This keystream can be used for encryption and decryption.

The execution of ZUC has two stages: initialization stage and working stage. In the first stage, a key and IV initialization are performed. During the first stage, the cipher is clocked without producing output. During the second stage, a 32-bit word of output is produced within every clock pulse.

2.1   Algorithm Description

ZUC has 3 logical layers which is shown in Fig. 1. The top layer is a linear feedback shift register (LFSR) of 16 stages. The middle layer is for bit-reorganization (BR), and the bottom layer is a nonlinear function F.

2.2   The Linear Feedback Shift Register

The linear feedback shift register (LFSR) has 16 cells. Each cell $S_i (0 \leq i \leq 15)$ has 31 bits. And each of them is restricted to take values from the following set $\{1, 2, 3, ..., 2^{32} - 1\}$.
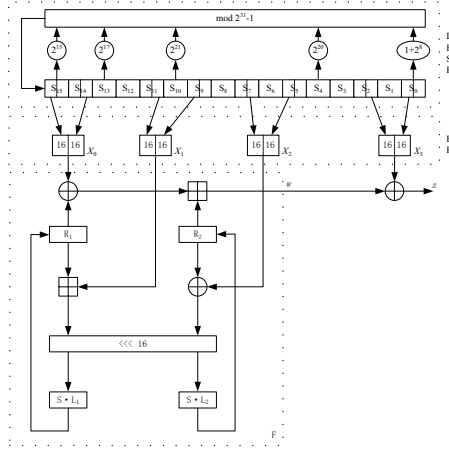
Fig. 1. General structure of ZUC

The linear feedback shift register has 2 mode of operations: the initialization mode and the working mode.

In the initialization mode, the LFSR receives a 31-bit input word u , which is obtained by removing the rightmost bit from the 32-bit output W of nonlinear function F , i.e., $u = W >> 1$. More specifically, the initialization mode works as follows:

LFSRWithInitialisationMode(u)

{

    1. $v = 2^{15}S_{15} + 2^{17}S_{13} + 2^{21}S_{10} + 2^{20}S_4 + (1 + 2^8)S_0 \bmod (2^{31} - 1)$;
    2. $S_{16} = (v + u) \bmod (2^{31} - 1)$;
    3. *if* $S_{16} = 0,$ *then set* $S_{16} = 2^{31} - 1$;
    4. $(S_1, S_2, ..., S_{15}, S_{16})-> (S_0, S_1, ..., S_{14}, S_{15})$.

  }

In the working mode, the LFSR does not receive any input, and it works as follows:

  {

    1. $S_{16} = 2^{15}S_{15} + 2^{17}S_{13} + 2^{21}S_{10} + 2^{20}S_4 + (1 + 2^8)S_0 \bmod (2^{31} - 1)$;
    2. *if* $S_{16} = 0,$ *then set* $S_{16} = 2^{31} - 1$;
    3. $(S_1, S_2, ..., S_{15}, S_{16})-> (S_0, S_1, ..., S_{14}, S_{15})$.

  }

2.3   The Bit-Reorganization

The middle layer of ZUC algorithm is the bit-reorganization (BR) procedure. It extracts 128 bits from the cell of the LFSR and forms 4 cells (X_0 ,X_1 , X_2 ,

X_3)each of which has 32-bit words. The first three words will be used by the nonlinear function F in the bottom layer, and last word will be used to generate the keystream.

The bit-reorganization forms 4 cells from the above cells as follows:

Bitreorganization()

{

1. $X_0 = S_{15H}||S_{14L}$;
2. $X_1 = S_{11L}||S_{9H}$;
3. $X_2 = S_{7L}||S_{5H}$;
4. $X_3 = S_{2L}||S_{0H}$;

}

3   The Proposed Cryptosystem

This section presents the proposed scheme for color image encryption in the framework of logistic chaotic map which is used as confusion and ZUC stream cipher which is used as diffusion. First, digital color image (P) of size $M \times N$ is converted into RGB components. Afterwards, logistic chaotic map is used to generate a chaotic shuffling sequence. And the position of pixels is shuffled by the logistic sequence. Each colors matrix (R, G, B) is converted into a vector of integers within. MD5 is used Each vector has a length of $L = M \times N$. Then, the stream cipher ZUC is used to encrypt the plaintext image.

3.1   Permutation Based on Logistic Chaotic Map

In this step, logistic chaotic map is used to shuffle the positions of the positions of the image pixels. As we known logistic chaotic map[2] is defined as follow:

$x_{n+1} = \lambda x_n (1 - x_n)$,

Where $\lambda \in (0, 4), n = 0, 1, ...$. The parameter λ and initial value x_0 may represent the key. The parameter λ can be divided into three segments.

The result shows that when $\lambda \in [3.5699465, 4)$, the characteristics of the logistic chaotic map is used to encrypt the image.

In this subsection, a position generator based on logistic chaotic map is used to shuffle the plain image pixels position. The detailed permutation process is stated as follows:

Step1: To facilitate the operation of positions path generation of an image of size $L = M \times N$, where M and N represent the width and the height of the image. The pixel of the image can be labeled by lable $(((i-1) \times M) + j)$, where $(i = 1..M$) and $(j = 1..N$).

Step2:To generate a logistic shuffling sequence, the chaotic map system parameter λ and initial value x_0 are given. A chaos sequence $[x_1, x_2, ..., x_L]$ will be obtained by iterating logistic equation.

Step3:Sorting this set from smallest to largest, a new set $[\overline{x}_1, \overline{x}_2, ..., \overline{x}_L]$ will be get. To obtain the position path, an array of length L is created and each element of this array takes its index array.

Step4:The original position of $\overline{x}_i$ in $[x_1, x_2, ..., x_L]$ can be found, the shuffling sequence is $MD = [m_1, m_2, ..., m_L]$ can be generated. When the shuffling sequence is generated, the pixel position of digital color image can be reordered according to $MD = [m_1, m_2, ..., m_L]$.

Here, we randomly set λ =3.66 and initial value x_0 =0.9. Fig. 2(a) is the original image of Lena, and Fig. 2(b) is encrypted image of Lena shuffled by logistic chaotic map. Fig. 3(a) is the decrypted image of Lena and the Fig. 3(b) is the decrypted image of Lena with fault parameter.
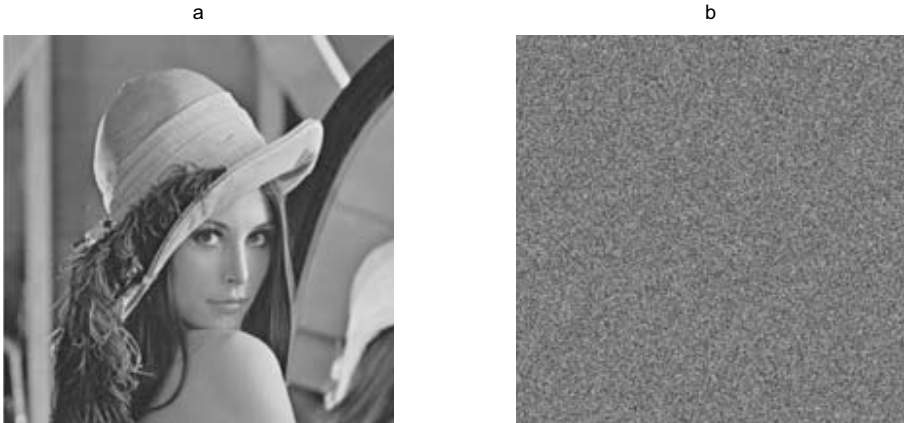
a   b

Fig. 2. (a) Original image of Lena (b) Encrypted Image of Lena shuffled by logistic chaotic map

3.2 Diffusion Based MD5 and ZUC Algorithm

Permutation based on logistic chaotic map is to shuffle the pixel of image which is shown in Fig. 5. The histograms of shuffled image is same to the original image. To overcome those limitations, this paper intends to propose MD5 hash and ZUC stream cipher to diffuse the digital color image.

The Message-Digest algorithm 5 (MD5) is wild used cryptographic function with a 128-bit hash value. This algorithm compresses packet message input with any length into a fixed 128-bit value. If one pixel of image changed, the result of output will also make great adjustment. Thus, it is difficult to reverse the original image depending on summary value. The detailed permutation process is stated as follows:

Step1: Use image which permuted based on logistic chaotic map described above go generate a 128-bit MD5 hash value $MD(i), (i = 1, 2, ..., 128)$ as the 128-bit initial vector iv;

Fig. 3. (a)Decrypted image of Lena (b) Decrypted image of Lena with fault perameter

Step2: During the initialization of ZUC algorithm, the $MD(i)$, $(i = 1, 2, ..., 128)$ and 128-bit initial key k are called by the algorithm. After the initialization stage, the algorithm moves into the working stage. Then the algorithm goes into the stage of producing keystream. For each iteration, a 32-bit word Z_i is produced as an output.

Step3: Reshape the sequence $Z = \{Z_1, Z_2, ..., Z_{M \times N}\}$ to two-dimensional value matrix as shown in the formula 1. Assume a consecutive sequence of plain image pixels P_{ij} as shown in the formula 2 where i $(i = 1..M)$ and j $(j = 1..N)$ donate the location of digital color image. Then the diffused image P' is generated as shown in the formula 3 where $\oplus$ represents the exclusive XOR operation bit by bit.

$$Z = \begin{bmatrix} Z_1 & Z_2 ... & Z_M \\ ... & ... ... & ... \\ ... & ... ... & ... \\ Z_{M \times (N-1)+1} & ... ... & Z_{M \times N} \end{bmatrix} \tag{1}$$

$$P = \begin{bmatrix} P_{11} & P_{12} & ... & P_{1N} \\ ... & ... & ... & ... \\ ... & ... & ... & ... \\ P_{M1} & ... & ... & P_{MN} \end{bmatrix} \tag{2}$$

$$P' = Z \oplus P = \begin{bmatrix} Z_1 & Z_2 ... & Z_M \\ ... & ... ... & ... \\ ... & ... ... & ... \\ Z_{M \times (N-1)+1} & ... ... & Z_{M \times N} \end{bmatrix} \oplus \begin{bmatrix} P_{11} & P_{12} & ... & P_{N1} \\ ... & ... & ... & ... \\ ... & ... & ... & ... \\ P_{M1} & ... & ... & P_{MN} \end{bmatrix} \tag{3}$$

Fig. 4. (a) is original image, and (b) is the encryption image

3.3   The Decryption Process

In the decryption procedure is similar in reverse way to that of encryption process which is mentioned above. Note that the receiver must have the same keystream to be able to decrypt the color image.

4   Performance and Security Analysis

A good encryption procedure should be robust against all kinds of cryptanalytic, statistical and brute-force attack. In this section, some experiments has been conducted to evaluate the performance of proposed encryption algorithm. These experiments include encryption and decryption process, histogram analysis of plain-image and cipher-image. In this section, a 512 x 512 digital color LenaRGB.bmp is used. Experiments are carried out and the data are analyzed using MATLAB.

4.1   Key Space Analysis

The initial key of ZUC and in addition to the seed of logistic chaos map are considered the key of this proposed cipher. As we known that a good encryption algorithm should have large key space to prevent brute-force attacks which is defined to exhaust all the possible key until the correct one. Then output of encryption system should be sensitive to the initial cipher keys. In the proposed algorithm, key space analysis and testing have been carefully performed and completely carried out.

The 128-bit initial key k and the 128-bit initial vector iv which is generated by the MD5 hash function of the ZUC stream cipher is needed. So this algorithm is a 256-bit encryption scheme, with the key space size $2^{256} \approx 1.1579 \times 10^{77}$.

Moreover, when a key is used to encrypt an image, another modified key is used to decrypt the ciphered image, the decryption is also completely fails.

4.2   Histogram Analysis

Image histogram depicts statistical distribution of color intensities. Encrypted color image can be characterized by flat histograms for all colors in which the intensities are distributed evenly over the whole color scale. Fig. 5 shows the histograms of RGB colors for the original image of Lena and the encrypted image respectively. The figures show clearly the uniformity and random-like appearance of the histogram in the ciphered image.
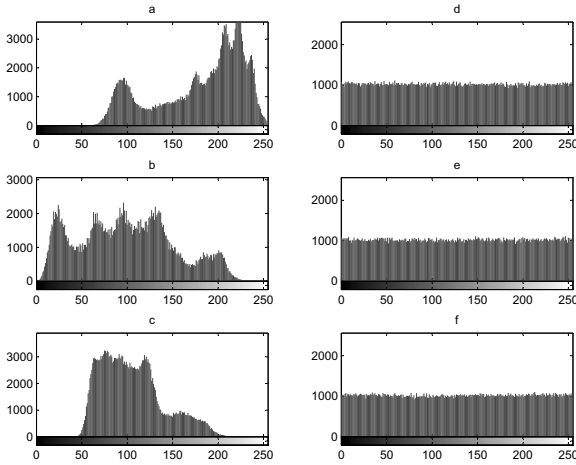

Fig. 5. Histogram of image, a,b,c is histograms of RGB components for the original image. d,e,f is histograms of RGB components for the ciphered the image.

4.3   Pixel Correlation Analysis

It is well known that adjacent image pixel are highly correlated either in horizontal, vertical and diagonal directions. An effective encryption algorithm should make the correlation between adjacent pixels in the images as minimally as possible.

The correlation coefficient between two adjacent pixels x_i and y_i can be calculated[11] between two vertically adjacent pixels, two horizontally adjacent pixels and two diagonally adjacent pixels respectively. Where x and denote two adjacent pixels and N is the total number of the pixels from the image for the calculation.

Table 1 summarizes the auto correlation coefficients for horizontal, vertical and diagonal orientations of the original and ciphered image.

Table 1. Correlation coefficients of two adjacent pixels

	original image			ciphered image		
	red	green	blue	red	green	blue
Horizontal	0.9798	0.9691	0.9327	-0.0022	0.0006	0.0005
Vertical	0.9893	0.9825	0.9576	0.0010	0.0018	0.0030
Diagonal	0.9697	0.9555	0.9183	0.0031	0.0017	0.0022

4.4 Differential Analysis

In general the opponent may make a slight change even one pixel of the encrypted image to observe the change of the result. In this way, the relationship between the plain-image and the cipher-image can be found. If one minor change in the plain-image can cause a significant change in the cipher-image, then differential attack would become very inefficient.

Such difference can be measured by means of two criteria namely, the number of pixel change rate (NPCR) and the unified average changing intensity(UACI). (reference)

Table 2 depicts the mean values of the NPCR and UACI tests for the image of Lena.

$$D(i,j) = \begin{cases} 0 \, , & C_1(i,j) = C_2(i,j) \\ 1 \, , & C_1(i,j) \neq C_2(i,j) \end{cases}$$

$$NPCR = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{i=1}^{N} D(i,j)$$

$$UACI = \frac{1}{M \times N} \sum_{i=1}^{M} \sum_{i=1}^{N} (\frac{|C_1(i,j) - C_2(i,j)|}{255}) \times 100$$

Table 2. Differential analysis result for the image of Lena

	Expected value(%)	proposed scheme(%)
NPCR	99.60937	99.611
UACI	33.46354	33.515

5 Conclusion

This paper has proposed a novel image encryption algorithm which uses a logistic chaotic map to confuse the image. The MD5 hash function and ZUC stream cipher to diffuse the image. This encryption system has enhance the cryptosystem security. And security analyses such as key space analysis, histogram analysis, pixel

correlation analysis and differential analysis have been conducted for several image to prove the security of the image encryption system. From the result, this technique outperforms other encryption techniques and can be used for real-time image encryption, real-time video encryption and other transmission applications.

References

1. Ahmed, H.E.H., Kalash, H.M., Allah, O.S.F.: Encryption quality analysis of the RC5 block cipher algorithm for digital images. Opt. Eng. 45, 107003-107003-7 (2006)
2. Barakat, M.L., Mansingka, A.S., Radwan, A.G., et al.: Hardware stream cipher with controllable chaos generator for colour image encryption. IET Image Proc. 8, 33–43 (2014)
3. Behnia, S., Akhshani, A., Mahmodi, H., et al.: A novel algorithm for image encryption based on mixture of chaotic maps. Chaos, Soliton. Fract. 35, 408–419 (2008)
4. Gao, T., Chen, Z.: Image encryption based on a new total shuffling algorithm. Chaos, Soliton. Fract. 38, 213–220 (2008)
5. Kwok, H.S., Tang, W.K.S.: A fast image encryption system based on chaotic maps with finite precision representation. Chaos, Soliton. Fract. 32, 1518–1529 (2007)
6. Matthews, R.: On the derivation of a chaotic encryption algorithm. Cryptologia 13, 29–42 (1989)
7. Mazloom, S., Eftekhari-Moghadam, A.M.: Color image encryption based on Coupled Nonlinear Chaotic Map. Chaos, Soliton. Fract. 42, 1745–1754 (2009)
8. Pareek, N.K., Patidar, V., Sud, K.K.: Image encryption using chaotic logistic map. Image and Vision Comput. 24, 926–934 (2006)
9. Kitsos, P., Sklavos, N., Provelengios, G., et al.: FPGA-based performance analysis of stream ciphers ZUC, Snow3g, Grain V1, Mickey V2, Trivium and E0. Microprocess. Microsy. 37, 235–245 (2013)
10. Zhu, H., Zhao, C., Zhang, X.: A novel image encryption compression scheme using hyper-chaos and Chinese remainder theorem. Signal Processing-Image 28, 670–680 (2013)
11. Liu, H., Wang, X.: Color image encryption based on one-time keys and robust chaotic maps. Compu. Math. Appl. 59, 320–332 (2010)

The Complexity Analysis of Chaotic Systems[⋆]

Wei Xu[1], Bingbing Song[2], Chunlei Fan[2], Qun Ding[2], and Shu-Chuan Chu[2]

[1] School of Computer Science and Technology, Heilongjiang University,
Harbin 150080, China
[2] School of Electronic Engineering, Heilongjiang University,
Harbin 150080, China
[3] School of Computer Science, Engineering and Mathematics, Flinders University,
Australia
4788216@qq.com, {songbingbing1988,Chunlei_Fan}@163.com,
qunding@aliyun.com

Abstract. The complexity of the sequence is an important index of quantify the performance of chaotic sequence. In order to select a higher complexity of chaotic sequence and apply it in hardware encryption system, this paper analyzes chaotic complexity quantitative analysis methods and presents the approximate entropy and permutation entropy as criterion of measuring the complexity of the chaotic sequences. Set tent, logistic and henon three kinds of chaotic systems as examples, and we analysis and comparison their complexity. It is proved that the two kinds algorithms are effective, and can distinguish different complex chaos and chaotic sequences. Researches show that the complexity of the Logistic map is greater than that of other chaotic systems. The results of the study provide the theoretical and experimental basis for the application of chaotic sequence in hardware encryption system and the information security communication.

Keywords: chaos, complexity, approximate entropy, permutation entropy.

1 Introduction

Chaos, as a classical complex phenomenon of nonlinear dynamic system, has attracted widespread attention for its broadband, noise-like, and sensitive features for initial state. In recent years, with more research on chaos, chaos has replaced the traditional pseudo random sequence in the high density of the commercial and most spread spectrum communication system [1,2].

The complexity of the sequence is not only a similarity degree of measurement between chaotic pseudo-random sequence and random sequence, but also a complexity degree of measurement by using part of the sequence to recovery

the whole. The bigger complexity of the sequence is, the smaller the possibility of recovery is. Therefore, the complexity of the sequence is an important index of quantify the performance of chaotic sequence. The researches of complexity have been attentioned by domestic and foreign scholars. Kolmogrov (1958) [3] defined a measure entropy and used it to measure the disordered degree of system movement. And then Lempel [4] et al. realized the measure entropy method by computer. Pincus (1991) [5] proposed the definition of approximate entropy through measuring the complexity of time series, and then Bandt [6] et al. proposed permutation entropy for measuring time series. Xiao Fang-hong [7] et al. proposed to apply a symbolic dynamics approach for the complexity analysis of chaotic pseudo-random sequences in 2004. Next year Larrondo [8] et al. proposed a intensive statistical complexity measure to quantify the performance of chaotic pseudorandom number generators. Chen Xiao-jun (2011) [9] et al. proposed a new complexity metric to evaluate the unpredictability of the chaotic pseudorandom sequences based on the Fuzzy Entropy.

Kolmogorov-sinai entropy proposed by Kolmogrov can measure the complexity of chaotic system, but it needs a lot of sample space and heavy computation. The symbolic dynamics approach can reduce the degree of dependence on the parameters, but before we measure the complexity, we must get the size of symbol space of the initial sequences, which is very difficult for us to obtain the priori knowledge in practice.

The article assesses randomness of Logistic, Tent and Henon mappings via approximate entropy and permutation entropy to find the better chaotic mapping, and then provides powerful basis for realization the chaotic encryption system by the hardware and the application of chaotic systems in cryptography and securet communication.

2   The Qualitative Characteristics of Chaos

Chaos, as one of the nonlinear dynamic systems has the geometry and statistical features that deterministic movement usually do not have, such as local instability while overall stability, strange attractor, continuous power spectrum, positive Lyapunov index, fractal dimension, positive measure entropy and so on. To sum up, the chaos has the following three main qualitative characteristics [10]:

1) Inherent randomness: From deterministic nonlinear system evolution process, they show random uncertainty behaviors in the chaotic sector. However, this kind of uncertainty is not from external environment random factors on the influence of the system movement, but from system of spontaneous. Another meaning of inherent randomness is local instability. Most of the chaotic system has the inherent instability and overall stability. The difference between chaotic and orderly state is stable overall while instable locality. The so-called local instability refers to the behavior for certain aspects of the system movement, which strongly depends on the initial conditions of the system.

2) Fractal dimension characteristics: Chaos has fractal dimension properties, its fractal dimension is not used to describe system geometrical shape, but to

describe the behavior characteristics of the system movement track in the phase space. Chaotic movement in the phase space within a given area of infinite time through constitutes infinite levels of self-similar structure - strange attractor.

3) Universality: Chaos is a kind of no cycle "advanced" orderly movement. In the research of chaos in the process, we can find some kind of scale invariance instead of the usual space and time periodic; the so-called universality refers to the system shown by the common characteristics in the trend of chaotic state, it doesn't depend on specific coefficient and the motion equation of system.

3   Analysis Method of Chaotic Complexity

There is no strict mathematical definition of the chaotic complexity so far, therefore, when we research the chaotic complexity, the main type of complexity parameters we use are entropies, fractal dimensions, Lyapunov exponents, and so on[11].

3.1   Lyapunov Exponents

Lyapunov exponent is an important quantitative index for measuring the characteristics of system dynamics, which characterizes the divergence or convergence average exponent rate between adjacent trajectories in phase space. The adjacent trajectory is convergence or divergent by rate less than exponent in the linear dynamic system, while divergent by rate of exponent in chaotic system. Therefore, we often use positive Lyapunov exponent as a criterion to estimate whether the system is chaotic. A continuous system given by

$$\frac{dx}{dy} = f(x(t)), x(t) \in \mathbf{R^n} \tag{1}$$

Initial condition of a trajectory and its neighboring in n-dimensional space respectively are x_0 and $x_0 + \Delta x$, which tangent vector is $W(x_0, t)$, then Lyapunov exponent can be defined as average exponent divergence rate between two trajectory

$$\lambda(x_0, W) = \lim_{t \to \infty} \frac{1}{t} ln \frac{\|W(x_0, t)\|}{\|W(x_0, 0)\|} \tag{2}$$

A discrete system given by

$$x_{n+1} = f(x_n), x_n \in \mathbf{R^n} \tag{3}$$

Define $A_n = [M(x_n)M(x_{n-1}) \cdots M(x_1)]^{\frac{1}{n}}$, where $M(x_n)$ is Jacobi matrix of $f(\bullet)$, $\delta_i(n)$ is eigenvalue of A_n , then Lyapunov exponent can be defined as

$$\lambda_i = \lim_{n \to \infty} ln| \delta_i(n) | \tag{4}$$

The former definition is mainly described the geometric meaning of Lyapunov exponent, while latter describe the calculation of Lyapunov exponent.

3.2   Fractal Dimension

The fractal dimension is a mathematical concept, which measures the geometrical complexity of an object.

The structure of chaotic attractor is self-similar with infinite levels, which disperse degree can be characterization by dimension. The trajectory in the phase space has characteristic of fractal dimension for its own fractal self-similar structure of chaos. The fractal dimension is effective measures to estimate whether the movement is chaotic.

Define S is a subset of n-dimension space, $M(\varepsilon)$ is the number of n-dimension cube under the scale of which used to cover S , if the limits exist, then the fractal dimension of S under the scale of ε can be expressed as

$$D(S) = \lim_{\varepsilon \to 0} \frac{ln M(\varepsilon)}{ln(\frac{1}{\varepsilon})} \tag{5}$$

The movement is chaotic if its dimension of attractor is fraction. For the calculation of Hausdorff fractal dimension is very difficult, Grassberg and Procaccia propose the GP algorithm to calculate another kind of fractal dimension, i.e. correlation dimension $D_o(S)$ in 1983.

$$D_o(S) = \lim_{\varepsilon \to \infty} \frac{ln C(\varepsilon)}{ln(\varepsilon)} \tag{6}$$

$$C(\varepsilon) = \lim_{n \to \infty} \frac{1}{n^2} \sum_{i,j=1}^{n} H(\varepsilon - \mid X(i) - X(j) \mid) \tag{7}$$

Where $H(\bullet)$ is step function of Heaviside.

The limitations of this method cant measure the complexity of chaotic system according to the value of fractal dimension, although according to the dimension it can determine whether the system is chaotic.

3.3   Metric Entropy

Metric entropy is also known as kolmogorov-sinai entropy, which describes the confusion degree of movement for dynamic system. The entropy can be used in recognition the characteristics of the chaotic and the confusion degree of chaos in chaotic system. The character of chaotic attractor is sensitive to initial value, which can be reacted from the concept of entropy. The adjacent trajectory is divergent by rate of exponent in chaotic system for its local instability, which causes the information of initial conditions fade away in the process of movement. In another hand, if two initial point very close so that can't distinguish from measure, but as the time evolutes, we can distinguish them by their distance increasing by rate of exponent. In this sense, we think that chaotic movement generates information. Put all the time information generation rate for the exponent averagewe get the kolmogorov-sinai entropy. Defining all the attractor is a set, which completes covered by a n-dimension cube with side length for ε .

Defining the number of cube is $M(\varepsilon)$, p_i as the probability through the cube, then metric entropy can be defined as

$$I(\varepsilon) = -\sum_{i=1}^{M(\varepsilon)} p_i ln P_i \tag{8}$$

Based on the principle mentioned above, it shows that the amount of information is related to the number of different trajectory which can distinguish. The number of trajectory N increased by rate of exponent in chaotic movement: $N \propto e^{kt}$,where k is Metric entropy. In a period of time, the number of different trajectory which can distinguish is more, then it has more complexity. In other words, the bigger k is, the stronger the randomness is and the smaller the possibility of recovery is.

Metric entropy can measure the complexity of chaotic system, but it needs a lot of sample space and heavy computation. It is hard for us to get the numerical value of metric entropy in practice.

4   The ApEn Algorithm

Approximate Entropy was proposed by Pincus in 1990 [12]. The algorithm developed from Limpel-Ziv algorithm. Approximate entropy algorithm count the random degree of sequence by the edge of the conditional probability statistical method and show the sequences complexity through the degree of adjacent track change.

Definition: Given a positive integer N and nonnegative integer m ,with $m \leq N$,a positive real number r ,and a time-series of data $x := x(1), x(2), ..., x(N)$,from measurements equally spaced in time, form a sequence of vectors $X(1), X(2),$ $..., X(N - m + 1)$ in R^m, define by $X(i) = [x(1), x(2), ..., x(N - m + 1)]$ next, define for each i, $1 \leq i \leq (N - m + 1)$, let the distance between two blocks $X(i)$ and $X(j)$ be defined by

$$d[X(i), X(j)] = \max_{k=1,2,...,m} | x(i + k + 1) - x(j + k + 1) | \tag{9}$$

Then let $C_i{}^m(r) =$(number of $j \leq (N - m + 1)$ such that $\frac{d(x(i),x(j)) \leq r)}{N-m+1}$ Now define

$$\Phi^m(r) = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} log C_i{}^m(r) \tag{10}$$

and

$$ApEn(m, r, N) = \Phi^m(r) - \Phi^{(m+1)}(r), m \geq 1 \tag{11}$$

ApEn measures the logarithmic frequencies with which blocks of length m that are close together for blocks augmented by one position. Thus, small values of ApEn imply strong regularity, or persistence, in a sequence. Alternatively, large values of ApEn imply substantial fluctuation, or irregularity.

The parameter m is the dimension of distance vector. The accuracy of complexity in the chaotic system is determined by parameter r which is a value of the distance. The value of r should be in the scope $0.1 \sim 0.25SD_x$.

4.1   The Logistic Map

The Logistic map is given by:

$$x_{n+1} = \mu x_n (1 - x_n), x \in (0, 1), \mu \in (0, 4] \tag{12}$$

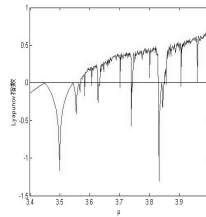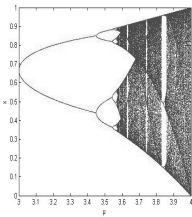Fig. 1 displays the bifurcation diagram of Logistic map. Fig. 2 displays the Lyapunov exponent trend of Logistic map with change of parameter μ.

Fig. 1. Orbit diagram for the Logistic map

Fig. 2. Lyapunov exponent for the Logistic map

4.2   The Tent Map

Tent mapping is defined as follows: The Logistic map is given by:

$$x_{n+1} = 1 - \mid 1 - \mu x_n \mid, x \in (0, 1) \tag{13}$$

The bifurcation diagram of Tent map on the interval $[1, 2.4]$ in Matlab is shown in Fig. 3.
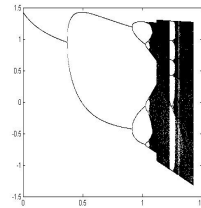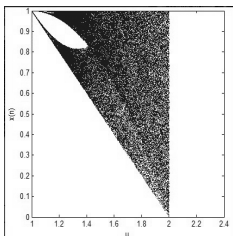

Fig. 3. Bifurcation diagram of Tent map   **Fig. 4.** Bifurcation diagram of Henon map

4.3   The Henon Map

The bifurcation diagram of Henon map in Matlab is shown in Fig. 4.

$$\begin{cases} x_{n+1} = 1 - ax_n^2 + y_n \\ y_{n+1} = bx_n \end{cases} \tag{14}$$

The parameter m is the dimension of Distance vector. We demonstrate the utility of ApEn by applying this statistic to Logistic, Tent and Henon map, observing the dependence of ApEn on m. The simulation results we obtained as shown in Fig. 5.
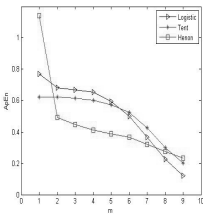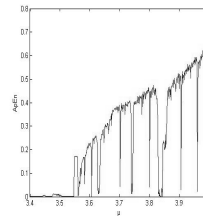

Fig. 5. ApEn of three maps on m

Fig. 6. ApEn of Logistic map on μ

The results in Fig. 5 has shown that the value of ApEn when m=2 is better then the others. According to the references [14], we choose parameters $m = 2$, $r = 0.1SD_x, N \geq 1000$, and tested the ApEn of sequence produced by Logistic, Tent and Henon map. The simulation results we obtained form three maps as shown in Fig. 6, Fig. 7 and Fig. 8 respectively.
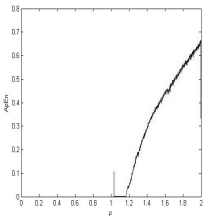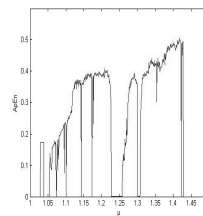

Fig. 7. ApEn of Tent map on μ

Fig. 8. ApEn of Henon map on μ

As shown in three figures, we noticed that the sequences' complexity of Logistic mapping is bigger than the other two mapping.

5   The PE Algorithm

Permutation entropy algorithm based on the complexity of Kolmogorov, by use the concept of information entropy, to statistics and calculation the complexity of the sequences. The algorithm by using multidimensional reconsitution space similarity to measure the complexity of the entire sequences, and analysis all the similar characteristics of the embedded dimension[13].

Definition: Consider a time series $\{x_t\}_{t=1..T}$.We study all $n!$ permutations of order n which are considered here as possible order types of n different numbers. For each we determine the relative frequency

$$p(\pi) = \frac{\#\{t|0 \leq t \leq T-n, (x_{t+1}, ..., x_{t+n}) has \qquad type \qquad \pi\}}{T-n+1} \qquad (15)$$

This estimates the frequency of π as good as possible for a finite series of values. To determine $p(\pi)$ exactly, we have to assume an infinite time series $\{x_1, x_2, \}$ and take the limit for $T \to \infty$ in the above formula. This limit exists with probability 1 when the underlying stochastic process fulfils a very weak stationarity condition: for $k \leq n$, the probability for $x_t < x_{t+k}$ should not depend on t.

The permutation entropy of order $n \geq 2$ is defined as .Where the sum runs over all $n!$ permutations π of order n. This is the information contained in comparing n consecutive values of the time series. It is clear that $0 \leq H(n) \geq log n!$ Where the lower bound is attained for an increasing for decreasing sequence of values, and the upper bound for a completely random system where all $n!$ possible permutations appear with the same probability. The time series presents some sort of dynamics when $H(n) < log n!$.

We demonstrate the utility of PE by applying this statistic to Logistic, Tent and Henon map, observing the dependence of ApEn on m. The simulation results we obtained as shown in Fig. 9.
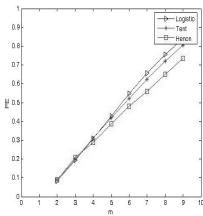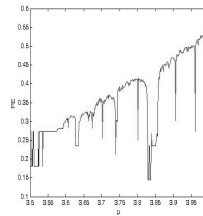

Fig. 9. PE of three maps on m   **Fig. 10.** PE of Logistic map on μ

From the results of Fig. 9, we can know the complexities of the various mapping have a linear relation with m. According to the references [14], we tested the PE of sequences produced by Logistic, Tent and Henon map. The simulation results we obtained form three maps as shown in Fig. 10, Fig. 11 and Fig. 12
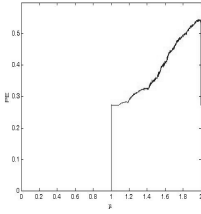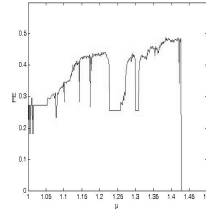
Fig. 11. PE of Tent map on μ

Fig. 12. PE of Tent map on μ

respectively. From the results of three figures, we noticed that the sequences' complexity of Logistic mapping is bigger than the other two mapping.

6   Comparison between ApEn Algorithm and PE Algorithm

Through analyze the computing principle and the physical significance of the two kinds of algorithms, we learn that the two kinds of algorithms has similarities except their own characteristics. The two kinds of algorithms based on the Kolmogorov complexity, and both describe the random degree of sequences. In addition, the two kinds of algorithms obtained the numerical values of sequences' complexity by using the concept of information entropy to count up the length and the probability of the sequences. For the ApEn algorithm, the bigger complexity of the sequences means similar reconstruct sequence is less. To obtain the reasonable calculation results, we must increase the length of the sequences and enlarge resolution parameter r. For the PE algorithm, the more complexity of the sequences is, the more the number of permutations is, the number of times for statistics are more, thus the PE of sequences is greater. The two kinds of algorithms are both based on the Kolmogorov complexity and information entropy.

The two algorithms also have their own characteristics. The ApEn algorithm show the complexity of the sequences based on the different embedded dimension. The calculation results vary according to the subjective factors for choice of embedded dimension and resolution parameter, while the PE algorithm show the complexity of the sequences based on the determinate embedded dimension. The concept of the embedded dimension is same as the one in the ApEn algorithm.

7   Conclusion

The approximate entropy is a method of quantize the complexity of time series based on edge probability distribution statistics. It can accurate calculation the complexity of the sequence, but the result is influenced by select the different parameters. The calculate of ApEn is fast, accurate and easy to be realized, in

addition the algorithm can calculation the sequence complexity through very short length of sample space. Permutation entropy is an appropriate complexity measure for chaotic time series, in particular in the presence of dynamical and observational noise, since the method is extremely fast, it seems preferable when there are huge data sets and no time for preprocessing and fine-tuning of parameters. In this paper, the ApEn algorithm and PE algorithm are used to calculate the complexity of the three different chaotic mappings, the result shows that the complexity of the Logistic map is bigger than the other two mapping, in other words the random degree of Logisitic sequences is greater. The research provides powerful basis for the application of chaotic systems in cryptography and securest communication.

References

1. Huang, L.L., Yin, Q.T.: A Chaos Synchronization Secure Communication system Based on Output Control. Electronics & Information Technology 31, 2402–2405 (2009)
2. Parlitz, U., Ergezinger, S.: Robust Communication Based Chaotic Spreading Sequences. Phys. Lett. A 188, 146–150 (1994)
3. Kolmogrov, A.N.: Three Approaches to the Quantitative Definition of Information. Problem in Information Transmission 1, 1–7 (1965)
4. Lempel, A., Ziv, J.: On the Complexity of Finite Sequences. IEEE Transactions on Information Theory 22, 75–81 (1976)
5. Pincus, S.M.: Approximate Entropy As a Measure of System Complexity. Proc. Natl. Acad. Sci. 88, 2297–2301 (1991)
6. Bandt, C., Pompe, B.: Permulation Entropy a Natural Complexity Measure for Time Series. Phys. Rev. Lett. 88, 174102-1–174102-4 (2002)
7. Xiao, F.H., Yan, G.R., Han, Y.H.: A Symbolic Dynamics Approach for the Complexity Analysis of Chaotic Pseudorandom Sequences. Acta Physica Sinica 53, 2877–2880 (2004)
8. Larrondo, H.A., Gonzalez, C.M., Martin, M.T.: Intensive Statistical Complexity Measure of Pseudorandom Number Generators. Physica A 356, 133–138 (2005)
9. Chen, X.J., Li, Z., Bai, B.M.: A New Complexity Metric of Chaotic Pseudorandom Sequences Based on Fuzzy Entropy. Electronics & Information Technology 147, 195–197 (1981)
10. Zhao, G., Fang, J.Q.: Modern Information Safety and Advances In Application Research of Chaos-Based Security Communication. Progress In Physics 23, 212–252 (2003)
11. Wang, Y.X., Weng, Y.F.: Study on the Complexity Analysis Methodology and Application to Logisitic Map. Beijing Technology and Business University 24, 38–41 (2006)
12. Pincus, S.M.: Approximate Entropy as a Measure of System Complexity. Proceedings of the National Academy of Sciences 88, 2297–2301 (1991)
13. Sun, K.H., Tan, G.Q.: The Complexity Analysis of TD-ERCS Discrete Chaotic Pseudo-Random Sequences. Acta Physica Sinica 57, 3359–3365 (2008)
14. Xu, W., Ding, Q., Zhang, X.G.: Detection Complexity of Chaotic Sequence. Information Technology 12, 5487–5491 (2013)

Implementation of Audio Data Packet Encryption Synchronization Circuit⋆

Hongbin Ma, Yingli Wang, and Gaoling Li

Electronic Engineering, Heilongjiang University,
Harbin, Heilongjia, China
`midaspeking@sina.com`

Abstract. The paper chooses the cyclone the second generation of the FPGA chip as the gate array (FPGA) used by AES encryption algorithm. The paper also uses speech codec WM8731 chip to realize eight voice and data coding. Then the paper refers to the PCM frame structure TS0 times lot function design code word synchronization. At last, the paper also uses AES128 grouping encryption algorithm to encrypt digital signal, and voice encryption is realized on the FPGA hardware.

Keywords: encryption algorithm, FPGA chip, speech coding, encryption synchronization.

1 Introduction

With the development of network applications, private speech information privacy is becoming more and more attention. The paper encrypts the audio data by grouping encryption technology[1,2] based on FPGA. But the routing of IP network transmission and delay characteristics make the received packet disorder, and affect the decryption key synchronization. The paper designs a code word synchronization unit of voice pretreatment and coding unit. It can ensure the encrypted declassified audio data synchronization. The grouping encryption technology is embedded in the FPGA kernel. The safety of voice communications is guaranteed by grouping encryption circuit.

2 Speech Data A/D Conversion

The WM8731 speech codec chip is used in the paper. It can collect voice data. When gathered, the data stream is transmitted bit by bit. The WM8731 chip inside integrates A/D conversion mode[3], and the analog signal data is converted digital signal in the A/D conversion mode. Figure 3-10 is data collection format. BCLK refers to the speech coding clock signal. ADCLRC refers to control signal

---

of A/D conversion. The LEFT CHANNEL refers to the left sound channel, and the RIGHT CHANNEL refers to the right sound channel. Left sound channel and the right sound channel completely control signals for ADCLRC. ADCDAT refers to data signals of A/D conversion, including ADCDAT 32 bits of data. LEFT CHANNWL refers to 16 bits of data and RIGHT CHANNEL 16 is bits of data. But the MSB is high data and LSB refers to low signal.
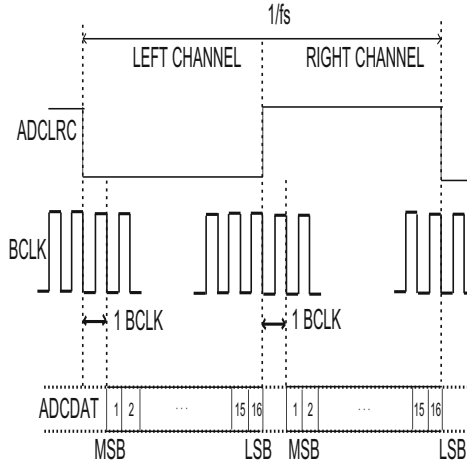

Fig. 1. The format of data collection

Speech signal input from the MIC[4], transfer to the WM8731 speech coding chip, then the data is in A/D conversion. Figure 1 shown is data forma after A/D conversion. The data are commonly controlled by the BCLK clock signal and ADCLRC control signal. BCLK refers to speech coding clock, and ADCLRC refers to control signal. When ADCLRC falling edge and the BCLK clock signal falling edge arrive at the same time and the BCLK next falling edge comes, it is time to begin collecting 16-bit data to the left sound channel (LEFT CHANNEL) speech signal. First collected data is on high bit, the latter collected data is in low bit. When ADCLRC rise along with the falling edge BCLK synchronous arrival and BCLK next falling edge comes, right sound channel began collecting 16-bit data (RIGHT CHANNEL). The clock signal BCLK finishes 32-bit data collection after 34 clock signal in the process. This is a collection signal and it sends 32-bit data. A serial ADCDAT send 32 bits of data. The signals are sent to the FPGA, and the data is encrypted.

3   Code Word Synchronization

How to deal with receiving synchronization must be considered for a communication system. It will directly related to the declassification of speech encryption

system receives the accuracy of the problem. Synchronous signal, the TS0 time slot as synchronization code, achieve voice synchronization signal. Referring to content of PCM frame structure, this paper uses TS0 times lot to tag code of voice and data. The voice synchronization is finished in the phonetic acquisition 32-bit code set aside a bit. In the paper, the voice signal acquisition is bit by bit the form of data flow. For A/D conversion of data format, it is known as serial 32-bit data after data collection. The two data of 32-bit data collected are marked.
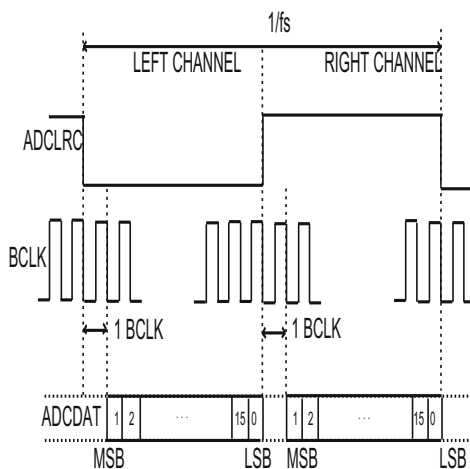

Fig. 2. The flag first bit of collection

As shown in figure 2, the left channel low 1 and the right channel low 1 are namely the 00. The next 32-bit data comes under low two namely 01 and so on. When the fourth 32-bit data arrives, the two low is namely 11. After the four 32-bit data are finished, they are sent encryption in nuclei and to be encrypted. At the receiving end, if the 10 data is missing, the decryption end would automatically use the key , which is need in the decryption process to solution the next data. As we know, the data is decrypted at the decoding end. The decoding order need rely on the labeling bit. In the process of data transmission, network congestion, or no signal will make the data miss. The missing data won't be code in the decoding end. Then it is necessary to use the mark bit to decide. The lost data, joining the 0010 mark bits on flag bit, will send a request. The sender will give back to the data with 10 mark. It can solve the problem of speech signal data loss in the process of transmission, realizing voice synchronicity.

4   Encryption Module Design and Implementation

This paper uses fast look-up table method to realize the AES algorithm [5,6,7] in the Verilog programming. The Case statement structure many things, such as, the S box form, the column hybrid forms in the process of encryption, simple cyclic shift, twice look-up table in encryption process, and mould operation. The look-up table method improves the efficacy of AES algorithm. AES encryption algorithms in the hardware circuit diagram and the design of circuit diagram AES in the FPGA [8] are shown in figure 3.
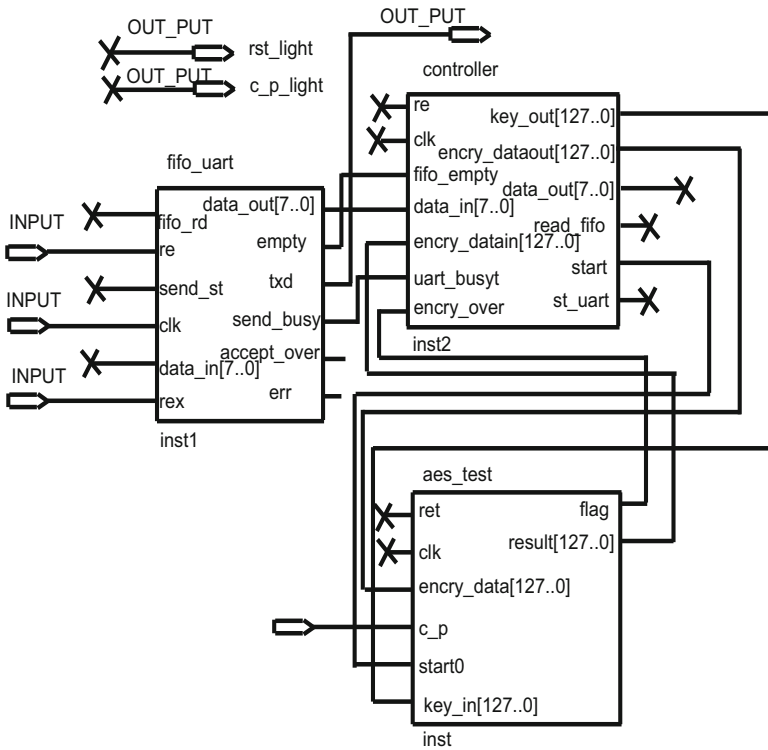
Fig. 3. AES encryption top nuclear

   The function of AES encryption: the digital signal of speech signal encrypts the data through the encryption module. The encrypted data is sent through the wireless and data is decrypted on decryption end. But AES encryption mainly consists with encryption module, key rotation, decryption module, key extension, etc. These modules can encrypt the receiving data with 128 - bit data. This is the function of encryption. The principle of the encryption process is shown in figure 4. Encryption principle is shown in figure 5. The encryption part is divided

Fig. 4. The principle diagram of the encryption part of the main frame
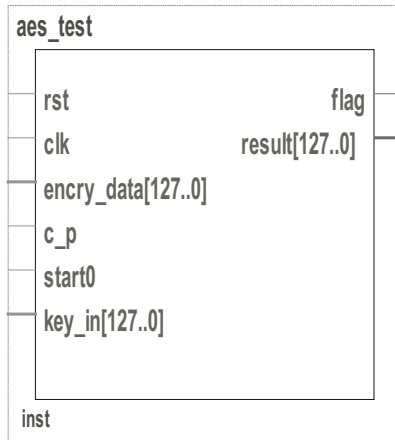
into four modules: the encryption module, decryption module, key extension module and RAM module. The AES encryption module contains 8 I/O ports. Six of the 8 I/O parts is input ports and other two ports is output ports. The encry_data[127..0] port can receive the 128 - bit digital signal of the voice signal. When coding data flow bit by bit is stored in AD_fifo, and it is converted to 32-bit data transmission. Before a 128 - bit AES encryption algorithm in the process of transmission, the data is stored in cache, namely in the fifo of the encryption algorithm. When the data arrives 128 bits of data, start0 port begins to work, the data is stored in SDRAM. After pressing the button for voice broadcast, the data need to be decrypted in the process of play, restoring previous digital signal. Then through the WM8731 decoding, voice encryption process is finished. The encryption part contains 6 input port: rst refers to clock reset signal, and it is effective under the high level. clk port mainly refers to the clock signal in the system. And encry_data [127..0] refers to the 128 data come clear speech signal input port. cp signal primarily refers to selective encryption. When cp is high electricity at ordinary times, the whole system is the encryption process. When cp is low electricity at ordinary times, the whole system is the decryption process. It is like the equivalent of buttons. When the button is not pressed, voice encryption process begins. When the button is pressed, the voice decryption process begins. start0 port refers to the start signal. The audio encryption of he whole system begins, It is effective high electricity at ordinary times. key_in [127..0] port is a key input port. At the beginning of the whole system, firstly inputting key, the system will automatically take the first data as the key data. After the data reaches 128 bits, the latter data is clear data. The system began to encrypt the latter 128 - bit data. Two output port: flag output port is signal that the add/decryption is over. Its effective level is high level effectively. When the flag is high level, data encryption process is finished. And result [127..0] port is the output result of add/decryption of the 128 speech signal.
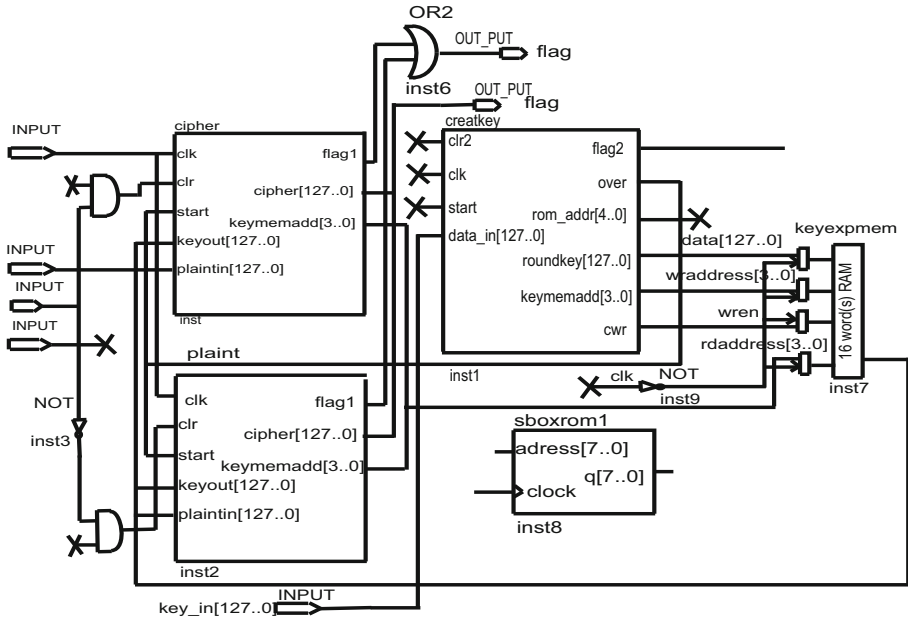
Fig. 5. The part diagram of Encryption principle

5 Software Process and the Environment

The system test environment mainly applies Quartus9.0 software simulation [9] and is encoded in the platform. The speech coding of audio WM8731 and the acquisition of voice data are base on Quartus software. The collected data through the WM8731 chip convert other data, namely the A/D conversion. The converted data through FPGA is under the encryption [10] of data. Then it is transmitted to the wireless network and received on the other side. Software process: design input, design, build, function simulation and timing simulation, hardware configuration and verification. The flow chart of the software overall design is shown as figure 6.

6 The Result and Stochastic Analysis

Set the initial key as hexadecimal 000102030405060708090A0B0C0D0E0F0 results of ten rounds of key are as follows:

Round1:D6AA74FDD2AF72FADAA678F1D6B76FE;
Round2:B692CF0B643DBDF1BE9BC5006830B3FE;
Round3:B6FF744ED2C2C9BF6C590CBF0469BF41;
Round4:47F7F7BC95353E03F96C32BCFD058DFD;
Round5:3CAAA3E8A99F9DEB50F3AF57ADF622AA;
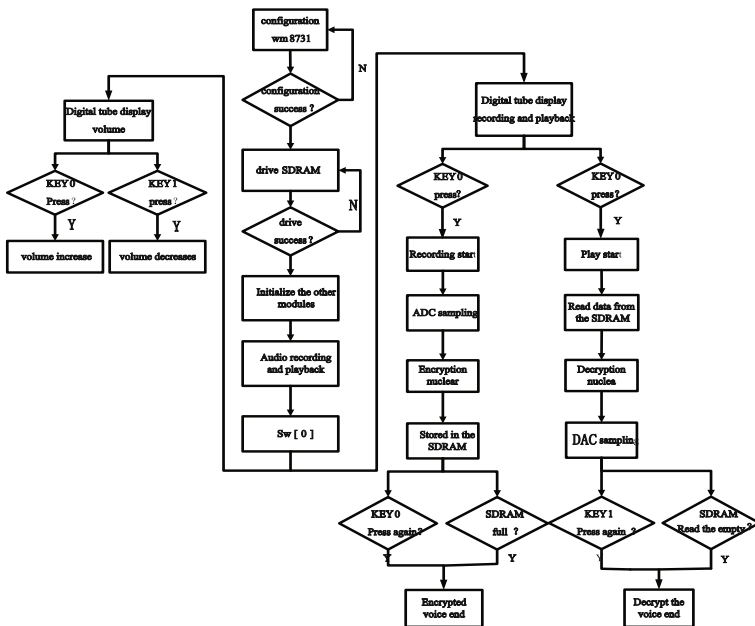Round6:5E390F7DF7A69296A7553DC10AA341F6B;

Fig. 6. The flow chart of software

Round7:14F9701AE35FE28C440ADF4D4EA9C1026;
Round8:47438735A41C65B9E016BA1F4AEBF7AD2;
Round9:549932D1F085577681093ED9CBE2C974E;
Round10:13111D7FE3944A17F307A78B4D2B30C5.

From ten key rotation wheel key extension, it will be found that the key of each round is different. The degree of data encryption is greatly increased in the encryption process. The round keys write number of round keys itself in the program. The key is 10 rounds in the paper. After key rotation, it is time to encrypt data. For captured data, because the intercept is not clear to key rotation, they can not to decrypt the data.

By Quartus software simulation data captures part of the AD conversion after the data. The data is passed to the encryption, the data for the cipher is text data. Then the cipher text data is put into a binary code, namely the 0 1 sequence. Through matlab simulates 0 1 sequence randomness, the number of run, inverse cumulative value of the standard normal distribution and the test results can be known.

011010000001001010101011000101001101011110010011110111101111010010
10 111001010011101000001011010111011100100001011010000100011011

The above data is a voice encryption data and can be taken as the input sequence. It is test by test principle of the runs. Setting parameters: r refers to the number of runs. Randomunm refers to the input sequence. zr refers to test results.

The inverse cumulative function value of the standard normal distribution is taken as a measure of random sequence by using Norminv.

The test sets up several special clear text input in the process.

Such as, clear data of full 1:

1111111 1111111111111111111111111,

clear data of the whole 0:

00000000000000000000000000 000000,

01 evenly distributed data:

0101010101010101010101010101,

F all the data:

FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,

The normal voice data:

11111010110010001111101011001000.

The above five clear data is encrypted, then the cipher text data.

full 1 cipher text data:

35D14E6D3E3A279CF01E343E34E7DED3,

All the zero cipher :

C6A13B37878F5B826F4F8162A1C8D879,

01 distributed evenly cipher text data:

C352805754237F311AC0FFF4E3E03E78,

all F the cipher text data:

3C441F32CE078 22364D7A2990E50BB13,

Voice cipher data:

6812AB14D793DEF4AE53A0 B5DC85A11B.

The cipher text is converted to binary. Then 0 1 series is under the discussion through the principle of runs test. It is a random sequence or not. The Inspection is that the value of the inverse cumulative of standard normal distribution is 1.96, proving that the encrypted cipher text sequence is random sequence.

References

1. Lian, S.G.: On the Secure Multimedia Distribution Scheme Based on Partial Encryption. In: International Conference on Taylor and Francis Group, vol. 9, pp. 23–27 (2010)
2. Martin, H., Oliver, P.: ADPCM with Adaptive Pre-and Post Filtering for-Delay-free AudioCoding. In: ICASSP 2007, vol. 3, pp. 269–272 (2007)
3. Chang, D.Y.: Design Techniques for a Low-Power Low-Cost CMOS A/D Converter. Journal of Solid State Circurts 33, 1244–1248 (1998)
4. Salamci, M.U., Ozgoren, M.K., Banks, S.P.: Sliding Mode Control with Optimal Slidingsurfaces for Missile Autopilot Design. Journal of Guidance Control and Dynamics 23, 719–727 (2000)
5. Nalini, C., Nagaraj: An FPGA Based Performance Analysis of Pipelining and Unrolling of AES Algorithm. In: International Conference on Source, vol. 24, pp. 477–482 (2006)

6. Lang, R.L., Xia, Y., Dai, G.Z.: Study of Advanced Encryption Atandard (AES) Algorithm. Small Microcomputer System 24, 905–908 (2003)
7. Orr, D., Nathan, K.: The effects of the omission of last round's Mixcolumns on AES. Information Processing Letters 11, 304–308 (2010)
8. Li, H.W., Yuan, S.H.: Based on Quartus FPGA/CPLD Design. Publishing House of Electronics Industry, Beijing (2006)
9. Altera Cyclone II Device Handbook, `http://www.altera.com`
10. Jiang, B.: Reduction of Primavera. Digital System Design and PLD Application Technology. Publishing House of Electronics Industry, Beijing (2010)

Discriminative Image Representation for Classification

Zhize Wu, Shouhong Wan, Lihua Yue, and Ruoxin Sang

School of Computer Science and Technology,
Key Laboratory of Electromagnetic Space Information,Chinese Academy of Sciences,
University of Science and Technology of China, Hefei, Anhui, China
{wuzhize,srx2007}@mail.ustc.edu.cn,{wansh,llyue}@ustc.edu.cn

Abstract. The Bag-of-visual Words (BoW) image representation is a classical method applied for various problems in the fields of multimedia and computer vision. During the process of BoW image representation, one of the core problems is to generate discriminative and descriptive visual words. In this paper, in order to represent the image completely, we propose a visual word filtering algorithm, which filters the lower discriminative and descriptive visual words. Based on the traditional method of generating visual words, the filtering algorithm includes two steps: 1) calculate the probability distribution of the various visual words, and then, delete the words with gentle probability distribution; 2) delete the visual words with less instances. In this way, the generated visual features become more discriminative and descriptive, furthermore, multiple cues fusion, such as shape, color, texture, is also taken into account, we compare our approach with traditional Bag-of-visual Words method applied for image classification on three benchmark datasets, and the performances of the classification all get improvements to some extent.

Keywords: Image Representation, Discriminative, Descriptive, Multiple cues fusion, Visual word filtering algorithm.

1 Introduction

Image category recognition is important to access visual information on the level of objects (buildings, cars, *etc.*) and scene types (outdoor, vegetation, *etc.*). In general, systems for category recognition on images [5], [9], [2] and video [15] use machine learning based on image descriptions to distinguish object and scene categories. However, there can be large variations in viewing and lighting conditions for real-world scenes, complicating the description of images and consequently the image category recognition task. Inspired by the success of textual words in large-scale textual information processing, researchers are trying to extract visual words from images which function similar as textual words. During the past few years, Bag-of-visual Words approaches have allowed significant advances in image classification [3].

Traditionally, visual words are created by clustering a large number of local features such as SIFT [10] in un-supervised ways. After that, each cluster center

is taken as a visual words, and a corresponding visual vocabulary is generated. With the visual vocabulary, image can be transformed as Bag-of-visual Words representation. This is simply achieved by extracting image local features and replacing them with their nearest visual words [8]. However, experimental results of reported works show that the commonly generated visual words [8], [11],[14], [18], are still not as expressive as text words. Mainly in the presence of noise in the generation of visual words, making visual word in different types of images appear mismatch. A toy example illustrating this finding is presented in Fig.1. In the figure, SIFT descriptors are extracted on interest points detected by Harris-Laplace. The three images are then represented as BoWs with a visual vocabulary containing 1000 visual words, by replacing their SIFT descriptors with the indexes of the closest visual words. In the figure, two interest points are connected with a red line if they share the same visual word. As we can clearly observe, although the visual appearances of the plane and ant are very different, there are still many matched visual words between them. In [8], Tian *et al.* analyze that the ineffectiveness of the traditional visual vocabulary might be largely due to its three innate shortcomings, *i.e.*, 1) the extracted local features are not stable, which means some local features are sensitive to affine transformations. 2) The classic visual words are generated from single local features, thus they cannot preserve enough spatial information. 3) The clustering process is unsupervised, thus, the generated visual words might be not semantically reasonable, and many noisy visual words can be generated. The reasons causing the low descriptive power of visual words are still being studied by researchers in related fields, and the ineffectiveness is not limited to these above reasons.

Initially, many methods only used the shape features, predominantly represented by SIFT [10] to represent an image [9],[6]. However, more recently the possibility of adding color information has been investigated. Bosch *et al.* [1] propose to compute the SIFT descriptor in the HSV color space and concatenate the results into one combined color-shape descriptor. Van de Sande *et al.* [16] performed a study into the photometric properties of many color descriptors, and did an extensive performance evaluation. The comparison performed in these studies [1],[16], [7] suggests that combining multiple cues usually improves final classification results.

In order to improve the discriminative and descriptive ability of visual words, in this paper, we propose a visual word filtering algorithm, which aim to filter the noise or low discriminative and descriptive visual words. The filtering process consists of two steps: 1) calculate the probability distribution of the various visual words, and then, delete the words with gentle probability distribution; 2) delete the visual words with less instances. Furthermore we also consider multiple cues fusion, making the representation more complete. After generating a descriptive and comprehensive visual vocabulary, we choose three experiments to demonstrate the performance of the proposed method of this paper.

The paper is organized as follows: in section 1, a brief review of the generation of visual words and related work is given. Section 2 presents the details of the proposed visual words filtering algorithm and process of multiple cues fusion.
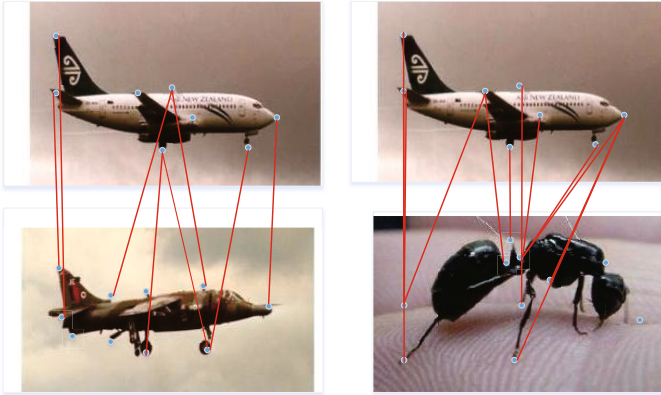
Fig. 1. Matched visual words between the same and different objects

Experimental results are presented in section 3 and finally, section 4 concludes the paper.

2   Visual Words Filtering Algorithm and Multiple Cues Fusion

In this section, we will first describe the relevant local features and the stages of the primary feature extraction pipeline used in this paper. Next, we will describe the details of the visual words filtering algorithm, which make the generated vocabulary more discriminative and descriptive. Finally, multiple cues fusion framework will be proposed. Based on these work, a complete representation method will be used for the image classification.

2.1   Local Features Extraction

Taking into account the extracted features are not stable, which means some local features are sensitive to affine transformations, in this section, robust shape, color, texture descriptors are presented. About shape, SIFT and Opponent-SIFT features are extracted, about color, Hue histogram is applied, about texture, LBP and RLBP are utilized.

SIFT. The SIFT descriptor proposed by Lowe [10] describes the local shape of a region using edge orientation histograms. To compute SIFT descriptors, the version described by Lowe [10] is used.

Opponent-SIFT. Opponent-SIFT describes all the channels in the opponent color space (*eq.* (1)) using SIFT descriptors. The information in the O_3 channel is equal to the intensity information, while the other channels describe the color

information in the image.

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{6}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix} \tag{1}$$

Hue Histogram. In the HSV color space, it is known that the hue becomes unstable near the grey axis. To this end, Van de Weijer *et al.* [17] apply an error propagation analysis to the hue transformation. The analysis shows that the certainty of the hue is inversely proportional to the saturation. Therefore, the hue histogram is made more robust by weighing each sample of the hue by its saturation.

LBP. The basic form of LBP is illustrated in Fig. 2 (a) and (b) [13] . Specifically, we use a local neighborhood around each pixel as input (see Fig. 2 (a)) and then threshold the neighborhood pixels at the value of the central pixel. The resulting binary valued string is weighted as follows:

$$LBP(I_c) = \sum_{i=0}^{P-1} 2^i s(I_i - I_c) \tag{2}$$

Where the parameter P runs over the eight neighbors of the central pixel I_c. I_c and I_i are the gray-level values at c and i, and $s(A)$ is 1 if $A \geq 0$ and 0 otherwise.
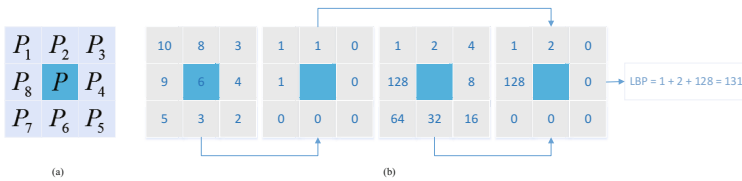
Fig. 2. LBP. (a) A pixel and its eight neighbors; (b) Basic LBP.

RLBP. An issue of LBP is that it is not so robust to the noise present in the image, about this question Chen *et al.* [4] improve the robustness of LBP by changing the coding bit of LBP. [4] assume that the case of sub-string (010) and (101) are noisy and respectively change them to a new sub-string: (000) and (111). An example of how to compute RLBP from LBP is shown in Fig. 3. And related experiments show that, this method, to some extent, improve the representation ability of image.

   The stages of the primary feature extraction pipeline used in this paper. First, the Harris-Laplace salient point detector is applied to the image. Then, for every point a local descriptor is computed over the area around the point. All the local
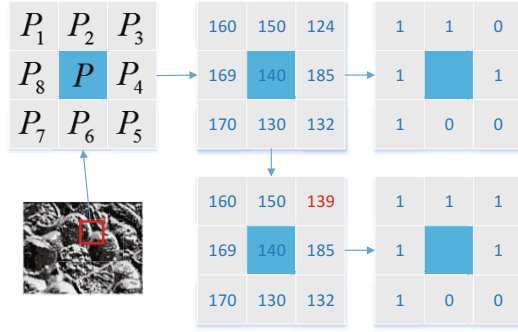
Fig. 3. Robust local binary pattern

descriptors of an image are subsequently vector quantized against a codebook of prototypical local descriptors. This results in a fixed-length feature vector representing the image.

2.2 Visual Words Filtering Algorithm

In order to generate the initial visual words, similar to existing works [12], [19], we train visual vocabulary by clustering a large number of local descriptors. We adopt classic K-means cluster to conduct the clustering. As we analyzed above, the visual vocabulary generated by the traditional method, there is a lot of noise, which greatly reduced the discriminative ability and descriptive capability, at the same time, it is bound to affect the late applications. Therefore, it is necessary to consider how to improve the descriptive power of visual words.

Motivation. Obviously, different classes of images include the discriminative subjects, for the same class of images, subject content is stable, while the background is unstable. Therefore, if the generated visual word have descriptive capability and discriminative power, it must possess two characteristics: 1) frequently match with the local features of a certain class of images, at the same time, with other classes is not so frequent; 2) the number of instances is more than training images of specified class. In view of these two properties, we propose visual words filtering algorithm, the details of this algorithm are showed in Algorithm1.

Details of the Proposed Algorithm. The pseudo-code of our algorithm, called **VWF** (for visual words filtering), is presented in Algorithm 1. **VWF** takes as input a training set of labeled local descriptors $S = \{X_k, \ y_k\}_{(k=1,...,m)}$. During the pseudo-code, there are two properties, named p_1, p_2, which p_1 means to calculate the probability distribution of the various visual words, and then, delete the words with gentle probability distribution; and p_2 means to delete the visual words with less instances.

Multiple Cues Fusion Framework. The Fig. 4 shows the multiple cues fusion framework, the main steps include local features exaction, visual vocabulary

Data: A set $S = X_k, y_{k(k=1m)}$ of labeled local descriptors, where X_k is the local feature descriptors set of kth image, $y_k \in \{1, \ldots, N\}$

Result: A final discriminative and descriptive visual vocabulary V'

Initialize the visual vocabulary V: Generate the initial visual vocabulary $V(v_1, \ldots, v_d)$ with K-means cluster

For Each visual word $v_i i = 1, \ldots, d$ do
        // Visual word frequency in each class of the image
        For Each class $j = 1, \ldots, N$ do
                Compute the frequency of occurrence of v_i: f_{ij}
        If(v_i doesnt meet the property p_1)
          Filter v_i
        Else
          Compute maximum of f : f_{im}, where m is class label
          //v_i doesnt meet the property p_2
          If (f_{im} ¡ the number of mth training class of images)
             Filter v_i
Return V', which is the rest of V

Algorithm 1. Pseudo-code of VWF

generation, feature quantification. During the process of the visual vocabulary generation, **VWF** algorithm is utilized. And the final representation of the image is weighted fusion of local features (shape, color, texture).

3 Experimental Results and Discussions

In the experimental phase, our visual words filtering algorithm and multiple cues combination method are tested for image classification. According to algorithms introduced in section 2 and section 3, we get a visual words histogram, then we employ SVM method to perform classification. In the experiments, we adopt SVM with RBF kernel in LibSVM library to train samples. The experimental environment is AMD Athlon X2 250 Processor, 4G memory, Windows 7 operating system. The programming languages used in the experiments are C++ and Python, and assistant libraries are OpenCV, LibSVM, numpy and scipy, among these, OpenCV was designed for computational efficiency and with a strong focus on real-time applications, numpy and scipy are scientific computation computational packages of Python, and LibSVM, which is SVM classifier library.

We perform image classification task on three image databases, which are Caltech_101 database, soccer team database and flower database, as shown in Fig.5. For Caltech_101 database, there are 101 classes, and contains 9146 images, in our experiment, we randomly select 10 categories, as for each category, including at least 40 images. The dominant cue of the Caltech_101 is unstable with the different categories. There are 280 images in soccer team database, which has 7 categories, and each category has 40 images. In soccer team database, color cue is the dominant cue. There are 1360 images in flower database, which has 17
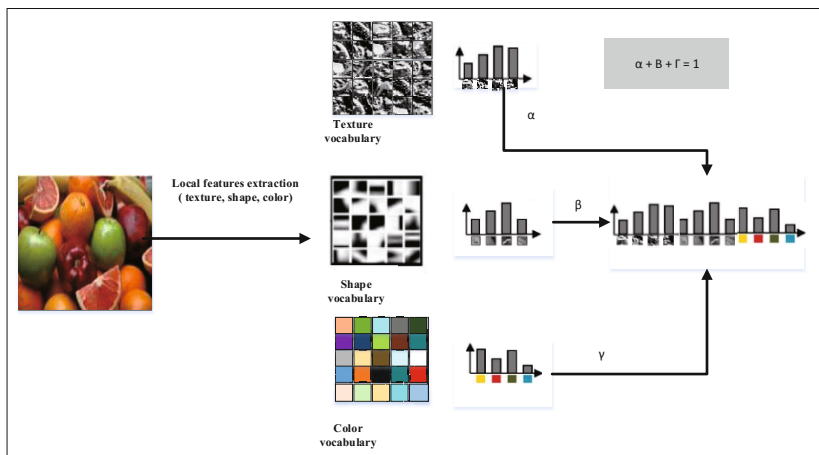
Fig. 4. Multiple cues (texture, shape, color) fusion framework

Fig. 5. Example images of Soccer, 17flowers and Caltech_101 databases

categories, and each category has 80 images. In flower database, color and shape cues are both important. We randomly select 70% of each category as training samples, with the rest as testing samples.

In the experiments, we totally considered five kinds of local features, they are SIFT, Opponent-SIFT, Hue histogram, LBP, RLBP, and we also choose SIFT, Hue histogram, LBP as a combination, the rest three features as another combination. These features and combinations are labeled with $1, 2, \ldots, 7$ respectively, as showed in the next figures and table.
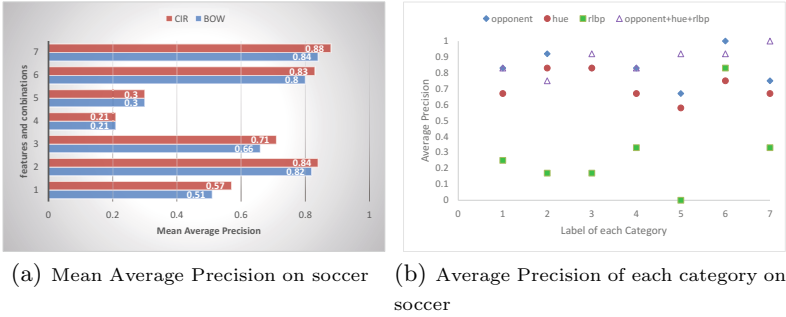
(a) Mean Average Precision on soccer   (b) Average Precision of each category on soccer

Fig. 6. (a) is the comparison of the proposed method with traditional method and (b) lists average precision of the proposed method on soccer database
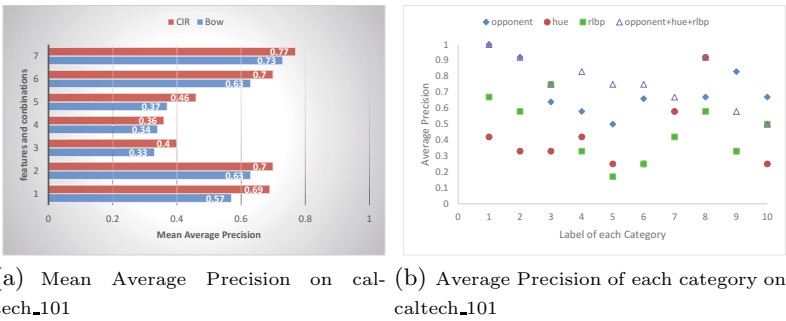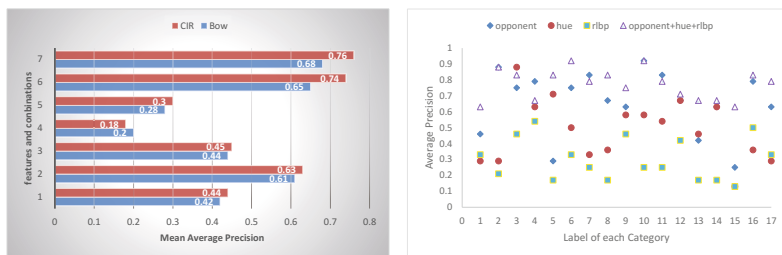

(a) Mean Average Precision on cal-   (b) Average Precision of each category on
tech_101   caltech_101

Fig. 7. (a) is the comparison of the proposed method with traditional method and (b) lists average precision of the proposed method on soccer database

Figure 6, 7, 8 shows the mean average precision and average precision of each category on the each database. As shown in the figures, compared with the traditional method of BoW, the classification rates improved considerably. Mainly due to the visual word filtering algorithm improves the visual word descriptive and discriminative ability to a certain extent.

Table 1 shows the classification rates on Soccer, 17flowers and Caltech_101 databases. In the table, columns named BOW list the results of traditional bag-of-words model, and columns named **CIR** (abbreviation of complete image representation) list the results of our method. We compare the algorithms with single feature, includes SIFT, opponent-SIFT, Hue histogram, LBP, RLBP, and feature fusion. From the table, we can see that when bringing in visual world filtering algorithm, the average accuracy will improves to some extent for both single feature and feature fusion, and the improvement range is $2\% \sim 12\%$.

(a) Mean Average Precision on 17flowers (b) Average Precision of each category on 17flowers

Fig. 8. (a) is the comparison of the proposed method with traditional method and (b) lists average precision of the proposed method on 17flowers database

Table 1. Classification rates on 17flowers, caltech_101 and soccer database

Feature and Combination	17flowers		Caltech_101		Soccer	
	Bow	CIR	Bow	CIR	Bow	CIR
1	0.42	0.44	0.57	0.69	0.51	0.57
2	0.61	0.63	0.63	0.70	0.82	0.84
3	0.447	0.45	0.33	0.40	0.66	0.71
4	0.2	0.18	0.34	0.36	0.21	0.21
5	0.28	0.3	0.37	0.46	0.30	0.30
6	0.65	0.74	0.63	0.70	0.80	0.83
7	0.68	0.76	0.73	0.77	0.84	0.88

4   Conclusion

In this paper, we analyzed the shortcomings of the traditional visual words, in order to improve the discriminative and descriptive ability of visual words and completely to represent the image context, the visual word filtering algorithm and multiple cues fusion method are proposed. In addition, we have test on three benchmark databases (Soccer, Caltech_101, 17flowers), validating our approach more.

Improving the descriptive and discriminative ability of visual word and combining multiple cues efficiently are still active research topics in multimedia and computer vision communities. In order to represent the image context completely, our future work will be carried out focusing on following two aspects: 1) the relationship between visual words and semantic will be explored; 2) weighting the cues with considering their statistical dependence in the application at hand.

References

1. Bosch, A., Zisserman, A., Muñoz, X.: Scene classification via pLSA. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 517–530. Springer, Heidelberg (2006)
2. Chang, S.-F., Ellis, D., Jiang, W., Lee, K., Yanagawa, A., Loui, A.C., Luo, J.: Large-scale multimodal semantic concept detection for consumer video. In: Proceedings of the International Workshop on Multimedia Information Retrieval, pp. 255–264. ACM (2007)
3. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods (2011)
4. Chen, J., Kellokumpu, V., Zhao, G., Pietikäinen, M.: Rlbp: Robust local binary pattern. In: BMVC (2013)
5. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (CSUR) 40(2), 5 (2008)
6. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 2, pp. 524–531. IEEE (2005)
7. Fernando, B., Fromont, E., Muselet, D., Sebban, M.: Discriminative feature fusion for image classification. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3434–3441. IEEE (2012)
8. Jegou, H., Harzallah, H., Schmid, C.: A contextual dissimilarity measure for accurate and efficient image search. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2007, pp. 1–8. IEEE (2007)
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2169–2178. IEEE (2006)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
11. Marszaek, M., Schmid, C.: Spatial weighting for bag-of-features. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2118–2125. IEEE (2006)
12. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 2161–2168. IEEE (2006)
13. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002)
14. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, pp. 1470–1477. IEEE (2003)
15. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 321–330. ACM (2006)

16. Van De Sande, K.E., Gevers, T., Snoek, C.G.: Evaluating color descriptors for object and scene recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 32(9), 1582–1596 (2010)
17. Van De Weijer, J., Gevers, T., Bagdanov, A.D.: Boosting color saliency in image feature detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(1), 150–156 (2006)
18. Winn, J., Criminisi, A., Minka, T.: Object categorization by learned universal visual dictionary. In: Tenth IEEE International Conference on Computer Vision, ICCV 2005, vol. 2, pp. 1800–1807. IEEE (2005)
19. Wu, Z., Ke, Q., Isard, M., Sun, J.: Bundling features for large scale partial-duplicate web image search. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp. 25–32. IEEE (2009)

Part VI

Database and Visualization Technologies

Design of VIP Customer Database and Implementation of Data Management Platform

Rui Li, Miao-Jie Sang, and Ke-Bin Jia[*]

Electronic Information & Control Engineering dept
Beijing University of Technology
Beijing 100124, China
kebinj@bjut.edu.cn

Abstract. Accurate and personalized analysis management for VIP Customer data are an important means for enterprises to make business decisions and win benefits. In this paper, take telecom industry as an example, SQL Server 2008 is applied to design the database according to the problem of traditional data analysis and management methods of VIP customers. Combined with Browser/Server structure, an intelligent VIP customer data management platform is designed and implemented, which achieves rapid extraction, effective analysis, early warning and graphic display for the data. In addition, it effectively guides customer managers to make business decisions, and provides managers with first-hand data on the customer managers' performance evaluation. The actual operation in a large domestic telecom company proves that the platform is stable, efficient and safe.

Keywords: VIP customer data, database design, B/S structure, graphical display of data.

1   Introduction

Nowadays, under the increasingly fierce market competition, enterprises improve business efficiency through information technology and create business value through data mining [1], VIP customer data with abundant amount of information plays an core role to enterprise's development. The way to make effective use of historical data and discover the deep inherent laws hidden in the data for extracting information oriented analysis and beneficial to decision has become the urgent attention problem for finance, telecom, oil and other core industries [2-4]. In this context, intelligent and systematic VIP customer data management platform has emerged.

Taking the telecom industry as an example, with the extensive application of computer technology, the telecom industry produces millions of customer data every month [5], including basic information, monthly income and arrears, situation of customer visit and maintain and so on. How to make use of the so much data effectively, and extract useful information for customer managers to make decision has become one of the main problems faced by the telecom's decision-makers [6].

---

[*] Corresponding author.

At present, the main way for telecom companies to analyze and manage VIP customers' data is still using the traditional manual record and Excel archive, applying simple statistical software, such as Excel to make report forms, then supporting business decision [7]. Such a way that does not play well on the supervision and management of customer data, as a result, the customer managers fail to grasp the customers' "health" status intuitively, which may hinder the business decisions, maintain and expand. In addition, lacking of the global analysis of data, the leaders cannot evaluate the customers' managers work reasonably, thus unable to fully mobilize the enthusiasm of customers' managers [8].

In this paper, we take telecom industry as an example, according to the problem of traditional data analysis and management above, we adopting SQL Server 2008 [9] to design the database, achieving useful data extraction from millions of data and effective calculation. Combined with Browser/Server structure, an intelligent VIP customer data management platform was designed and implemented. Setting up personalized "health profile" cards for each VIP customer and achieving effective analysis, early warning and graphic display for the data, which can effectively guide customer managers to make business decisions, and provided managers with first-hand data on the customer managers' performance evaluation.

This paper is structured as follows: Section 2 describes the ideas for designing database and the way for accessing data; Section 3 provides the integrated framework of the data management platform, and according to the data category, the function of each module is introduced; Section 4 discusses how to graphically display the results and guide business decisions. We will show the real interfaces of the platform in section 5. Finally, the characteristics and performance of the platform are summarized.

2   Database Design and Data Accessing

2.1   Database Design

Database design is mainly based on the elements involved in the system functional modules and the relationship between them [10]. Our platform applies SQL Server 2008 to set up the database. Due to the database is designed for VIP customers to store millions of data, it is not in place to optimize the database will directly affect the fluency and stability of the platform. Table design is the basis, which determines the robustness of the integrated platform.

We designed the data sheet according to the original data categories and the demand of the data management platform. As the data quantity of customers ownership, resources information, obstacles information, visit information and basic information of platform management is smaller, which means the load for database is lower, so we packaged those information in the same DB database file, formed resource occupied data sheet, obstacles information data sheet, service and visit information data sheet and customer information data sheet, the first three data sheet associated with the last data sheet by primary key ID and foreign key customers' name. The income and
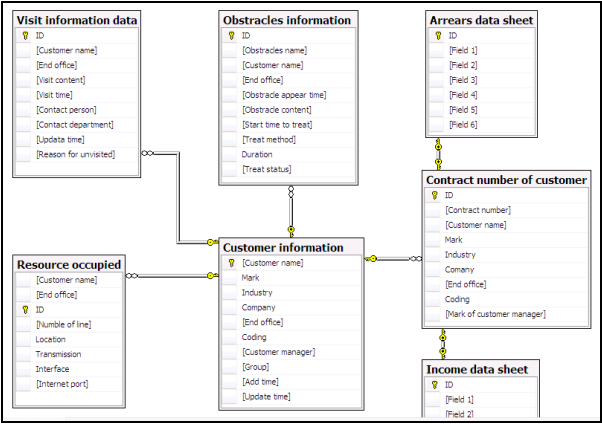
Fig. 1. Database structure

arrears data sheets contain millions of data, the operation is complex and resource consumption is serious, requiring high capability of hard ware, so we building an independent database for those data. The structure of the database is shown below in Fig. 1.

2.2   Data Accessing

As the analysis for income and arrear is an important means for VIP customer managers to make business decisions, in the same time, the database load is heaviest for data access, so here we focus on the access to income and arrear data.

We use database index to access income and arrears data. A database index is a data structure that improves the speed of data retrieval operations on a database table. Indexes are used to quickly locate data without having to search every row in a database table every time a database table is accessed. Indexes can be created using one or more columns of a database table, providing the basis for both rapid random lookups and efficient access of ordered records [11].

Due to the huge amount of income and arrears data of telecom VIP customers, if we calculate the Year and Chain and statistics calculate for each individual event over a period of time according to the original monthly data, that will greatly increase the load on the database, and also affect the loading speed of internet page. Therefore, we choose a manual way to generate summary worksheet every month, that is, when the monthly income and arrears data is automatically stored in the database through the interface, the customer manager generates a summary worksheet by the button shown in the home page, then the summary worksheet will store in the database. When the users want to calculate the statistics data which during a period of time, the database will extract the related data from the summary worksheet directly. That greatly improves the loading speed of system statistics page and contributes to the operation's efficiency of telecom companies. The generated summary worksheet is shown in Table 1.

Table 1. Summary worksheet

Cus-tomer name	Total revenue	Reve-nue increase on a year-on-year basis	Revenue increase on a month-on-month basis	Fixed telephone basic income	Internet income	Fixed telephone value-added service income	Data and network element revenue
Co. A	10735.57	0.00	0.00	5904.24	4583.33	248	0
Co. B	15080.81	0.00	0.00	15065.81	0	15	0
Co. C	29218.95	-100.00	2921895.00	21803.85	6270.1	145	1000
Co. D	65017.8	-9.23	10.17	25373.7	39004.1	640	0
Co. E	29852.77	0.00	0.00	27052.77	1791.27	8.73	1000
Co. F	10365.69	0.00	0.00	4320.69	5500	245	0
Co. G	57964.92	-100.00	5796492.00	55924.92	0	965	1075

3   System Framework and Functional Module

3.1   System Framework

The VIP customer data management platform applies the widely used B/S three-tier structure to separate presentation layer from business logic layer. The three-tier framework is shown below in Fig. 2.
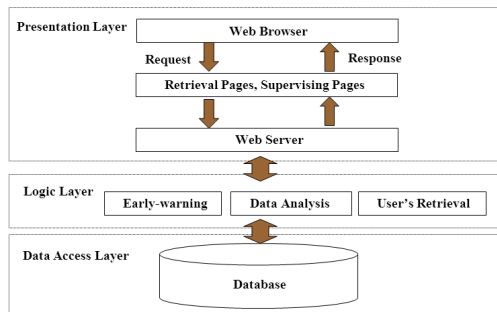
Fig. 2. B/S three-tier structure

B/S is the abbreviation of Browser/Server. It is a special example of the application of three-tier Browser/Server structure developed from the traditional two-tier Client/Server structure in the Web.

Compared with Client/Server structure, which required to install appropriate soft-ware in the users' computer, the B/S structure put major business logic in the server. Therefore, client not need to install any software, having a browser is enough. Thus

greatly simplifies the loads for client computer, reducing the cost of system mainten-ance and upgrades [12].

3.2   Functional Module

According to actual needs, the VIP customer data management platform is divided into front display unit for customer managers privilege and background management unit for leaders privilege. The former mainly consists of six modules, and the latter consists of two modules, each module holding corresponding sub-modules, which are shown in Fig.3.
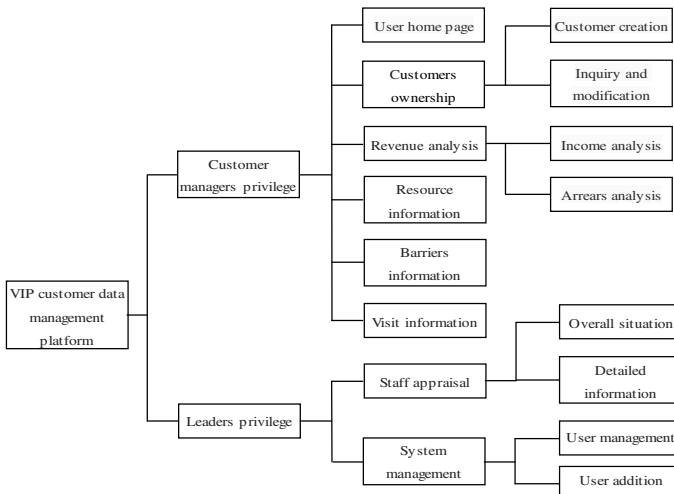

Fig. 3. Function modules of VIP customer data management platform

Next, the function of each module are described briefly:
For customer managers privilege:

- User home page module: Displays the warning information, provides function to generate the monthly summary worksheet, and shows the customers' information by "health card".
- Customers ownership module: To add, delete, change and inquire the customers' information and ownership. Each manager is responsible for several customers, they needs to reclaim their customers and manage their information.
- Revenue analysis module: The manager can inquire their own customers' income and arrears, either single month or multi-month, the result will be graphically dis-played.
- Resource information module: Record and manage the resource the customers occupied (such as equipment, cables).
- Barriers information module: Record and manage the barriers appeared in the ser-vices and the way to deal with them.

- Visit information module: Record the communication between managers and customers, include the visit time, content, results and other information.

  For leaders privilege:

- Staff appraisal module: Analyze the maintain state within the scope of all the customer managers, including income and arrears status, the trend of income and arrears, and so on. Those data support customer managers to make business decisions correctly.
- System management module: The leaders have the highest privilege to manage the users' information, they can add, delete, change, and inquire the users' information in this module.

4   Graphically Display the Result

Whether customers can bring benefits to the enterprise has become the concern focus for telecom operators. Fully understanding the income and arrears status of customers and the situation of customer maintaining is the basis to make business decisions. Therefore, the data management platform applies the graphical approach to display the revenue arrears and customer maintaining with pie charts and line charts, giving the intuitive understanding to managers, and then the managers can make better business decisions.

The platform displays data graphically with pie chart and line chart which come from Highcharts library. Highcharts is a charting library written in pure JavaScript, offering intuitive, interactive charts to web site or web application. It is solely based on native browser technologies and doesn't require client side plugins like Flash or Java [13].
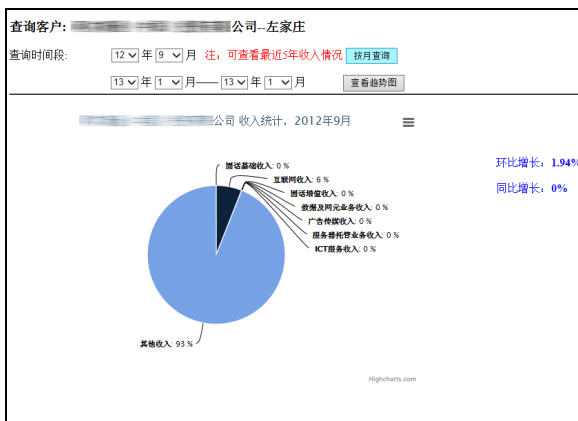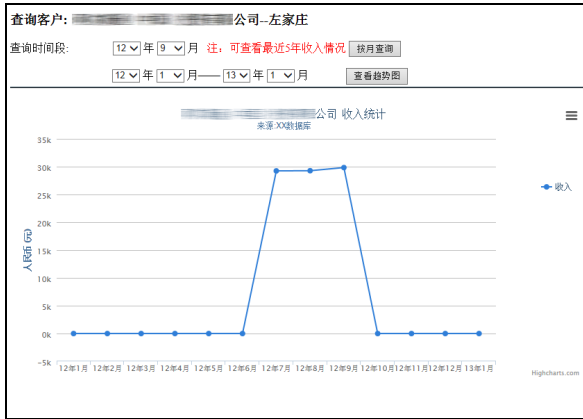

Fig. 4. Single month by pie chart

Fig. 5. Multi-month by line chart

In the revenue analysis module, in order to get an intuitive display effect, we show the income and arrears data from single month by using pie chart, which includes the specific amounts of each part. Income and arrears data from multi-month is displayed with a line chart, the customer managers get the trend of data from it. The pie chart and line chart are shown below in Fig. 4 and Fig. 5.

5   Example of System Interface

In the system, we use the basic development of Browser/Server structure. Combined with HTML and CSS, we achieve a static display of user information. Besides, we apply JavaScript, Ajax and JQuery technology to display user information and exchange data dynamically.

After login, the customer manager enters the home page, which is shown in Fig. 6.

Fig. 6. Home page for customer manager

The bar in the left side of the page shows the early warning information, which will notify customer managers the name list of delinquent customers in last month, delinquent customers for three consecutive months, customers' basic information without updating for three consecutive months and unvisited customers for three consecutive months, so the managers can take measures in first time to make sure the customers' information is fresh and effective.

The main content in the middle of the page is the "health card" for each customer. The manager can add, delete, change, and inquire all aspects information of the customer, thus achieving intelligent management of data and guiding business decisions.

After login, the leader enters the home page, that is the staff appraisal page, which is shown in Fig. 7.

Fig. 7. Home page for leader

The staff appraisal page statistical analyzes the overall status of maintaining and the detailed maintain result of each manager. Displaying those results by tables and graph can effectively provide leaders with first-hand data on the customer manager's performance evaluation.

6   Conclusion

According to the problem of traditional data analysis and management methods of VIP customers, and in order to meet the enterprises' needs of intelligent data management platform, we adopt SQL Server 2008 to design the database, and use the basic development of Browser/Server structure to implement an intelligent VIP customer data management platform, which achieves rapid extraction, effective analysis, early warning and graphic display for the data, effectively guides customer managers to make business decisions, and provides managers with first-hand data on the customer managers' performance evaluation. This platform has been applied in a large domestic telecom company. The actual operation proves that the platform is stable, efficient and safe. In future studies, the platform may be further extended to other industries with a broad application prospects.

References

1. Liao, P.-Y.: Optimal Pricing Strategy for Queuing Systems with Capacity Constraint Problem. In: Third International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP), pp. 561–564 (2007)
2. Li, R., Kong, X.: Practice of Data Mining in Modern Banking. In: The 2011 Asia-Pacific Youth Conference of Youth Communication and Technology, pp. 310–313. Atlantis Press, Hangzhou (2011)
3. Zhang, Y.: VIP Customer Segmentation Based on Data Mining in Mobile-communications Industry. In: 5th International Conference on Computer Science & Education (ICCSE 2010), pp. 156–159. sn, Hefei (2010)
4. Niu, Q.: Examine the Importance of Large Customers from a Strategic Height. Chinese Petroleum Corporation (6), 99 (2012)
5. Miaomiao, M.: The Research and Implementation of the Algorithms of Massive Data Processing in Data Mining. Xi'an University of Architecture and Technology, Shanxi (2012)
6. Qi, C.: The Strategies of Big Data Analysis, Application, Development and Management in Telecom Industry. Telecommunications Science (3), 12–16 (2013)
7. Dong, L., Tuo, C.: Design and Implementation of Telecommunications industry's VIP Customer Relationship Management System. Guide of Sci-Tech Magazine (27), 190–191 (2013)
8. Wang, H., Gao, M.: Evaluation and Application of VIP Customer Managers' Work in Telecommunications Industry. Computer CD Software and Applications (20), 138–140 (2013)
9. Chen, N.: Analysis of SQL Server 2008 Database Security. Computer Development & Applications 25(1), 64–66 (2012)
10. Zhuang, X.: Design and Implementation of Management System for Small and Medium Enterprises Based on B/S Structure. Ocean University of China, Qingdao (2010)
11. Liu, H., Chen, Z., Tu, G.: Database Indexing and Optimization. Modern Computers (12), 21–26 (2012)
12. Peng, P.: Library Management System Based on B/S Structure. Shandong University, Shandong (2013)
13. Zhang, Y.: Research and Implementation of Remote Monitoring System Based on Web BES In Detector. Zhengzhou University, Henan (2011)

Design and Implementation of Operation Energy Management System Based on AJAX-SSH2

Ye Yuan, Ke-Bin Jia, and Qi-Te Wang

College of Electronic Information and Control Engineering, Beijing University of Technology,
Beijing 100124, China
`yuanye_jason@126.com, kebinj@bjut.edu.cn`

Abstract. With the development of communication network construction, the operation consumption of communications industry grows significantly in recent years, and the high cost of electricity is hard to be ignored. Therefore, it is necessary to develop an operation energy management system to ameliorate the cost control of electricity. By integrating AJAX with SSH2 framework, AJAX-SSH2 lightens the burden of server and optimize users' interactive experience with the intrinsic performance of low-cost system maintenance and function extension. This paper designs and implements the Operation Energy Management System based on AJAX-SSH2 framework, achieving the functions include data entry, data query, report export, budget analysis, energy-saving evaluation, tendency prediction, and user management. This system provides a platform for centralized management and analysis in order to make the process of energy-saving and consumption controllable, which finally achieves the power optimization of communication network.

Keywords: SSH2, AJAX, Energy Management.

1   Introduction

With the rapid growth of the construction scale of communication network, power consumption of communications industry increases more than 15-20%, and the high cost of electricity is hard to be ignored. Because of the scattered distribution comprised of physical locations, it is difficult to achieve timely data integration and analysis with current off-line management system. Delicacy management is the core project of enterprise operation [1], which means it is necessary to establish an energy delicacy management mechanism to ameliorate management to control electricity cost.

With the development of information processing technology, B/S system, a web-based technology, becomes a magnet for system development [2]. SSH2 (Struts2+ Spring+Hibernate) is a lightweight Java EE development framework, bringing low-cost coping strategies of system maintenance and function extension, which has been attracted more attention by technical personnel at home and abroad in recent years [3]. However, SSH2 still use the traditional method of exclusive request in view layer, which may not only bring the burden to sever, but also cause page flickering when

frequently refresh the webpage using DIV+CSS, a mainstream technology in HTML5 [4], and finally limits the technique feature of SSH2 itself.

Therefore, this paper optimizes SSH2 framework by integrating AJAX technology to ease the burden of server and optimize users' interactive experience with the intrinsic performance advantages of SSH2. Then we design and implement the Operation Energy Management System based on AJAX-SSH2 framework to provide a platform for centralized management and analysis with the data sources came from all physical locations of fixed and mobile network. It makes the process of energy-saving and consumption controllable, which finally achieves the power optimization of communication network.

2   System Function Design

The operation energy management system is established in a dual channel management model between cash management channel and energy consumption channel, represented in three-dimensional space of application service axis, data information axis and time management axis. (Fig. 1). As for users, the functions are mainly embodied in the application service axis, which, according to the service logic, can be divided into seven modules include data entry, data query, report export, budget analysis, energy-saving evaluation, tendency prediction and user management. Each function module is independent from each other in application service axis, but shares data with five relational data tables include basic table, energy consumption record table, analysis record table, management table and classification table in data information axis, and processes in three cycles include month, season and year in time management axis, which finally implements energy delicacy management.
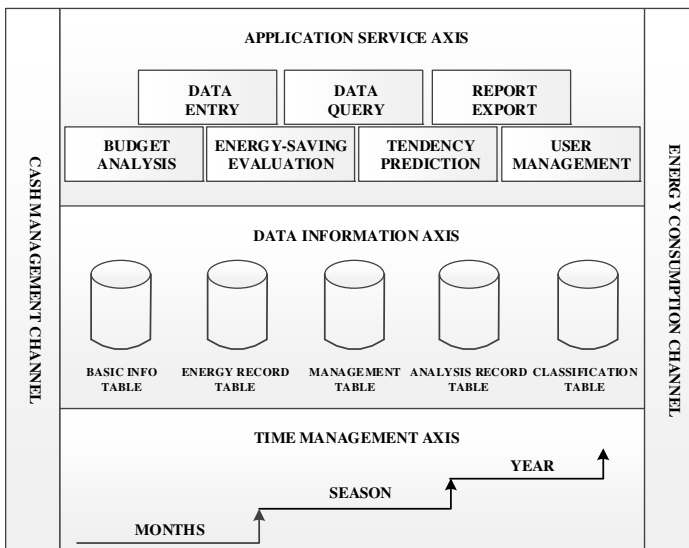

Fig. 1. Structure of System Function

1. Data Entry

Data Entry module accomplishes data classification among heterogeneous data of energy consumption, and data record to database server by either entry from single physical location or Excel files. Users can check the entry status of each physical location. The summarized data is supplied to other function modules.

2. Data Query

Data Query module accomplishes query function from historical data of energy consumption in database server. Users can choose date interval and type or input key words as index for one query operation.

3. Report Export

Report Export module accomplishes data export function by downloading Excel file from web server. Users can select user-defined attribute and content of data stored in database, and export as an Excel file, which satisfy the need of various forms of report.

4. Budget Analysis

Budget Analysis module accomplishes budget analysis function and provide 3 sub-modules classified by time management axis. Each sub-module provides analysis result include budget completion percentage, budget balance, and cumulative percentage, calculated by the data in database server. Users can view statistical graph charts as result to complete cost analysis. Users can also check the cost status from each physical location, which satisfy the need of budget monitoring in different time cycle in order to adjust power scheme and control cost in time.

5. Energy-Saving Evaluation

Energy-Saving Evaluation module accomplishes the management of energy-saving measures deployed in each physical location, and the evaluation of energy-saving effect by calculating the parameters of energy-saving ratio, average value, mean square root, and so on. Users can view statistical graph charts as result for aided optimization of energy-saving scheme.

6. Tendency Prediction

Tendency Prediction module accomplishes the prediction function by using the Exponential Smoothing prediction algorithm to predict the future trend of electricity cost. Users can view statistical graph charts as result to prepare budget for the future.

7. User Management

User Management module accomplishes user management. Only administrators can change users' permission include data access permission and function access permission in order to ensure all users perform their own duty.

3   System Structure Design

3.1   Physical Environment Model

According to practical situation, the physical environment model with tree structure is given based on the distributed management mode (Fig. 2). Every single physical location is monitored by corresponding end office pertained to fixed network, or base station pertained to mobile network. All end offices and base stations belong to different cost center classified by geographical position. Cost center for fixed network have one single sub-office, MSs (Module Station) and PASs (Personal Access System), while for mobile network include micro-base stations and macro-base stations.
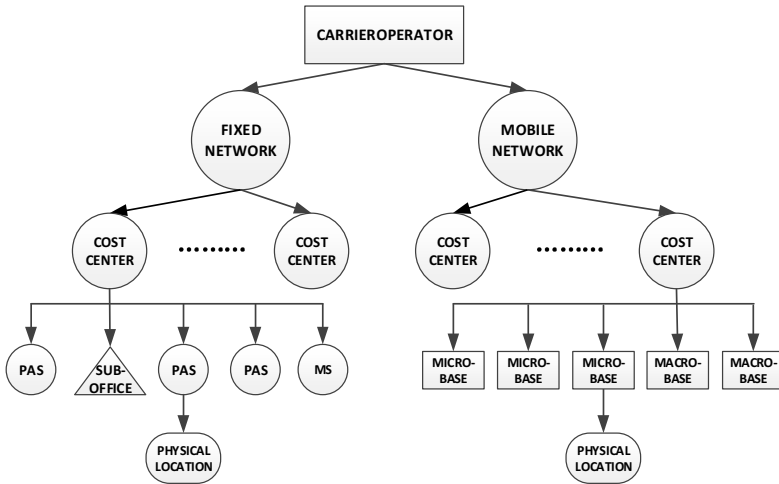
Fig. 2. Physical Environment Model of System

3.2   System Architecture

According to the physical environment model and actual demand, the operation energy management system uses B/S distributed multiplayer architecture as thin client. Application view is separated from actual business logic that means interactive operation directly faces users instead of business logic and database, which ensure the safety of background process and database. The system consists of four layers include application layer, view layer, business logic layer and data layer.

1. Application layer

Application layer provides different levels of access operation entrance for users with different authority. Users can visit system from any web browsers of PC, tablet computer and mobile device, instead of terminal client software dealing with different operation system.

2. View layer

View layer refers to web pages in web server, which provide various functional interactive interface for users. This layer only execute transfer and display of data between web server and web browser.

3. Business logic layer

Business logic layer embodies the main business functions of the system in web server, which is responsible for data package and data exchange from database server.

4. Data layer

Data layer refers to the database server, which consists of database engine and a variety of data tables. Database server and Web server can either be integrated into one single server or be placed separately.

4   Key Technologies

4.1   AJAX-SSH2 Framework

SSH2 (Struts2+Spring+Hibernate) is a Java EE development open source framework with lightweight MVC (Model-View-Controller) standard design pattern, which regard Spring framework as a core container to non-invasive integrate other frameworks with loose coupling structure [5]. Hibernate, an ORM (Object-Relational Mapping) framework, maps java class to the relational data table from database, treating as a thinking of object-oriented programming development [6]. Comparing with the early SSH framework, Struts2 provides integration interfaces for multi-technologies in view layer and completely frees between servlet API and business logic because of the interception mechanism based on WebWork core [7]. However, SSH2 still use the traditional method such as exclusive request in view layer, which may not only bring the burden to sever, but also cause page flickering when frequently refresh the webpage using DIV+CSS, the mainstream technology in HTML5, and finally limit the performance of SSH2 itself.

AJAX (Asynchronous JavaScript and XML) is a set of web development technologies, which has an independent engine between users and web server. AJAX engine [8] sends HTTP requests asynchronously to the server in the background by using JavaScript, and then uses DOM to update the content within a webpage marked by DIV instead of reloading the entire page. By integrating AJAX technology with SSH2 framework, AJAX-SSH2 can lighten the burden of server with the intrinsic performance advantages and make web applications more interactive and dynamic.

4.2   jQuery

jQuery is a lightweight open source JavaScript library, which provides a convenient API to simplify dynamic webpage development such as animation, event handling and AJAX [9]. jQuery is applied to this system to integrate website layout using

DIV+CSS and data transfer using AJAX, which works across a multitude of browsers with a combination of versatility and extensibility.

4.3   Highcharts

Highcharts is an open-source charting library written in pure HTML5/JavaScript, providing interactive charts to the web application [10], which, not like JFreeChart, is independent of Java, Flash or another plug-in. Highcharts currently supports more than 18 chart types include line, area, bar, pie, and so on by using SVG and VML technologies, and supports to export into PNG, JPG, PDF or SVG format with the exporting module enabled. Highcharts is supplied to this system to make dynamical statistic charts integrated by AJAX (Fig. 3).

Fig. 3. Dynamic charts using Highcharts in System

4.4   JXL

JXL (Java Excel API) is an open source JAVA API, offering Excel data operation [11]. It can handle complex Excel files except graphics and chart operation, and make it possible to edit Excel correctly in non-windows system. JXL is supplied to this system to dynamically read and write Excel files in web server (Fig. 4).

Fig. 4. Report Export Interface of System

5   System Implementation

5.1   Implementation Environment of the System

The operation energy management system is implemented in Myeclipse2013 with JAVA JDK-7u51, MySQL5.6 and Tomcat7, and accomplishes all business functions based on AJAX-SSH2 framework, Figure 4 shows the data entry interface of system as an example.

Fig. 5. Data Entry Interface of System

5.2   Structure implementation based on AJAX-SSH2

The work flow structure of system implemented by AJAX-SSH2 is shown in Figure 5. AJAX is used to optimize the interactive and dynamic of web view. Struts2 is used to control the correspondence between view and business logic. Hibernate is used to map the JAVA class to the relational data tables from database to persistence data. Spring, as the core IoC container of the AJAX-SSH2 framework, manages all instantiations from controller layer, business logic layer, DAO layer and POJO layer through a xml configuration file (ApplicationContex.xml). When a user from web browser sends a HTTP request, the controller components (Struts2 Controller) intercept the request and calls the business logic components of system. Then business logic components use bridge-connection pattern to call DAO components which is rely on SessionFactory and DataSource from POJO components to accomplish the data interaction from database.
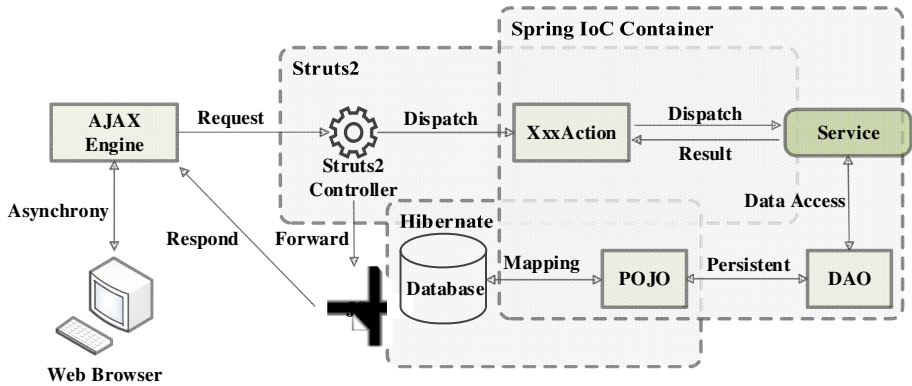
Fig. 6. Work Flow Structure of System

1. View Layer

View layer consists of AJAX engine and user interfaces of system written in JSP. Figure 3-5 show some examples of system interfaces. jQuery is used to integrate website layout using DIV+CSS and data transfer using AJAX, and use JSON format to execute data transition between users and web server.

2. Controller Layer

Controller layer firstly analyzes the execution data from view layer, then sends it to Struts2 core controller after filtering by a series of interceptors. According to the xml configuration file (Struts.xml), Struts2 core controller calls the correlative action class which, however, is actual created by Spring IoC container according to Bean ID. Spring uses dependency injection to inject instantiation into business logic layer, and Struts2 controller finally sends the result data to view layer after packaging JSP.

3. Business Logic Layer

Business logic layer, consists of all actual complete models of system function, is dependent on atomic operations provided by DAO components. The java classes in this layer, also needed to deploy in Spring in the xml configuration file (Application-Contex.xml), do not care how to get data from database which is the responsibility for DAO components. In this way, it is possible to make database applications freely change among different persistence technologies.

4. DAO Layer

DAO components in this layer encapsulates the atomic CRUD (Create, Read, Update and Delete) operations again based on POJO components, and only has the responsibility for caring the changes of persistent technology, which implement bridge mode between business logic layer and DAO layer. Spring supports interfaces for common persistence technologies such like Hibernate. Hibernate needs to inherit 'Hibernate-DaoSupport' class to implement the DAO components. Spring IoC container injects 'SessionFactory' for Hibernate and 'DataSource' for database configuration.

5. POJO Layer

POJO components in this layer maps the JAVA class to the relational data tables in database by xml files (*. Hbm.xml) corresponding to the tables. So, the programmer of JAVA can manipulate relational database with the thinking of object-oriented programming, which, as a result, reduce the development workload for writing SQL code.

6   Conclusion and Prospect

In this paper, we present an energy delicacy management model, optimize SSH2 framework by integrating AJAX technology called AJAX-SSH2 to lighten server and optimize users' interactive experience on the basis of the intrinsic performance advantages, and finally design and implement the Operation Energy Management System based on AJAX-SSH2 framework, achieving the functions include data entry, data query, report export, budget analysis, energy-saving evaluation, tendency prediction and user management. This system provides a platform for centralized management and analysis in order to make the process of energy-saving and consumption controllable, which finally achieves the power optimization of communication network.

This system will be further developed as the following:

1. Remote smart metering [12] will be added in data entry module, which integrates to establish an internet of things system.
2. Prediction algorithm need to be optimized to improve the accuracy of prediction.
3. Some personalized functions for users will be considered such as stations text messaging.
4. The system will be further tested for system performance, and we will pay more attention to data backup and disaster recovery.

References

1. Guan, B.: Delicacy Management under the New Economic Situation of Chinese Enterprises. China Management Magazine 2012(7), 573–574 (2012)
2. Hamed, O., Kafri, N.: Performance Testing for Web based Application Architectures (.Net vs. Java EE). In: 2009 IEEE 1st Networked Digital Technologies Int. Conf. (NDT), pp. 218–224 (2009)
3. Wang, X., Ruan, H., Li, H.: A Software-reuse Method from Model1 to SSH2. In: 2013 IEEE 37th Annual Computer Software and Applications Conf. Workshops (COMPSACW), pp. 470–474 (2013)
4. Li, A.: HTML5 Standard Implementation and Future Development. Information Technology & Standardization 2012(11), 9–12 (2012)
5. Walls, C., Breidenbach, R.: Spring in Action. Manning Publications, Greenwich (2011)
6. Linwood, J., Minter, D.: Beginning Hibernate. Apress, New York City (2010)
7. Li, G.: Lightweight Java EE Enterprise Application Actual. Publishing House of Electronic Industry, Beijing (2012)

8. Ying, M., Miller, J.: Refactring legacy AJAX applications to improve the efficiency of the data exchange component. The Journal of Systems and Software 86(2013), 72–88 (2013)
9. Bibeault, B., Katz, Y.: jQuery in Action. Manning Publications, Greenwich (2010)
10. Kuan, J.: Learning Highcharts. Packt Publishing Limited, Birmingham (2012)
11. Li, Y.: Implementation of Excel Documents Read and Write in Java Web Development. Computer Programming Skills & Maintenance 2012(8), 17–18 (2012)
12. Benzi, F., Anglani, N.: Electricity Smart Meters Interfacing the Households. IEEE Transactions on Industrial Electronics 58(10), 4487–4494 (2011)

Network Operation and Maintenance System of Meticulous Management Based on Data Analysis

Qi-Te Wang, Ke-Bin Jia, Ye Yuan, and Yu-Xin Song

College of Electronic Information and Control Engineering,
Beijing University of Technology, Beijing100124, China
wangqite1@163.com, kebinj@bjut.edu.cn

Abstract. Due to rapid development of the domestic communication industry, energy consumption on network operation and maintenance has become one of the main energy consumptions in China. With the aim of obtaining meticulous managementon enterprise power, raising electricity availability and evaluating the effect of energy-saving measures, it's crucial for us to develop an intelligent system for data analysis. We first introduces a management system based on B/S architecture and MVC framework with multi-functions of information inquiry, budget analysis, energy management and etc.; and then focus on key-technologies such as database modeling, database index, stored procedure and trigger and least square method. This intelligent system has been successfully employed in a communication enterprise and has been proved accurate, stable and efficient.

Keywords: MVC, Database Model, Database Index, Stored Procedure, Trigger, Least Square Method.

1 Introduction

In recent years, with the rapid economic growth, energy consumption growth is becoming a serious problem, which limits economic forward. Energy conservation is considered to be the key for solving this problem. Within communication industry, the large demand of information in communication network leads to constant power consumption increasing with average growth rate between 15% and 20% in recent years. The trend increases the country energy burden. So the management of energy consumption meets great challenges when dealing with problems such as high power consumption, lower power-efficiency and high cost. It obtains great significance to collect electric quantity statistics, budget electricity charge and predict electricity trend.

Based on these situations, we propose an intelligent network operation and maintenance system of meticulous management. The system mainly has six function modules: information record, information query, statistical analysis, data prediction, energy management and reports generation. According to the actual electricity consumption, statistical algorithms are applied to analyze effective intelligent energy-saving measures.

Least square method is employed to predict the energy consumption in the future time. Related finance reports are generated from database.

This paper is divided into five parts. Part I introduces the overall system. Part II describes structure and function of the system. Designing principles and key technology are explained in Part III. Part IV shows the system function. At last we summarize the characteristic and advantage of this system.

2   The System Structure and Function

2.1   System Structure

In order to facilitate system software development, configuration deployment, operation and maintenance, our system applies Model-View-Controller framework. As shown below in Figure 1, the system contains three layers: the view layer, business logic layer and data persistence layer [1, 2].
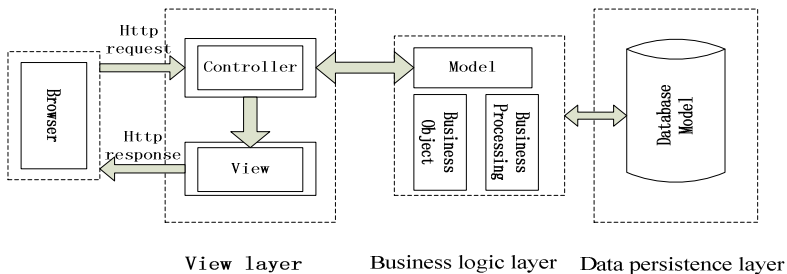
Fig. 1. The System Framework

In MVC framework, users access browsers to send requests. Controller answers the requests from users and then accesses logic layer to process data. Based on users' requests, logic layer gets data from data persistence layer and returns to controller [10]. Controller delivers results to specific views. Finally, users obtain what they want on the browsers.

2.2   System Function

Based on the requirement analysis of actual conditions, this system mainly consists of six functions7: information record function, information query function, budget analysis function, data prediction function, energy management function and report generation function. Function diagram is shown below in Figure 2.
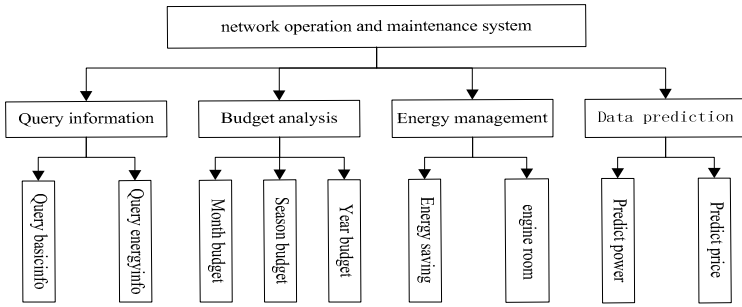
Fig. 2. The System Function

Here, we will make a brief introduction to each function module:

1. Information record function

This function records power information, electricity charge information and basic information. Generally, system only records basic information once unless it is modified.

2. Information query function

Corresponding to information record function module, this function module includes two sub-functions: basic information query and energy information query. Both sub-functions have the same query conditions: engine room name, engine room classification and cost center. Users could choose the appropriate conditions to query information.

3. Budget analysis function

Based on actual budget electricity charge, system would compute parameters including percentage of budget and budget balance, which are defined to accomplish budget warning.

4. Data prediction function

By building proper mathematical model, data prediction function could predict future power and electricity charge based on Least Square Method. This function could help users grasp the trend of energy consumption.

5. Energy management function

Users can add or delete energy-saving measures and update the time when certain measure begins or ends in this module. For a single engine room name, this function helps users analyze energy consumption conditions by calculating percentage of energy-saving.

6. Report generation function

Report generation function holds two sub-functions: report upload and report download. Report upload function helps users record information to the system by EXECEL forms so that users could upload batch data to the system. Report download function helps users download information from database by choosing attributes that satisfy the demand in different periods.

3   Principle of Design and Key Technology

3.1   Design Principle

Time-Data-Application Model in Three Dimensional Space
From the logical perspective of the system functions, the design of this system is based on a three dimensional space of time management, data service and application operation [8].

Time axis refers to time information, which helps check and management. Data axis refers to electric quantity and electric charge information, which is easy to calculate, count, analyze and predict the business logic computation.

The Idea of Electric Quantity and Charge Separation
In the process of this company's operation, energy consumption is normally divided into two sorts of data attributes, which obey the proportional relationship, and the index is the electric charge unit price. However, in the real life, the equipment records the real quantity of electric consumption, which is easy to calculate according to the electric charge unit price given by the administration of power supply. At that time, the electric quantity and electric charge have the congruent relationship. From the financial point of view charge does not mean quantity. If the company overpays the electrical bill in one month, the extra part will be amortized in another month, which is called Electric Charge Amortization. The payment on behalf will be collected by the financial department. This process is called Recovery of Electric Charge.

Because of the arising of electric charge amortization and electric charge recovery, electric charge can't be compared with electrical quantity in daily work of the company. For this reason, the system separates the electric charge from the electric quantity.

3.2   Key Technology

Database Model

Database Tables Building
Designing database tables is equivalent to building database structure. Because database tables have great influence on system efficiency, in the processing of designing tables, sufficient consideration of database norm form are taken to meet requirements such as reasonable allocation of storage location, easy information query and scalable interfaces etc.[4].

In this paper, we create 15 different kinds of tables based on application function re-spectively used to store basic information, fixed network power information, fixed-line electricity information, base station power information, the base station electricity information, fixed-line recovery information, amortization fixed network information, base station recycling information, amortization base station information, energy con-servation information, user information management, etc.[12].
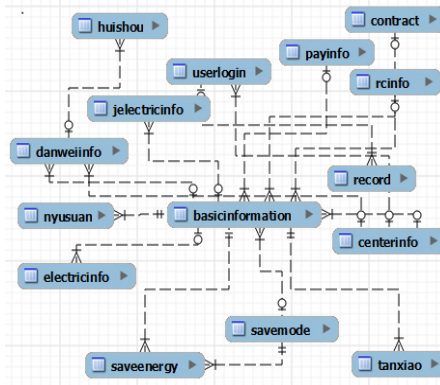

Fig. 3. The System Database Model

The system of the selected database of the MySQL database system is suitable for small and medium-sized enterprises, which has many characteristics including unli-mited number of users at the same time, fast, simple user permissions settings, and so on[3]. This system has the function of modeling using the workbench, the data table is shown in figure 3.

In the process of the design of the data table, we use the advantages and characte-ristics of foreign keys flexibly, effectively prevent double redundancy of data, and update foreign key at any time [13]. For example, the cost center of the basic informa-tion table of engine room, the attribute of engine room, electricity type, payment me-thods, etc. established the corresponding information connection by foreign key, avoiding the fixed information repeated of the basic information table of engine room, facilitating information update at the same time.

Database Index

In relational database, database index is a data structure related to database tables. During the processing of querying information, sql statements execute much faster if they correspond to database index [7].

In this system, in order to speed up query, we add index to the attributes which are the common query condition: room name, room classification and cost center. We use the B-Tree index in this system. B-Tree is a balanced binary tree structure [11]. Data are stored by the leaf nodes of tree structure. The minimum distances between root node and leaf nodes are the same.

Data are stored in B-Tree in key-data form. In the processing of B-Tree retrieval, binary search begins from root node. If result is found, return the result. Otherwise, the node which is pointed by pointer begins recursive search until result or null poin-ter is found.

Database Stored Procedure

Database stored procedure is a set of SQL statements to complete specific function in large database systems. After compiled, procedure stored in database is stored in the database. Users call stored procedure by name (and setting parameters if necessary) [5].

Stored procedure is a kind of modular design. It could be called repeatedly to accomplish business logic. If several sql statements are used to update database, stored procedure could contribute to reducing access and connection to database aimed to improve the efficiency. Furthermore, operations become much safer by setting stored procedure permissions.

In the system, the algorithm of budget analysis module is as follows:

Set B_i gross total electricity budget of engine room in month i; A_i total cost of the total actual electricity of engine room in month i; K total power of engine room year before in month i; and set K^* total power of engine room year after in month i;

Parameter 1: budget completion percentage in month i. $C_i = A_i / B_i$

Parameter 2: Budget balance in month i. $D_i = B_i - A_i$

Parameter 3: budget completion percentage by the end of month i.

$$E_i = \sum_{i=1}^{i} A_i / \sum_{i=1}^{i} B_i$$

In energy management module, system computes parameters below. Take budget completion percentage for example, system calculate electric quantity before and after adopting energy conservation. The process is below in Figure 4.

```
mysql> call energy
('3Q200903051003',1,@a,@b);
Query OK, 0 rows affected

mysql> select @a;
+------------+
| @a         |
+------------+
| 13000.0000 |
+------------+
1 row in set

mysql> select @b;
+------------+
| @b         |
+------------+
| 9901.0000  |
+------------+
1 row in set

mysql> select round(@b/@a,2);
+----------------+
| round(@b/@a,2) |
+----------------+
| 0.76           |
+----------------+
1 row in set
```

Fig. 4. System Stored Procedure

Database Trigger

Database trigger is a specific database stored procedure by event-triggered execution instead of program-call execution. When data change occurs in the database tables caused by update, insert or delete operation, corresponding triggers execute automatically.

In order to reduce entry items and make system more intelligent, indirect items such as electric charge unit price, PUE and payment cycle can be calculated by triggers based on basic entry items.

Triggers are divided into two types: Pre-trigger and post trigger. Pre-trigger execute sql statements before insert, update or delete operation. If sql statements are executed unsuccessfully, insert, update or delete won't operate. In contrast, post trigger execute sql statements after insert, update or delete operation.

In this system, pre-trigger is applied to prevent mistakes. For example, when users enter electric information into system, trigger sql statements will examine electric charge unit price if reasonable. If the result can't meet the conditions, insert operation is not performed. System will inform users of checking this record information.

Least Square Method

Least square method is a mathematical optimization algorithm. We predict data by minimizing the sum of least squared error. From the aspect of geometric, the target is to solve the curve fitting equation, which holds minimum distance from point such as $(x_i, y_i)(i = 0,1,2...k)$.

According to the enterprise's historical data records, the electric charge from January 2011 to July 2013 is shown in table I.

Table 1. Electric Charge Data（uint：yuan）

Year Month	2011	2012	2013
1	736592.04	798022.66	986120.02
2	725773.21	785483.62	995621.33
3	718965.90	775634.51	987456.50
4	738876.87	796677.49	993465.33
5	759045.34	804322.05	999564.11
6	749855.66	795643.44	987745.33
7	738184.84	792627.30	991662.10
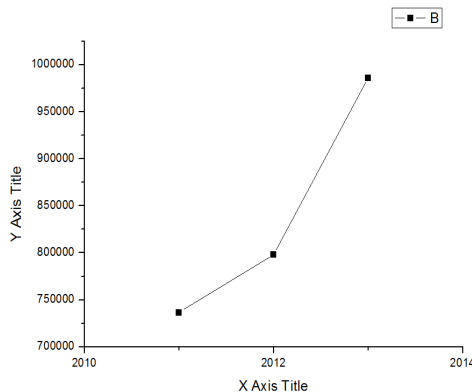
Fig. 5. 2010-2013 January Electric Charge Tracing Point Graph

Take January data from 2011 to 2013 for example, after drawing graph which is shown in Figure 5, it can be approximated by an exponential curve. We can choose the fitting function: $I = ae^{bv}$ (a, b are constants) and then we solve value of a and b.

To solve coefficients in the fitting function, we should transfer linear model to non-linear model [6]. By logarithmic operation, we can get the function: $\lg I = \lg a + \lg b$ .To simplify ,we can assume it $m = A + Bn$.

Table 2. Caculate Table

n	*1*	*2*	*3*
m	5.87	5.90	6.00

After computing, we determine the parameters *a* and *b*. and then we get the fitting equation: $I = 620860034e^{1.161v}$. Furthermore, we predict the electric quantity in January 2014. The predicted value is 1000120.02, and the true value is 1023010.11. The predicted error is only 2.3%. So the model is very reasonable.

4   System Implementation

The system administrator can access to the module of query, budget, energy-saving management and so on via the Private Network.

In the query information module, the users can select between the starting time and finishing time and choose the data they want from the whole system. In order to divide further data volume, the system sets three ways, names of locations, types of engine rooms and cost centers respectively. As is shown in Figure 6:

Fig. 6. Query Module Example

In the budget analysis module, when making specific analysis of budget, taking the consequence of network room in January 2013 for example, as shown in figure 7, the system has shown the whole budget of the year, in the meantime, we see the actual number of electric charge. There are parameter values of completion percentage, remaining budget and accumulative percentage.
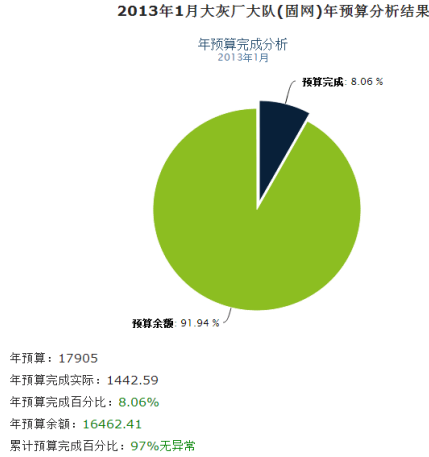
Fig. 7. Budget Module Example

In the energy-saving module, aiming at single engine room, the system gives the concrete energy-saving situation and explains the implementation effect of the particular energy conservation measure. Taking aptitude heat exchange technology for example (shown in Figure 8), the system reveals all the comparisons between electric usages before and this technology is applied. With the table below, users can see the total amount of electric quantity before and after energy saving. Moreover, it has listed each percentage that energy has been saved.
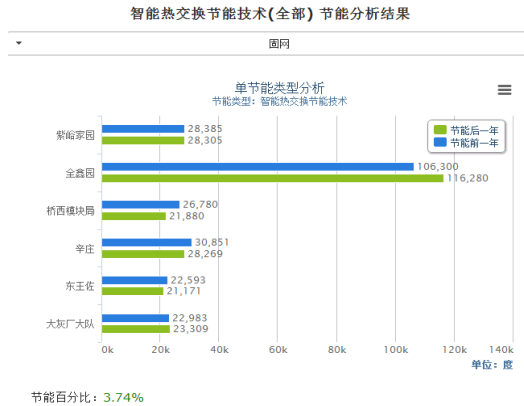
Fig. 8. Save-Energy Module Example

5   Summary

This system has acquired the function modules of budget management, energy-saving analysis and electricity prediction. Currently it has been used in a company's private network. The database and data analysis play an important role in this system. It helps this company improve the work efficiency of energy consumption management.

References

1. Ren, Y.C., Xing, T., Xing, Z.F.: Application research for integrated SSH combination framework to achieve MVC mode. In: Computational and Information Sciences, pp. 499–502 (2011)
2. Mcheick, H., Qi, Y.: Dependency of components in MVC distributer architecture. In: IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), pp. 691–694 (2011)
3. Bovenzi, A., Cotroneo, D., Pietrantuono, R., Russo, S.: On the Aging Effects due to Concurrency Bugs: a Case Study on MySQL. In: Software Reliability Engineering (ISSRE), pp. 211–220 (2012)
4. Chen, A., Liu, L., Shang, S.: A Hybrid Strategy to Construct Scientific Instrument Ontology from Relational Database Model. In: Computer Distributed Control and Intelligent Environmental Monitoring (CDCIEM), pp. 25–33 (2012)
5. Li, P., Lu, C.: Data analysis system research and implementation based on stored procedure. In: Communication Software and Networks (ICCSN), pp. 689–691 (2011)
6. Dutta, S., Overbye, T.J.: Prediction of short term power output of wind farms based on least squares method. In: Power and Energy Society General Meeting, pp. 1–6 (2010)
7. Wu, D., Wu, S.: Dynamically Maintain The Teaching Examples of Triggers and Stored Procedures about The Course of Database. In: Application. Education Technology and Computer (ICETC), pp. 525–527 (2010)
8. He, S., Liu, G.: Design and Implementation of Log Management Module in Three-dimensional Spatial Database Management System. In: Geoinformatics, pp. 1–5 (2010)
9. Zhao, E., Li, Z.: Multi-based Database System Development about Dam Safety Monitoring. In: Computational Intelligence and Design, pp. 422–425 (2008)
10. He, Z., Zheng, J.: Design Implementation of Student Attendance Management System Based on MVC. In: Management and Service Science, pp. 1–4 (2009)
11. Mukherjee, P., Kr., Nasipuri, M.: Indexing and searching in multimedia database management system. In: India Annual Conf., pp. 143–146 (2014)
12. Jayanthi, S.K., Prema, S.: Referenced attribute Functional Dependency Database for visualizing web relational tables. In: Electronics Computer Technology (ICECT), pp. 1–5 (2011)
13. Cagiltay, N.E., Topalli, D., Aykac, Y.E.: Abstract conceptual database model approach. In: Science and Information Conference (SAI), pp. 275–281 (2013)

Image Annotation and Refinement with Markov Chain Model of Visual Keywords and the Semantics

Zhong-Hua Sun and Ke-Bin Jia

College of Electronic Information and Control Engineering,
Beijing University of Technology, Beijing, China
`sunzh@bjut.edu.cn,kebinj@bjut.edu.cn`

Abstract. This paper presents a discriminative stochastic method for image annotation and refinement. We first segmented the images into regions and then cluster them into visual blobs with a small number than the whole training image regions. Each visual blob is regarded as a key visual word. Given the training image set with annotations, we find that annotation process is conditioned by the selection sequence of both the semantic word and the key visual word. The process could be described in a Markov Chain with the transition process both between the candidate annotations and the visual words set. Experiments show the performance of this annotation method outperforms the state of art methods.

Keywords: Image annotation, Markov Chain.

1   Introduction

Automatic image annotation refers to labelling image into obvious content descriptive words. Although it is easy for humans to recognize an image content, it has been a proven challenge for computers. At present, automatic image annotation solutions could be classified into two categories. The first one is generative model, such as Co-occurrence Model proposed by Mori et al [1], which annotated the images with words by counting the probability of associating words with segmented image regions. Translation Model [2] improves the annotation performance by describing images with a vocabulary visual words. Jeon J. et al proposed Cross-Media Relevance Models to take advantage the joint distribution of words and blobs, in such way, image content is described both with image features (blobs) and text (words) [3].The second category methods are discriminative models, such as Bayesian point estimate method [4], 2D-MHMM model [5] and SVM based model [6].

The general process of image annotation could be carried in the following steps:
1) Image Visual Unit Segmentation. In this step, images are usually segmented into equivalent blocks on a regular grid [7] or into regions covering main objects in the image [2]. The irregular regions are further clustered with a criterion. Only the one that is larger than a threshold could be reserved or else be discarded. The clustered regions are named visual blobs. All images are represented with a set of blobs. Each blob is numbered with a unique integer.

2) Descriptive Visual Content Feature Computation. For each image visual unit, low level visual features are extracted such as color, texture, position and shape information. All the computed features are stored in a high-dimensional vector. For example, in [2] 33 features are computed from the images.

3) Visual Appearance and Text (words) Correlation Model. In this step, effort is focusing on finding the relationship between the distribution of visual unit and annotated words from probability aspect or mapping from visual unit to annotated words based on machine learning theory. How to look at the relationship between visual information in an image and the semantic descriptive keywords affects the performance of image annotation results. For example, the Co-occurrence Model requires large number of training samples and every visual blob tends to be annotated with frequent words. Translation Model will require building more classifiers according to the number of training words.

To decide an annotation scheme, the above steps are interrelated. In CMRM, Jeon did not assume that there was an underlying one-to-one correspondence between visual blobs and the annotating words. The only assumed that a set of keywords $\{w_1 \dots w_n\}$ is related to the set of object represented by blobs$\{b_1 \dots b_m\}$. The number of them may not be equivalent. Although its annotation performance outperforms Co-occurrence Model and Translation Model, we find that CMRM neglect the correlation between the annotated words. There may be obvious conflict word topic in an annotation. For example in Figure 1, some annotated words may be irrelevant to the topic described by the word set. It could be seen that the word Beach may be treated as a noisy result. For this reason, the connection between the annotation results could be refinement further.

Fig. 1. Dear, Tree, Sky, Snow, Beach

Another problem in CMRM is the lack of consideration of the spatial distribution among the visual blobs. Lu et al in [7] proposed to use spatial Markov chain to represent the correlation between the blobs. Each image is segmented into equivalent blocks on a regular grid and scan the blocks in row wise raster direction. In this way, 2D blob sequence is formed. However, Lu's method only describe the spatial layout of the succeeding image blocks. In another side, the equivalent segmentation of the image would introduce object over segmentation, which affects the distribution of blob to blob transition probability. Zhou et al [11] proposed a multi-instance multi-label (MIML) learning framework that can successfully represent complicated object which have multi-label semantic meanings. In [8] Guang et al considered label correlation as an undirected bipartite graph, in which each annotation was considered correlated by a common hidden topics. In this way, the irrelevant labels could be removed from the results. Similar work is reported in [12], which linearly combined

different graph Laplacians through a set of optimal combination coefficients. In [13] Wu et al proposed a tag matrix completion method. The probability of a tag is assigned to the given image is described in the matrix, whose column corresponds to a tag and the row corresponds to the image. Both tagged images and untagged images are exploited to find the optimal tag matrix.

Our primary purpose in this paper is to show that the annotation process could be presented in Markov chain by considering both the relationship between the candidate keywords and the sequence selection of the annotation word. The following paper is organized as follows. We discuss the Markov chain model of visual blobs and text in section 2. In section 3 we show the experiment result of our method. Section 4 concludes this paper.

2 Image Annotation Based on Markov Chain Model

We propose an extended generative model, which uses the Markov transition probability instead of the joint probability for image annotation. In generative methods, images are first segmented into regular blocks or the irregular visual object regions. The right automatic visual object segmentation will affect the correctness of the visual words construction. Here we use the rival penalized competitive learning (RPCL) algorithm [9] to refine the results when extracting image visual vocabulary of blobs.

2.1 Visual Blob Clustering and Refining

Suppose we have a training image collection T with annotated word set W. Each image $J \in T$ could be represented with the combination of visual regions and the semantic descriptive words: $J = \{b_i; w_j\}$, where b_i represents the blobs corresponding to regions of the annotated image. For an annotation work, an un-annotated image $I \in C$ could be modeled in a probability distribution $P(W \mid I)$, which is named as the relevance model of I. The training annotated words $\{w_i\}$ is selected that maximizing $P(W \mid I)$, which could be described as $P(W \mid b_1...b_m)$. The probability of selecting a word is conditioned on the visual blobs, which are extracted from the training image collection. In [14] Duygulu et al used Normalized-cuts [15] to segment images into object regions and then extracted features such as color, texture, position and shape of the object region. Regions with smaller features than a threshold are ignored. Then K-means clustering (k=500) is applied to cluster the regions based on these features. However, the K-means clustering algorithm has two obvious limitations: one is that K-means is too sensitive to the initial cluster centers; the other is K-means cannot get the appropriate number of clusters in advance. Therefore, the K-means algorithm can seriously affect clustering and image annotation results.

In this paper we use rival penalized competitive learning algorithm (RPCL) [9] to optimize K- means clustering results. As is described in [9], the salient advantage of the RPCL algorithm is that it can automatically determine the number of clusters

while avoiding errors caused by the preset. Meanwhile, it also can solve the K-means limitation of selecting initial cluster centers. Consequently, we can create a reasonable visual vocabulary $\{b_1, b_2, ..., b_n\}$ to represent the visual content of image regions. In Section 3 experimental results showed that this improved algorithm can effectively improve the annotation performance of annotation with the CMRM.

2.2   Cross-Media Relevance Model and Hidden Markov Model for Image Annotation

In cross-media relevance model (CMRM), a set of keywords $\{w_1 ... w_n\}$ is assumed to be related to the set of objects represented by visual blobs $\{b_1 ... b_m\}$ in the training image collection, which could be described in a joint distribution as is shown in equation (1).

$$P(w, I) \approx P(w, b_1, ... b_m) = \sum_{J \in T} P(J) P(w, b_1, ..., b_m \mid J) \tag{1}$$

Here they assumed that the observing of w and $b_1, ..., b_m$ are mutually independent once an image J is picked from T, as follows:

$$P(w, b_1, ..., b_m) = \sum_{J \in T} P(J) P(w \mid J) \prod_{i=1}^{m} P(b_i \mid J) \tag{2}$$

Here we found that the annotation process depends more on the correlation between the visual keywords and the training text label for the image set. The correlation between the annotated words is neglected, which brought some obvious semantic conflict. For example, in Fig.1 word snow and beach have low probability to appear together from people's experience.

In [16] Ghoshal et al introduced a hidden markov model to describe the correlation between the annotated words as is shown in equation (3).

$$f(x_1, ..., x_T, s_1, ..., s_T \mid s_0) = \prod_{t=1}^{T} f(x_t \mid s_t) p(s_t \mid s_{t-1}) \tag{3}$$

Image was segmented into regions i_t $(t = 1, ..., T)$ according to regular grid over the image. In eq. (3), $x_t \in IR^d$ represents the color, texture, edges, and shape visual features of region i_t. The probability density function $f(\cdot \mid s_t)$ describes the generation of x_t given s_t, which denotes the state of a selection of word from the annotated word set. $p(s_t \mid s_{t-1})$ computes the transition probability from state s_{t-1} to s_t, $\{s_1 ... s_T\} \in W$. Here visual features $\{x_t\}$ of adjacent image regions are assumed to be conditionally independent given the candidate annotation. This

assumption neglects the correlation between neighboring regions that determines the semantic meaning by the layout of the objects or the regions.

2.3   Markov Chain Model Combining Visual Keywords and the Semantic Annotation

In order to fully utilize the potential correlation among the annotations and image visual characteristics, we treat the image annotation process as a Markov chain. Each keyword in the annotation set and each region from the training image set are treated as a state in Markov chain. Consequently, image automatic annotation can be realized by comparing Markov transition probability between keywords and image visual characteristics.

Images are first segmented and clustered into visual blobs with RPCL+K-Means clustering. Each blob is assigned a unique integer to serve as its identifier. All images in the training set then could be represented as the combination of blobs from this visual vocabulary, as is shown in Fig. 2.
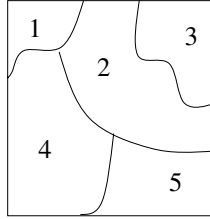
Fig. 2. Visual Keywords Layout

In this paper, our model calculates the Markov transition probability instead of joint probability for image visual features and the corresponding annotations. Given an un-annotated image I , after image segmentation and clustering, image annotation process can be described in the following equations:

$$P(w \mid I) \approx P(w^{t+1} \mid I, W^t) \cdot P(s_{j+1}^t \mid s_j^t) \tag{4}$$

Here $P(s_{j+1}^t \mid s_j^t)$ is transition probability of the visual keywords from state s_j^t to state s_{j+1}^t given an annotation set W^t. To compute the visual keywords transitional probability, we record the visual keywords layout into a visual blob matrix $V_{m,n}$. m and n depends on the number of valid visual regions in each image. $P(s_{j+1}^t \mid s_j^t)$ is defined as the occurrence frequency of available blobs recorded in matrix V .

$$P(w^{t+1} \mid I, W^t) = P(W^t \mid I) \cdot P(w^{t+1} \mid W^t) \tag{5}$$

Where $P(w^{t+1} \mid I, W^t)$ is the Markov transition probability of w^{t+1}. The parameter α and β are the weights of $P(w^t \mid I)$ and $P(w^{t+1} \mid W^t)$ respectively. In the experiment, they are empirically set to be 0.62 and 0.43. In (5), $P(w^{t+1} \mid I)$ denotes the conditional probability of candidate annotation w^{t+1} and reflects the correlation of candidate annotation and image content. $P(w^{t+1} \mid W^t)$ is the conditional probability of candidate annotation w^{t+1} given annotations W^t. It reflects semantic correlation between different keywords in the same image. Considering that not all annotations have correlation in the same image and most have the strong correlation with the image topic or salient object, we at most choose the first two annotations with the highest probability $\{w_1, w_2\}$ as W^t.

When the image content I and the annotations W^t are given, comparing Markov transition probability in (4) is equivalent to compare the joint probability.

$$P(w^{t+1} \mid I, W^t) \propto P(w^{t+1}, I, W^t) \cdot P(s_{j+1}^t \mid s_j^t) \tag{6}$$

$$P(w^{t+1}, I, W^t) = P(W^t \mid I) \cdot P(w^{t+1} \mid W^t) \cdot P(s_{j+1}^t \mid s_j^t) \tag{7}$$

Where $P(w^t, I)$ can be estimated by (2) at the initial annotation state. $P(w^{t+1}, W^t)$ reflects the co-occurrence of annotations, which can be estimated by the TF-IDF method [17] and described as follows:

$$P(w^{t+1}, W^t) = K_c(w^{t+1}, W^t) * \log(\frac{N_T}{n_w}) \tag{8}$$

Here, $K_c(w^{t+1}, W^t)$ denotes the number of co-occurrence of w^{t+1} and W^t. n_w is the number of training image labeled w^{t+1}. N_T reflects the total size of the training set. Therefore we can rewrite (7) as follows:

$$P(w^{t+1}, I, W^t) = \sum_{J \in T} P(J) P(w^{t+1} \mid J) \prod_{i=1}^{m} P(b_i \mid J) \cdot K_c(w^{t+1}, W^t) \cdot \log(\frac{N_T}{n_w}) \cdot P(s_{j+1}^t \mid s_j^t) \tag{9}$$

Thus, we can calculate the probability of each candidate annotation according to (9) and select fixed-length annotations for image I. Compared with the joint probability of CMRM [3], our proposed Markov Chain Model not only considers the potential correlation of different annotations, but also it is related to the image visual content I.

3   Experimental Results

3.1   Dataset

In experiment we use the Corel Image dataset in Jeon et al [3]. The dataset includes 5000 images from 50 Corel Stock Photo CDs. Each CD includes 100 images of the same topic. In segmentation process, we adopts Normalized Cuts [8] similar to CMRM. We use the RPCL [10] competitive learning algorithm to optimize K-means clustering algorithm, which is different from CMRM. Each image was annotated 1-5 keywords. Finally, we got 371 words and 573 blobs in the dataset. We divided the dataset into 3 parts – 4000 training images, 500 evaluation images and 500 test images. The evaluation set is used to find system parameters which include parameters of CMRM, rival learning rate of RPCL and correction coefficients in Markov model. After setting the parameters, the evaluation set was merged into the training set. This corresponds to the training set of 4500 images and the test set of 500 images used by Jeon et al [3].

3.2   Evaluation Measure

For each method, we take the top five words as the final annotation. We evaluate the annotation performance by the average *precision* and *recall* over all the testing images, which can be calculated as follows:

$$precision(w) = \frac{N_{correct}}{N_{automatic}} \tag{10}$$

$$recall(w) = \frac{N_{correct}}{N_{manual}} \tag{41}$$

Where $N_{correct}$ denotes the number of images correctly annotated with w. $N_{automatic}$ reflects the number of images automatically annotated with w. While, N_{manual} denotes the number of images manually annotated with w.

3.3   Results and Evaluation

We compared the annotation performance of the four models: CMRM, the improved CMRM with RPCL, the improved by WordNet for CMRM with RPCL and the Markov model proposed in this paper. In experiment, precision and recall are averaged for both the 49 words with the best performance and the all 260 words that occur in the testing set. Figure 1 and 2 show the performance of our proposed method compared with CMRM, CMRM+RPCL and CMRM+WordNet. To be fair, the final annotations of all the models are fixed to be 5. The results of Figure 3 and 4 clearly show that our proposed method outperforms all the other three. In Figure 4 we provide sample annotations for the two best methods, CMRM+WordNet and our prosed, showing that our proposed method in this paper is considerably more effective.
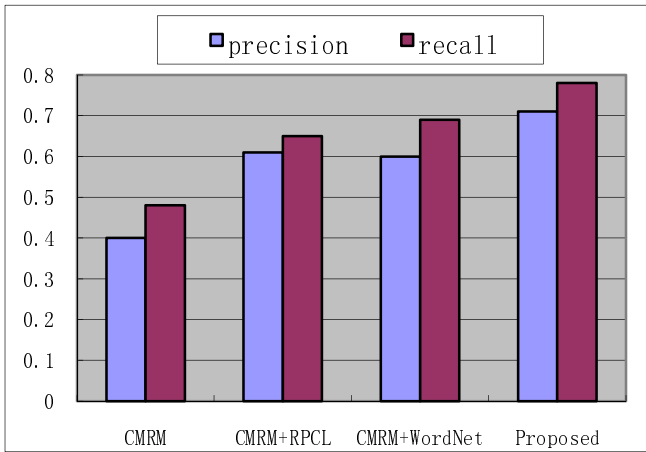
Fig. 3. Performance comparison averaged for the 49 words with the best performance on Corel dataset
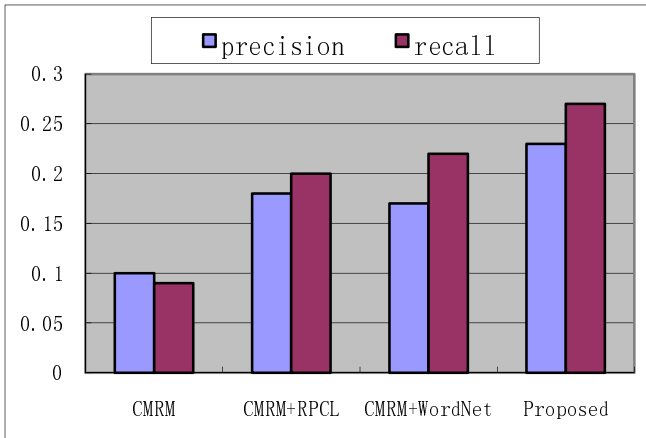

Fig. 4. Performance comparison averaged for all the 260 words on Corel dataset

Images				
CMRM+RPCL	horses, grass, tiger, tree, forest	cars, tracks, street, plane, sky	sea, beach, sand, sky, clouds	sky, tree, clouds, grass, water
Proposed	horses, grass, foals, field, fence	cars, tracks, formula, wall, grass	mountain, snow, sky, clouds, water	buildings, street, cars, sky, clouds

Fig. 5. Sample images and annotation results

4   Conclusions

In this paper, we proposed an image annotation method in considering the spatial relationship of the visual regions and the rationale of the annotated words set. Especially, when deal with the annotation process, we describe the correlations between the visual keywords and the semantic meaning that have annotated. When people recognize the meaning from an image, they would build their understanding based on the semantics they have gotten from the visual appearance. Thus we use the Markov Chain model to simulate the annotation process. Image automatic annotation can be realized by comparing Markov transition probability between states of keywords and image visual characteristics. On the one hand, compared with classical algorithms, the proposed method stops making the assumption that each keyword is independent to each other; instead, they are related to the existing key words and the visual keywords that have been located. On the other hand, not only considering correlation between keywords improves results of image annotation, but our proposed approach also takes image visual content into account. Experimental results show that the proposed algorithm outperforms other existing annotation refinement algorithm and could effectively improve the original annotation results.

References

1. Mori, Y., Takahashi, H., Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words. In: Proceedings of MISRM 1999 First International Workshop on Multimedia Intelligent Storage and Retrieval Management (1999)
2. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
3. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings of SIGIR 2003, Toronto, Canada, July 28-August 1 (2003)
4. Chang, E., Kingshy, G., Sychay, G.: Gang Wu, CBSA: content-based soft annotation for Multimodal image retrieval using Bayes point machines. IEEE Transactions on Circuits and Systems for Video Technology 17, 26–38 (2003)
5. Jia, L., Wang, J.Z.: Automatic linguistic indexing of pictures by a statistical modeling approach. IEEE Transactions on Pattern Analysis and Machine Intelligence. Intell. 25, 1075–1088 (2003)
6. Cusano, C., Ciocea, G., Sehettini, R.: Image annotation using SVM. In: Proceedings of SPIE-The International Society for Optical Engineering, Internet Imaging, pp. 330–338 (2004)
7. Lu, Z., Ip Horace, H.S.: Spatial markov kernels for image categorization and annotation. IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics 41(4), 976–988 (2011)
8. Jiang, G., Liu, X., Shi, Z.Z.: Multi-label image annotation based on neighbor pair correlation chain. In: Perner, P. (ed.) MLDM 2012. LNCS, vol. 7376, pp. 345–354. Springer, Heidelberg (2012)

9. Xu, L., Krzyzak, A., Oja, E.: Rival penalized competitive learning for clustering analysis, RBF net, and curve detection. Neural Networks 4(4), 636–649 (1993)

10. Salton, G., Buckley, C.: Term weighting approaches in automatic text retrieval. Information Processing and Management 24(5), 513–523 (1988)

11. Zhou, Z.-H., Zhang, M.-L., Huang, S.-J., Li, Y.-F.: Multi-instance multi-label learning. Artificial Intelligence 176(1), 2291–2320 (2012)

12. Xia, T., Tao, D., Mei, T., Zhang, Y.: Multiview spectral embedding. IEEE Trans. Syst. Man, Cybern. Part B, Cybern. 40(6), 1438–1446 (2010)

13. Wu, L., Jin, R., Jain, A.K.: Tag completion for image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(3), 716–727 (2013)

14. Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part IV. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)

15. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888–905 (2000)

16. Ghoshal, A., Ircing, P., Khudanpur, S.: Hidden Markov models for automatic annotation and content-based retrieval of images and video. In: SIGIR 2005, Salvador, Brazil, August 15-19, pp. 544–551 (2005)

17. Salton, G., Buckley, C.: Term weighting approaches in automatic text retrieval. Information Processing and Management 24(5), 513–523 (1988)

The Power Spectrum Estimation of Signal Based on Neural Networks

Guang-Min Sun, Wei Liu, Hai-Tao Yan, Fan Zhang, and Hao-Cong Ma

Department of Electronic Engineering, Beijing University of Technology, Beijing, China
gmsun@bjut.edu.cn,
{liuwei1402,haitaoyan}@emails.bjut.edu.cn

Abstract. In this paper, in order to extract the needed information from Wireless Signal Detector, some different frequently–used methods and the related parameter estimation algorithms are compared. Besides, a new method is proposed based on signal spectrum estimation of neutral net which was used to compute signal power spectrum estimation. By simulating the practical data on MATLAB, the factor of neural network can be trained to extract the signal spectrum. The velocity of spectrum estimation is far better than some other algorithm of AR model spectrum estimation method when using proposed method, and spectral estimation quality consistent.

Keywords: spectrum estimation, neural network.

1   Introduction

Along with increased number and variety of the wireless communication and satellite communication technology, the monitoring of real-time wireless signal is becoming more and more urgent. In the actual work environment, even the deterministic signal in transmission process inevitably affected by external interference and noise inside the equipment, so at the receiver observed signal is a random signal. Therefore, how to extract useful information from the signal and express in the right form is the basic content of signal processing [1]. Spectrum estimation is a method that mainly research the random signal in frequency domain, it's mainly purpose is to take into account the random signal interference and carries on the detection and removal, so as to extract the useful signal, its mainly object of study is zero mean Gaussian process[2]. The method of power spectrum estimation has a lot of kinds, usually fall into two categories, one is a classic spectrum estimation [3], also known as linear spectrum estimation; the other kind is a modern spectral estimation, also known as nonlinear spectral estimation. Data were collected through the article [4], we use the different spectral estimation method to analysis data, and puts forward a kind of a signal spectrum estimation algorithm based on neural network, and use actual data to simulate in MATLAB, the result of simulation show that the proposed spectrum estimation algorithm based on neural network can really effective signal spectrum estimation for the subject, and has the very high accuracy.

2   Power Spectrum Estimation

Classic spectrum estimation is based on Fourier transform spectrum estimation method, generally can be divided into two kinds, one kind is BT (Black-Tukey) method, and the other is a period-gram method. Modern spectral estimation [5] is based on the signal model, by the observation data to calculate the parameters of signal model, again by the corresponding formula to calculate the power spectrum of signal and estimate the power spectrum question becomes the observation data to estimate the signal parameters. There are a lot of models, such as AR/MA model and so on, according to different situation to select appropriate model, the quality of the power spectrum estimation has greatly improved than the classic spectrum estimation. In this system, according to previous studies, the algorithm used including BT method, cycle diagram method, Levinson-Durbin algorithm, Burg algorithm, feature vector estimation method, MUSIC estimation method and so on.

In this project, the signal frequency range from 0 Hz to 6G Hz, so it involve a large amount of data, in order to detection signal better, the signal needed truncated and segmented detection method, convert all the bands to baseband signal. The signal of 68.79 MHz as center is used in experiment, as the general monitoring signal usually mixed in several dozens to several hundred MHz signal range, so we use signal 68.79 MHz as an example, we scan and analysis the period of 68.79MHZ.

We select frequency of 68.79 MHz as center frequency to collect experimental data, after convert the signal to the baseband signal use "wireless signal detector" equipment and acquisition signal of 512 points. According to the period-gram method and BT method for the baseband signal spectrum estimation, power spectrum estimation shown in figure 1 and figure 2. The method of period-gram is for the 512 point to direct Fourier transform power spectrum, BT method is used to calculate the autocorrelation function of 512 points and then the power spectrum, as can be seen from the graph, the BT method has low resolution, period-gram method is relatively high, but the amount of calculation is larger. Hence need using modern spectral estimation method for improved.
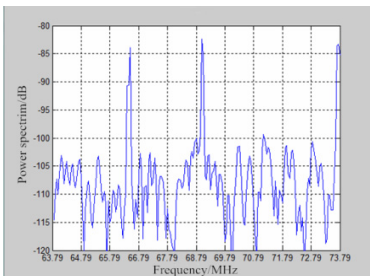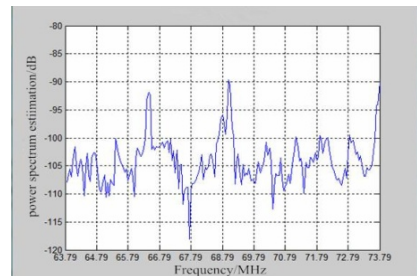

Fig. 1. Period-gram method   **Fig. 2.** BT method

The parameter model method of modern power spectrum estimation mainly includes the L-D recursive method (the Yule-Walker method), the Burg method and modified covariance method. The several methods of simulation data is shown in figure 3 to 5. As the parameter model method can be according to the order of the model to determine the resolution of the experimental data, after repeated experiments, when the model of order is 70, the resolution is higher, can achieve the desired effect.

Because L-D recursive method is based on the Yule - Walker equation to calculate, the calculation is simple, but can only be calculated according to limited observation data, the error will be relatively large. Modified covariance method is improvements to the covariance method, but due to the need to compute the autocorrelation function, calculation relatively large amount. The Burg method needn't to calculate the autocorrelation function, so it involves relatively small amount of calculation, spectrum estimation speed is more faster. The resolution is basically the same as modified covariance method. As can be seen from the graph, the Burg method of spectral line effect is slightly better than the previous two methods, and accurately reflects the signal power of three frequency points.
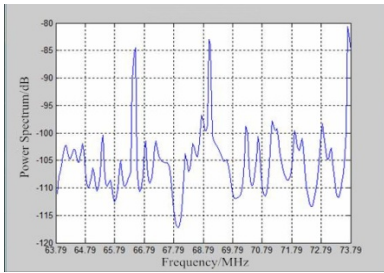
Fig. 3. L - D recursive method   **Fig. 4.** Modified covariance method
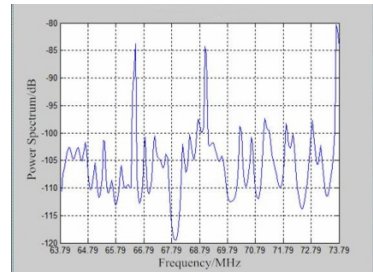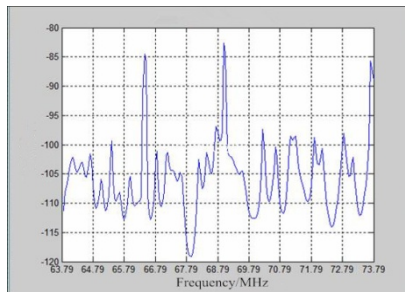

Fig. 5. Burg method

Another method of modern spectrum estimation is the characteristic matrix decomposition method, including the MUSIC algorithm and feature vector method (AV). MUSIC algorithm is suitable for most sine signal parameter estimation method, but because of data in this system obtain from the RF signal, after transformed into

baseband signal is relatively complex, MUSIC algorithm can't reach the expected result. Feature vector method (AV) is a non-parametric method that based on the characteristics of matrix decomposition, which suitable for sine signal mixed with noise. AV method is more applicable to the system data than MUSIC algorithm. However, it unable to distinguish the peaks of signal effectively.

In summary, the signal of this project to carry on the analysis, found that the classical power spectrum estimation resolution is poor, and the feature vector composition of spectrum estimation method for the prediction ability of peaks and troughs are based on a difference, so the modern spectral estimation methods of AR model is suitable for the task of data, can describe the signal power spectrum.

3   The Principle of Neural Network Spectrum Estimation

According to the AR model of the differential equations and the theory of neural network, first, we establish a three layer neural network model, the inputs of the model is the function of the front (n-j), the output as a function of the item n, the order of the model is the number of input layer element [6], the neural network structure as shown in figure 6:
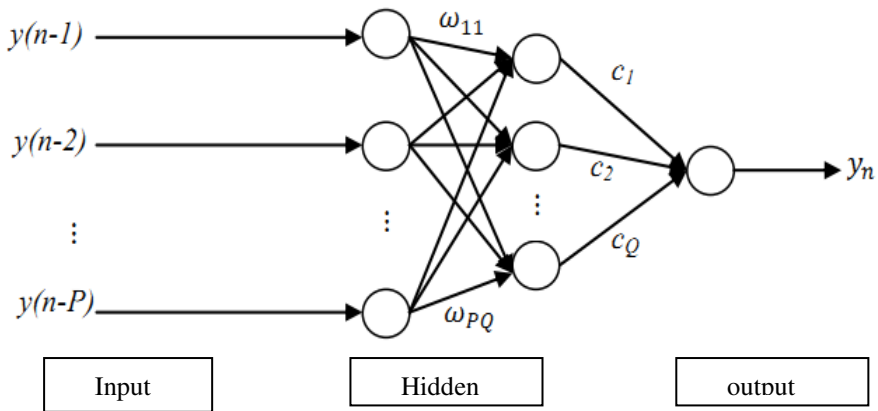

Fig. 6. Neural Network Spectrum Estimation

The output signal AR model can be expressed as:

$$y(n) = \sum_{j=1}^{p} a_j y(n-j) + e(n). \tag{1}$$

Based on neural network is established, we can write the relation between output signal and the input data for the:

$$y(n) = \sum_{i=1}^{Q} c_i \varphi_i(x_i) + e(n).$$ (2)

In (2), Q is hidden layer unit number, c_i is the hidden layer to output layer weights, $\varphi_i(x_i)$ the transfer function of each node in the i layer is $\varphi_i(x_i)$, x_i represent the input layer to the hidden layer weighted sum:

$$x_i = \sum_{j=1}^{p} w_{11} y(n-j)$$ (3)

The transfer function is assumed to be a linear function, namely, the type and (2) and (3) can be written as follows:

$$y(n) = \sum_{j=1}^{p} (\sum_{i=1}^{Q} c_i w_{ji} y(n-j)) + e(n).$$ (4)

Contrast (1) and (4) can be found, neural network hidden input weights and the output weights and added is the parameters of the AR model, as shown in:

$$a_j = \sum_{i=1}^{Q} c_i w_{ji}.$$ (5)

So you can know, as long as the neural network model is established, and the sample for neural network training, after the network stability, can the value of the weighted according to the system and to calculate the parameters of AR model, which based on AR model of formula to calculate the power spectrum of signal.

4   Analysis of the Results

The number of observation data in each bandwidth is 512, so the neural network input and the output is the 512 sets of data. Because the order of neural network estimation method is the input layer and hidden layer weights weight product sum. so in order to guarantee the training data, order number, respectively take 10, 15, 20 in three cases. For example a case of 20 order neural network, when set input is 20, three layer neural network model with single output, which input layer nodes to 20, hidden layer nodes is 5. The results indicate simulation of the neural network is trained 209 times to satisfy the value of the error system 0.000005. Through the comparison of experimental data with training, Levenberg-Marquardt algorithm is the best training function of the network, after 209 iterations are required to achieve error, we found that spectral estimation error in the p=15 order case is small, and the parameters are very small differences in Burg algorithm. The training process of the neural network as shown in Figure 4-11, can be seen, the training effect is more ideal.

In order to verify the accuracy of the algorithm, the experimental data of 55.79MHz(Fig.9-Fig.10) and 65.79MHz(Fig.11-Fig.12) and 65.79 MHz(Fig.13-Fig.14) were taken as the center frequency signal, deal with data the Burg algorithm of AR model spectrum estimation and neural network spectrum estimation to comparison and analysis.
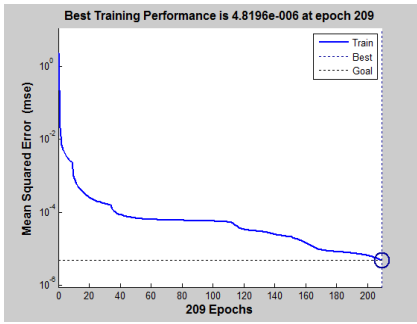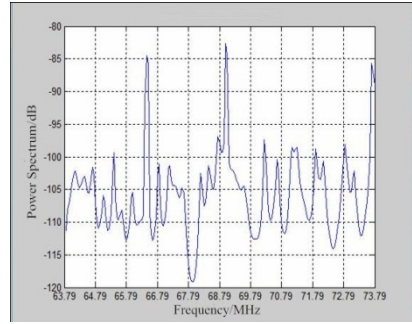

Fig. 7. Neural Network training
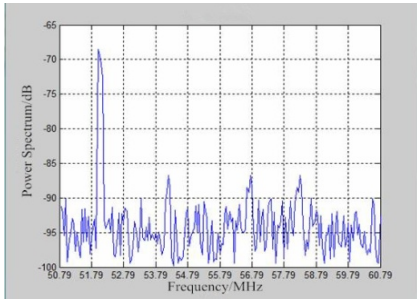

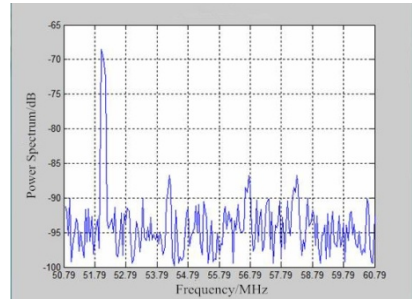**Fig. 8.** Neural Network method

Fig. 9. Burg method

Fig. 10. Neural Network method
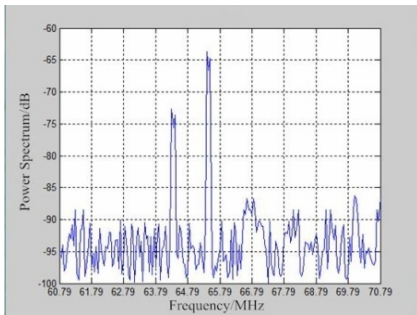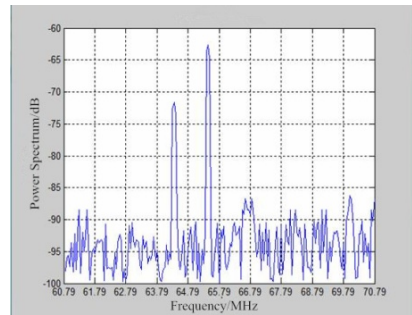

Fig. 11. Burg method
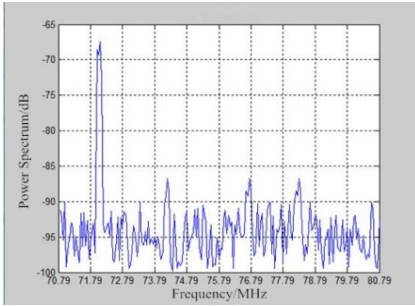

Fig. 12. Neural Network method
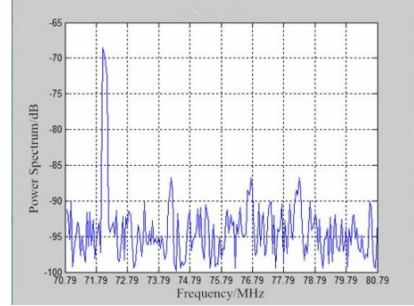
Fig. 13. Burg method

Fig. 14. Neural Network method

The spectrum estimation of the neural network algorithm takes about 0.02 seconds and the Burg algorithm takes about 0.1 seconds, and to be able to distinguish between frequency signal power spectrum estimation obviously. Burg algorithm has the shortcoming of spectral line splitting, but neural network algorithm overcomes this shortcomings.

5   Conclusion

For the experimental data of velocity spectrum estimation is far better than some other algorithm of AR model spectrum estimation method using neural networks, and spectral estimation quality consistent. So it is, in this paper based on spectral estimation spectrum estimation algorithm is better than the AR model of neural network.

References

1. Robert, J.: Modern Statistical Signal Processing. Tinghua University Press, Beijing (1990)
2. Li, P.: The Research Status about Spectrum Estimation Method. China High Technology Enterprise, pp. 76–78 (2008)
3. Yao, W.C.: The Matlab Analysis of Classical Spectral Estimation Method. Journal of Huazhong University of Science and Technology 4(4), 45–47 (2000)
4. Zhang, F., Sun, G.M.: A Novel Ring Buffer Transmission Architecture Based on DM6446 and FPGA. In: International Conference on Control Engineering and Communication Technology (ICCECT), pp. 974–977 (2012)
5. Fu, G.C.: The Application of Matlab in Modern Spectral Estimation. Intelligent Computer and Applications 12, 6–7 (2003)
6. Chen, H.Z., Wu, X.Y.: The Neural Network Method for Adaptive Evaluating Spectrum. Modern Radar (1995)

Algorithm of Laser Spots Recognition Based on "Cat Eye Effect"

Qiang Wu[1], Li-Xiao Yao[1], and Xu-Wen Li[2]

[1] College of Electronic Information and Control Engineering
of Beijing University of Technology
Beijing, China
`yaolixiao1988@126.com`
[2] College of Life Science and Bio-engineering of Beijing University of Technology

Abstract. The paper designs an efficient recognition algorithm to detect laser spots with high Real-time. An improved Otsu is proposed to identify the target and exclude the noise on the condition of atmospheric turbulence. The problem that spot center is not accurate caused by the phenomenon of "supersaturated" is efficiently solved as well. Firstly, the difference the foreground image and the background image is calculated, the morphological filtering is done to segment the target. The elliptic characteristic of the target is used for preliminary identi-fication. Many steps are designed in order to remove False-alarms and im-prove recognition accuracy. The algorithm is applied in the hardware TMS320C6455 system. Extensive experiments show that the algorithm not only ensures the matching accuracy but also improves the time response.

Keywords: Spots Recognition, Improved Otsu Algorithm, Ellipse Detection, Target Selection.

1 Introduction

Laser spots detection, which is based on the spot recognition technology, converts the signal location of the target into the position of the Imaging System [1]. The Algo-rithm of this article proposed is applied in the active laser surveillance system.
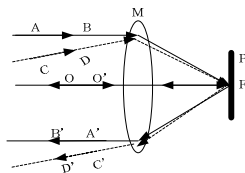

Fig. 1. Principle of "Cat Eye Effect"

Usually, the riflescope is used to aim at the target, so we can make use of "cat eye effect" of optical system to detect the threat targets. The photoelectric equipment has strong reflecting characteristics for the incident laser beams, which is known as "cat

eye effect" [2]. As shown in figure 1, the   A'B' is the outgoing light of AB just like C'D' is the outgoing light of CD, where M is the focusing lens and P is the focal plane. The Outgoing lights have more energy than the diffused lights. In this paper, the algorithm makes use of "cat eye effect" and laser imaging technology to detect the laser spots.

2   Spot Detection Algorithm

The laser beam emitted towards the riflescope is reflected by the riflescope, as shown in figure 2. In the system, the camera 1 receives lights at visible wavelengths to generate background images including visible light spots. While camera 2 receives lights at visible and laser wavelengths to generate Real-time images including visible spots and laser spots. The integration of two cameras can eliminate self-luminous objects and reflected visible light objects to reduce time and improve efficiency. The process of this algorithm is shown in figure 3.
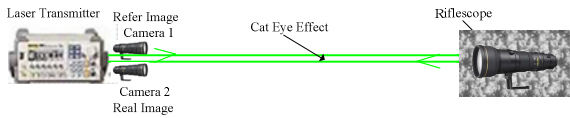

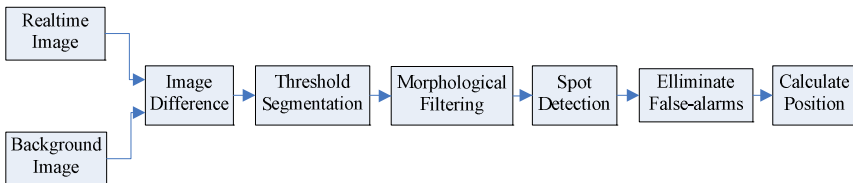**Fig. 2.** System Structure

Fig. 3. Algorithm Flowchart

   As the recognition system is used in complex environments, many spots and highlighting regions generated by untargeted objects present in the image. Impulse and Gauss noises are existed in the image because of the inherent feature and electrical interference at the same time. The difference between the foreground and the background image can remove noise and untargeted spots, and the spots features are enhanced. The background is suppressed and the target is upgraded in a certain extent [3]. Threshold segmentation compresses gray levels and speeds up operations effectively [4-5]. Morphological filtering is useful to erase small isolated points, burrs and small connected regions, making edges smooth and filling small space. The ellipticity of spots is used to detect the suspicious targets [6]. Then the algorithm takes advantage of target characteristics to eliminate False-alarms. At last, the algorithm combines centroid method with structure method to calculate the target position.

2.1   The Otsu Algorithm

The threshold segmentation is a kind of high speed and widely used image segmentation technology [7]. Because the algorithm is to detect spots, not all gray levels are necessary. Selection of appropriate threshold to separate the target from complicated background is the precondition.

L is image grayscale, t is the threshold, $C0 = \{0,1,2......,t\}$, $C1= \{t +1,t+2...... L-1\}$ are two kinds of gray, $W0$ and $W1$ are probability of $C0$ and $C1$, $U0$ and $U1$ are the mean of $C0$ and $C1$ respectively.

$$
\begin{cases}
W_0 = \sum_{i=0}^{t} P_i = W(t) \\
W_1 = \sum_{i=t+1}^{L-1} P_i = 1 - W(t) \\
U_0 = \sum_{i=0}^{t} iP_i \Big/ W_0 = U(t) \Big/ W(t) \\
U_1 = \sum_{i=t+1}^{L-1} iP_i \Big/ W_1 = \left[U(L)-U(t)\right] \Big/ \left[1-W(t)\right]
\end{cases}
\tag{1}
$$

In the equation, $U(t)$ equals to $\sum_{i=0}^{t} ip_i$, $U(L)$ equals to $\sum_{i=0}^{L-1} ip_i$. $D0$ and $D1$ are variances of $C0$ and $C1$. Db means Between-class variance.

$$
\begin{cases}
D_0 = \sum_{i=0}^{t} \left(i-U_0\right)^2 \frac{P_i}{W_0} \\
D_1 = \sum_{i=t+1}^{L-1} \left(i-U_1\right)^2 \frac{P_i}{W_1}
\end{cases}
\tag{2}
$$

$$
D_b = W_0\left[U_0 - U(L)\right]^2 + W_1\left[U_1 - U(L)\right]^2 = W_0 W_1 \left(U_1 - U_0\right)^2
\tag{3}
$$

Calculating the Between-class variance where t in the interval 0 ~ L-1. Corresponding to the maximum variance, t is the optimal threshold for image segmentation [8]. Otsu threshold segmentation algorithm is based on image histogram. It has remarkable eff-ect when target grays and background grays are evenly distributed.

But Otsu algorith-m has two major problems when it is used in the laser spot recognition system.

1. The False Peak Phenomenon

In the figure 4, the grass is lightened up by the laser, leading to the "false peak" phenomenon in the histogram, because the gray level of the grass is between the target and the background. The grass is misjudged as the ratio of the spot region and the grass is too small.
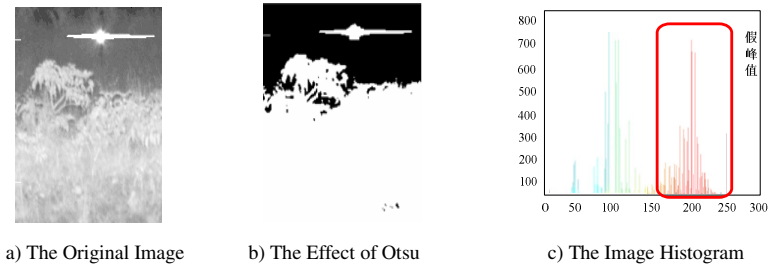
a) The Original Image   b) The Effect of Otsu   c) The Image Histogram

Fig. 4. The False Peak Phenomenon

2. The Speckle Effect

In the figure 5, the speckle effect emerged due to atmospheric turbulence and air suspension. There is no bimodal curve of the target and background in the histogram, leading to that too many untargeted pixels were mistaken for targets.
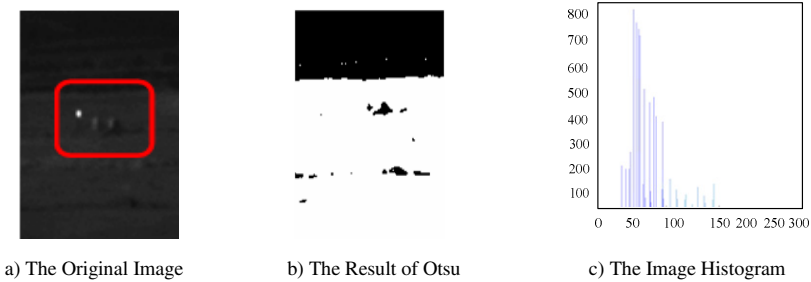

a) The Original Image   b) The Result of Otsu   c) The Image Histogram

Fig. 5. The Speckle Effect

2.2   The Improved Otsu Algorithm

There are two aspects to improve the Otsu algorithm.

3. Threshold Selection Function

The speckle effect causes to peak aliasing. Because the classic Otsu threshold function didn't consider the interclass variance, leading to the difference between the calculated threshold and the theoretical threshold. Especially when under severe conditions of atmospheric motion, atmospheric turbulence will magnify this difference. The speckle phenomenon will make the histogram appear two peaks resulting segmentation difficulty. It needs to modify the threshold selection function, balancing interclass variance and Between-class variance to reduce the impact of the speckle. The function is improved as formula 4 that interclass variance and Between-class variance is the numerator and the denominator respectively. The value of t is the optimal threshold when $G(t)$ is the maximum. It can reduce the influence of the speckle on choice of the optimal threshold in some extent.

$$G(t) = \frac{W_0 W_1 (U_1 - U_0)^2}{W_0 D_0 + W_1 D_1}$$

(4)

4. Iterative Feedback Tuning

Iterative feedback method effectively reduces impact of background and objects ratio imbalance. The steps are as follows:

a) Use Otsu algorithm dividing the image.
b) Calculate minimum bounding rectangle of target region. The rectangle is *Si* and the centre is *Gi*.
c) Divide the rectangle area of *Si* to generate the next segmented region *Si +1*.
d) Stop iterating when the distance between two rectangles of *Gi* and *Gi+1* is smaller than two pixels.
e) Calculate external rectangles of all unconnected regions. As shown in figure 6, assume that there are two rectangular regions and a stopping area *S*. Calculate histogram of three regions to segment the regions.
f) Repeat steps b ~ e until all regions satisfy the convergence condition. The threshold value will be obtained at this time as the optimum segmentation threshold.
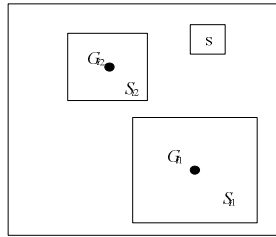

Fig. 6. The Iteration

It can be seen that the principle is that reducing the search scope and increasing the ratio between the target and the background. Images segmentation is shown in the figure 7 using the improved Otsu algorithm.
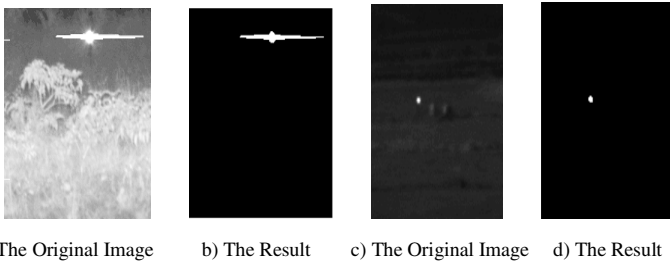

a) The Original Image   b) The Result   c) The Original Image   d) The Result

Fig. 7. The Segmentation of Improved Otsu

2.3   Target Detection

Typically, spots are circular as the riflescope is circular. The spot can be seen as an ellipse in the condition of saturation that exists when the laser energy is too strong or the target is close. As circle is one kind of ellipse, target detection is equivalent to ellipse judgment.
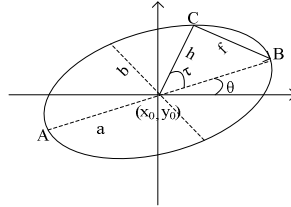

Fig. 8. Elliptic Geometry Flow

The suspected targets are detected by using the ellipse geometric feature. Assuming that vertex coordinates of longer axis are known as *A(x1,y1)* and *B(x2,y2)*, an arbitrary point on the ellipse is *C(x, y)*, the distance between *C* and the ellipse center is *h*, the distance between *C* and *B* is *f*. Other parameters such as the longer axis and the short axis can be calculated by the above known parameters, as shown in figure 8 and equation 5.

$$\begin{cases} (x_0,y_0)=[\dfrac{x_1+x_2}{2},\dfrac{y_1+y_2}{2}] \\[2mm] q=tan[\dfrac{y_1-y_2}{x_1-x_2}] \\[2mm] t=arccos[(a^2+h^2-f^2)/2ah] \\[2mm] a=\dfrac{\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}}{2} \\[2mm] b=\sqrt{\dfrac{a^2h^2\sin^2 t}{a^2-h^2\cos^2 t}} \end{cases} \qquad (5)$$

Rules to determine whether the spot is elliptic are as follows in equation 6 [9]. Firstly, the average *b_avg* of 80% short axis is calculated, then the absolute differences are calculated between short axis and *b_avg,* the number of pixels that the absolute differences less than *T* is counted,   the ratio of all selected pixels and the counted pixels is obtained. If the ratio is greater than the threshold *K*, the spot is elliptic. The *T* and *K* are selected according to the image.

$$\begin{cases} b\_avg = (\sum_{i=0.1M}^{0.9M} b_i)/0.8M \\ d_i = |b_i - b\_avg| \\ ratio = \dfrac{select\_num}{0.8M} \end{cases} \tag{6}$$

2.4 Eliminating the False-Alarms

As the laser power is known and the detection range is certain, spots in the image are of regular shape and high brightness. The algorithm makes use of target spots characteristics to eliminate False-alarms.

There are two steps to eliminate False-alarms. The first step is to eliminate untargeted spots before target detection using spot size. The rules are as follows. The maximums of the circumference, the area and the Length-width ratio are set. Then the corresponding values of the spot are calculated after contour extraction. If any of the values exceed the corresponding maximum, the spot is excluded as a target. This step can exclude large spots and irregular spots, as figure 9 shown. Spot in image a) is large spot that circumference and area exceed the maximum. Spot in image b) is longitudinal tensile that Length-width ratio exceeds the maximum.

Then, the neighborhood method is used to eliminate False-alarms after elliptic judgment. Practically, the target spot is isolated that multiple spots don't exist in the small area. If there are multiple spots in a small area, these spots are untargeted. In the algorithm, the area is a rectangle that radius is of five pixels. In the figure 9, there are many spots in image c) in a small area. They are eliminated to be targets.
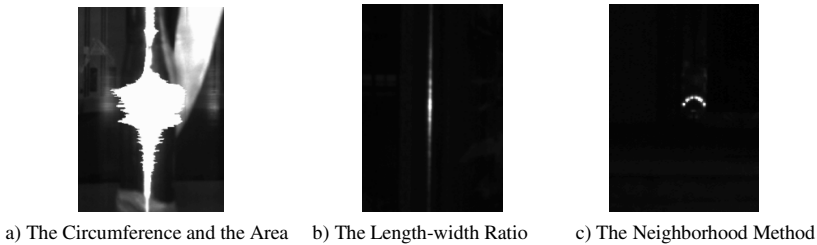

a) The Circumference and the Area   b) The Length-width Ratio   c) The Neighborhood Method

Fig. 9. Elimination of False-alarms

2.5 Calculating Target Position

The centroid is regarded as the position of the spot with regular shape. Equation for calculating centroid is $\bar{x} = \dfrac{\int x \, dA}{A}$, $\bar{y} = \dfrac{\int y \, dA}{A}$ which $(\bar{x}, \bar{y})$ is the coordinate of centroid, (x,y) is the coordinate of boundary point and A is the sum of boundary points. For the spot show in figure 10, the centroid tends to the stretching direction. Analysis of spot characteristics, we can see that the vertical protrusion can be seen as the position. As shown in equation 7.

$$\begin{cases} Center[0] = (Xmin + Xmax) / 2 \\ Center[1] = (Y\_Xmin + Y\_Xmax) / 2 \end{cases} \tag{7}$$

Center[0] and *Center[1]* are center coordinates, *Xmin*, *Xmax* are minimum, maximum row coordinates of boundaries and *Y_Xmin*, *Y_Xmax* are corresponding column coordinates.
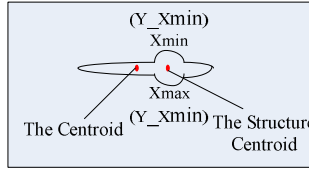

Fig. 10. Calculate Position by Target Shape

The algorithm integrates the centroid method and the structural method to calculate the position. The centroid method is used if the spot is structured while the structural method is used if the spot is stretched.

3   Experiment and Result

In the Real-time system, the row includes 512 pixels and the column includes 200 pixels, the time is 14.5 milliseconds per frame. The algorithm is applied to the TMS320C6455 DSP platform. Firstly, optimize the algorithm, and then test time for each module. The data shows that the algorithm satisfies time performance of the system.
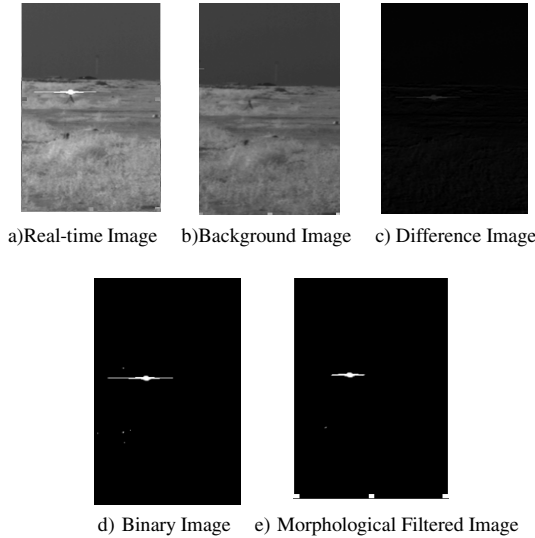

a)Real-time Image   b)Background Image   c) Difference Image

d) Binary Image   e) Morphological Filtered Image

Fig. 11. The Experimental Image

Table 1. Each part of the system running time(single)

Operation module	Cycle	Time(ms)
Image difference	1215681	1.215681
Threshold segmentation	4228387	4.228387
Image corrosion	2.054756	2.054756
Image expansion	1998541	1.998541
Ellipse judgment	415782	0.415782
Total		9.913147ms

4   Conclusion

The paper presents an efficient algorithm to detect riflescope on the basis of "cat eye effect". Firstly, the difference between Real-time image and the background image is done which can eliminate spots generated by reflecting sunlight. Then, the threshold segmentation and the morphological filtering are used to remove noise and smooth the image. At last, the ellipse shape characteristic of the target is adopted to detect the spots. The integration of centroid and structural features is used to calculate the position. Meanwhile, the algorithm eliminates untargeted spots in two steps in order to improve the accuracy efficiently. Experimental results show that the detection accuracy is more than 90%.

References

1. Lin Xibo, A.: Summary of Laser Technology Application. J. Aerospace Shanghai 32(3) (2006)
2. Tong, L.-J., Jiang, X.-Y.: Target Detection Based on Laser Imaging with "Cat Eye Effect". J. Laser & Infrared 9, 982–985 (2009)
3. Kamerman, G.W.: Laser Radar Signals the Civilian Sector. J. Laser Focus World (4), 81–87 (1996)
4. Zhang, Y.-J.: Image Segmentation, vol. 2, 3, pp. 43–60, 149–153. M. Sciences Press, Beijing (2001)
5. Lecocq, C., Deshors, G.: Sight Laser detection modeling. In: Proc. SPIE, vol. 5086, pp. 280–286 (2003)
6. Haibo, L., Jing, S.: Digital Iimage Processing Using Visual C++. pp. 245–251. M. China Machine Press, Beijing
7. Castlemen, K.R.: Digital Image Processing. Publishing House of Electronics Industry (1996)
8. Li, S.: Image Segmentation Methods Based on ROI and Adaptive Otsu Threshold Algorithm. J. Modern Electronics Technique 34(6) (2010)
9. Xie, Y.H., Ji, Q.: A New Efficient Ellipse Detection Method. In: J. Proceedings of the 16th International Conference on Pattern Recognition. IEEE Computer Society, Los Alamitos (2002)

Part VII

Computer Networks and Mobile Computing

A Synchronization-Free Cooperative Transmission Protocol to Avoid Energy-Hole Problem for Underwater Acoustic Networks

Ying-Hong Wang[1], Yao-Te Huang[2], and Kai-Ti Chang[3]

[1] Dept. of Computer Science and Information Engineering,
Tamkang University, Taiwan, R.O.C.
inhon@mail.tku.edu.tw
[2] Distance Education Development Section, Tamkang University, Taiwan, R.O.C.
Jacob@learning.tku.edu.tw
[3] Dept. of Computer Science and Information Engineering,
Tamkang University, Taiwan, R.O.C.
Funkyhome@gmail.com

Abstract. Underwater acoustic sensor network with terrestrial wireless sensor network characteristics distinct, underwater acoustic sensor network uses sound waves to be transmitted, resulting in underwater network with high propagation delay characteristics, if applied directly to land-based routing mechanism water network, will give rise to a number of problems. Therefore, this paper designed for underwater network environment routing protocols, thereby reducing the underwater acoustic sensor network to send packets to the end-to-end delay and for collaborative work on underwater acoustic transmission sensor network, encountered the challenges of high propagation delay characteristics discussed. This paper presents the work last sleep period, coordinated and collaborative collection of neighbor information transfer protocol. Through simulation, this paper found that the proposed routing protocol SCTP, the delay in the point to point, network transmission performance and power consumption of a significant improvement in performance.

Keywords: cooperative transmission, routing, energy hole, underwater sensor networks.

1 Introduction

Due to the booming network technology, wireless network come into existence as another choice other than cable network, providing people more and more multiplication of network type and making wireless network become a hot issue in recent years. In order to meet different requirements, types of wireless specification and technologies have been developed, especially Wireless Sensor Networks (WSNs) [1], which is a successful example of combine sensor and wireless network.

In WSNs, routing mechanism need to achieve separate data transport, power saving and ensure data can be complete sent to sink. The main function of the routing mechanism is to transfer data separately, saving power, and to ensure data can be

completely sent to the sink. However, many researches proposed before [2-4] used wireless sensor node with GPS as an anchor node to localization. However, high power consumption must be considered. When the anchor node power is exhausted, other nodes could not replace the original one, resulting in the interruption of data transmission.
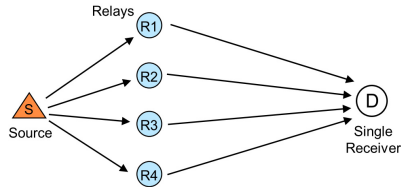

Fig. 1. Cooperative Transmission diagram

Therefore, under the premise of reducing the power consumption, GPS device is not applied in this research. On the other hand, since most the objectives of the many applications of WSNs are in a static environment, the mobility of the wireless is not the first concern and calculation ability is not strong calculation ability is not the necessary. In this regard, power consumption and cost are expected to be minimized. In this paper, we propose a novel routing mechanism "A Routing Mechanism Using Virtual Coordination Anchor Node", which applied Cooperative Transmission. The concept of cooperative transmission is shown as figure 1. With this approach, sensors are randomly deployed and the data is transferred through the specially designed routing mechanism, which is expected to expand the lifetime of WSNs, achieve the efficient transmission; meanwhile, the use of additional devices can be avoided to reduce energy consumption and cost. The mechanism can work with simple computing capability.

2   Related Work

Due to the blooming of microelectronic mechanical systems in recent years, the size of sensor nodes becomes smaller and smaller. How to effectively use the limited battery power to achieve the highest efficiency is an important issue. Since the main factor of power consumption is hardware and data transmission path, to reduce hardware and select a valid path is the focus of this paper research. The current routing algorithms, can be basically divided into four directions: Flooding [6] , Chain [7] , Clustering [8] , and tree [9] . Each route has its own advantages and disadvantages. In this section, we will be in the following detailed descriptions about relevant literature and technical.

2.1   Concurrent Cooperative Transmission (CCT))

CCT[10] use the clustering architecture, it also has the characteristics of the active routing protocol. This method will divide wireless sensor network into many different cluster regions, member node only can communicate each other in the same region. The whole process of long-distance transfer requires the following steps: Firstly, every region will choose a node as cluster head and collect the data collected in the region; then, it will send the date to the base station or data sink.

The choice of cluster head is using random method self-generated and power consumption of cluster head will be higher than other nodes. In order to prevent the death of the previously-selected cluster, a new cluster head will be re-elected after each round of data transfer.

Each round will compare and select a new cluster head, but it will shorten the lifetime of WSNs.

2.2   Time-Division Cooperative Transmission (TDCT)

TDCT[11] is based on grid, it will create virtual grid in network environment. This method will create virtual grid in Advertisement phase as to find relative path. There will be many small grids, knows as "cell", in the grid structure. Data source will be the first Dissemination Nodes, then grids cross node as other Dissemination Nodes. Each grid node will be linked to a Dissemination Node and each Dissemination Node knows its upstream and downstream Dissemination Node. When Dissemination Node needs data, it will send a query message and this message will be transferred by Dissemination Node to data source. Then, data source transfers sensing data to Dissemination Node in reverse, shown as figure 2.
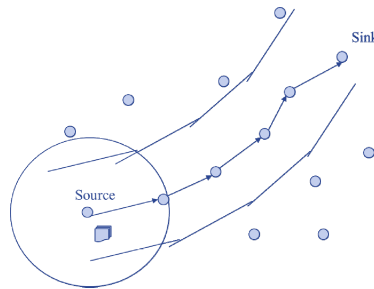

Fig. 2. TDCT architecture diagram

With this method, the best path is not applied; instead, it will select the path which is along the grid crossing node, extending path and need to maintenance periodically. Before the grid is created, it needs to position, and position is cost-consuming. Each time when there is a need to recreate grid or node as Dissemination Node, these nodes will cause the increase of power consumption. Overall, this method can be achieved fairly stable in transmission success rate.

3   A Synchronization-Free Cooperative Transmission Protocol to Avoid Energy-Hole Problem for Underwater Acoustic Networks (STCP)

This chapter will discuss the routing mechanism proposed above to overcome the problem caused by holes of random deployment and unknown boundary in WSNs. The objective of this paper is as following: (1) to create VCS and find virtual anchor

node; (2) to create data transfer path; (3) to find alternative virtual anchor node and path node; (4) to reduce the cost of hardware; (5) to extend the lifetime of WSNs.

3.1   Network Environment Settings

In this paper, the environment of WSNs, lots of sensors are randomly deployed in the range wherever you want to monitor. When sensors are deployed, they can't move anymore and the computing, memory and battery are limited. Each sensor nodes has wireless communication, which can transmit detected data via wireless. There have a sink node, which has more power and better computing capability then other nodes. Normal nodes will send detected data to sink node, and sink node will use collectively. Each node has its own Node_ID to be recognized.

Each node has its transmission range. In our architecture, R is the maximum transmission range; m is the unit of meter, shown in figure 4(a). The transmission range is cut into four equal portions, each portion has its send priority, show in figure 4(b), it will send the packet or data according to their priorities via directional antenna to the direction.
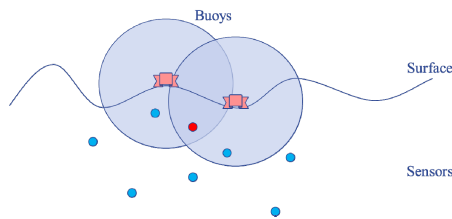

Fig. 3. Underwater positioning

3.2   Basic Principle

When wireless sensor nodes settle down, sink node starts to create virtual coordinate, then it will find the four directions of the entire network environment as the virtual anchor nodes. Therefore, in this phase, sink node will set itself to act as a start node, and create a special ANS packet. The packet has many different fields supply sensor nodes to record coordinate status.

After sink node creates ANS packet, the Self_Coor will be set as (0,0) and according to the priority of sensor node to send beacon toward the priority1 direction, show in figure 4.

When other sensor nodes receive beacon, response ack message back to sender. The Sender will choose the nearest node as the next ANS packet, and according to the definition of priority1 X-1, take out (0,0) from Self_Coor then execute X-1 to get (-1,0). Then, set Next_Coor of ANS packet as (-1,0) then send the fastest reply node.

When sensor nodes received ANS packet, it will copy Self_Coor from ANS packet to Pre_Coor as the coordinate of previous node, and copy Next_Coor to Self_Coor as the current coordinate to save to memory of node. We do this to ensure the link
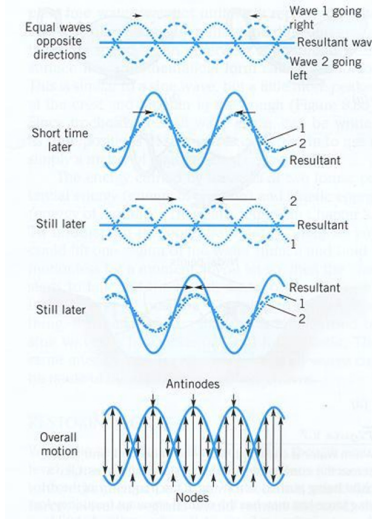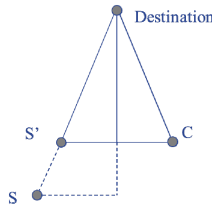
Fig. 4. Underwater wave adding diagram

Fig. 5. System model of S,C co-op transmit to destination

between each sensor node is correct. Then, we take Self_Coor respectively compare with the Left_Node, the Up_Node, the Right_Node and the Down_Node to see if that should Self_Coor replace anchor nodes of four corner or not. Checking logic is shown in figure 5.

If there is no ack packet response when sensor node send beacon toward the priority1 direction, it means priority1 direction doesn't have any sensor nodes, or the status of sensor nodes is bad. Sensor node A doesn't sense any node from priority1 direction; likewise, sensor node A doesn't receive ack packet.

At this moment, the status of node A will be set as Warning, then it will seek the node in sequence according to its priority until it find the next node, it will find node D and node E.

Then it will seek according to the priority. The purpose of this mechanism is by giving the priority1 direction a chance again owning to the random distribution of sensor node, it is possible that the current direction temporary doesn't have any nodes and it will revise its direction by switching to the next priority, finding the sensor node again.

If sensor node in priority1 direction can't find any sensor (include node status is Bad) and status is W, then priority will change to the next set priority, and clean the value of status, re-find nodes from priority1 direction.

Shown as Figure 9, the switching sequence of the Priority group is from top-left to bottom-right. The purpose of the design is to seek the boundary node in the network counterclockwise, which is the way that Priority1 is deigned. The rest of the Priorities are designed according to the extra chance of the previous Priority1 given by the sequence of the previous group of Priority.

3.3   STCP Algorithm

The process of establishing virtual coordinates and searching of virtual anchors will last repeatedly until the ANS packet is received repeatedly by wireless sensor node. Those wireless sensor node receiving ANS packet repeatedly will produce Final Coor according to the packet it has received, shown as Figure 2. The repeatedly-received packet will be abandoned after the production of Final Coor; meanwhile, this packet will be sent to the previous wireless sensor node according to the record in the previous coordinate recorded in ANS packet in the wireless sensor node. When the wireless sensor node receive the last two Final Coors, the packet transfer will come to a halt and the stage of the establishing of virtual coordinate and the searching of virtual anchor is then accomplished.

After the stage of the establishment of the virtual coordinate in wireless sensor network, all the sensor on the borders of wireless sensor network are given virtual coordinates; meanwhile, each sensor can obtain information of the virtual anchors and their coordinates of the sensors through ANS packets, shown in figure 6.
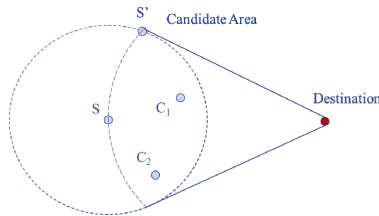

Fig. 6. Candidate area of cooperative selection.

In this stage, the establishment of routing begins with the four virtual anchors in Final Coor. If the virtual nodes are Left_Node or Right_Node, Pre_Coor and Next_Coor in ANS packet of the exact virtual node will be applied to search for the Y coordinate equal/lager than that of Self_Coor until it reaches the virtual node when Y coordinate is 0. If the virtual node is Up_Node or Down_Node, Pre_Coor and Next_Coor in the ANS packet of the exact virtual node will be applied to find the X coordinate which is smaller or equal to that of Self_Coor until it reaches the virtual node when X coordinate is 0.

Next, Path_Link will be sent to Sink Locatin (0,0) according to the Self_Coor in the ANS packet of the virtual anchor, shown in Table 3. When the wireless sensor node receives the packet, it will record its Node_ID and Anchor_Coor. Then, it will

search for Pre_ID and Next_ID. Finally, the packet will be sent to the next wireless sensor node until the packet is transferred to the Sink Node.

Node_ID：The only identification number of the node

Anchor_Coor：The coordinate of the source anchor

Next_ID：The identification of next node

Pre_ID：The identification of previous node

Figure 10 illustrates the virtual coordination anchor nodes in the wireless sensor network after the establishment of virtual anchors and Sink Node routes. A route begins with four virtual anchor nodes link to each Sink Note separately and serves as substitute to send the data detected on the wireless sensor nodes which are on the boundary of non-wireless sensor network.
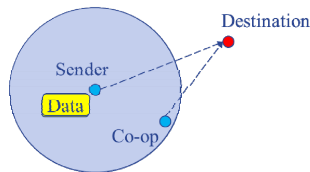

Fig. 7. Scene 1 of best hop-count alhorithm

When an event is detected by a wireless sensor node, the node will send Request Packet to its One-Hop neighbors. While the wireless sensor node receives the packet, it will send a Response Packet back to the one detected the event. If the Request Packet is sent to the wireless sensor node that is on the border of wireless sensor network or on the route of Sink Node, a mark will be attached on the Response Packet so as to raise its priority. If the Response Packet is marked, the date will be sent to the wireless sensor node so as to send data back to Sink Node as soon as possible; if not, data will be sent back to the first wireless sensor node which responds to the Request Packet. The node which an event takes place will record the sent Node_ID and the request of the Node_ID will be neglected in the future. At the same time, the node will take itself as the wireless sensor node with an event detected and repeat the process of the stage.

3.4   To Solve Energy Hole Problem

When data transfer is taking place within wireless sensor nodes, a Request Packet will be sent to a receiving wireless sensor node. If no response is received by the wireless sensor node which sending the Request Packet, a hole is recognized. In this situation, date should be transferred back to the mechanism immediately to ensure that it can be sent to the destination.

When a hole is found, the wireless sensor node which sends the Request Packet will regard itself neglected to any Response Packet and sent Rescue (see Table 4) to its One Hop. In this scenario, only Pre_Node will receive packet since no neighboring nodes exist other than the sending node. Therefore, when Pre_Node receives a Rescue, it will choose a new transfer node to send Request Packet.
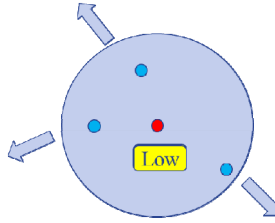
Fig. 8. Avoid energy hole diagram.

Avoid energy hole algorithm:

Input: Set of randomly deployed sensor nodes
Output: Link/Mapping of nodes-to-nodes
foreach Task *m* **do**
foreach sensor *i* **do**
Calculate Two-hop neightbors using PT(*i,j*);

if *LoE(i,j)* $\neq$ *LoE(j,i)* **then**
i = *θj*;
j = *θi*;
while *non-candidate sensor remaining* **do**
foreach *sensor m* **do**
Existing energy through planar trigonometry, PT(*i,m*);
if *PT(i,j) ≤ PT(i,m)* **then**
m = *θj*;
Mapping(sensor *i* on sensor *j*) ;
for *sensor field* **do**
Low in the sensor field;
if *Packet from pervious node* **then**
Set Lm, L1 ;
Continue until Low, Lm
for *sensors inside the low* **do**
new link established.

4   Comparison and Analysis of Simulation

In this chapter, "A Routing Mechanism Using Virtual Coordination Anchor Node" proposed above will be programmed and stimulated in C++ language and compared with Flooding, TTDD, LEACH, and Anchor Node Based Sink Location Dissemination Scheme for Geographic Routing. Analysis of the results will be provided in the following sections to prove the mechanism proposed in this paper is more sufficient than the approaches above.

4.1   Simulation

Setting and parameters of the environment of wireless sensor network used in the stimulation:

- Range of the network: 100M X 100M
- Number of wireless sensor: 50
- Range of wireless transfer: 10M
- Transmission cycle: 10 seconds
- The deployment of wireless sensor: random

4.2   Analysis and Comparison of The Results

In the beginning, time is taken as the parameter to see its impact to the packet control by the wireless sensor node. The stimulating time ranges from 20 seconds to 200 seconds, shown as Figure 9.

From the result, it is clear to find that the number of the controlled packet is much higher than that in other routing, mainly owing to reason that each wireless sensor node needs to participate in Flooding. For TTDD, a grid is established by the wireless sensor nodes which serve as the sources. The grid need to be maintained regularly, therefore the number of its controlled packet is lower than that in Flooding. In each round of LEACH, cluster head need to be compared and re-elected, resulting in the number of controlled packet lower than the previous two kinds of mechanisms. In "Anchor Node Based Sink Location Dissemination Scheme for Geographic Routing", a sink location announcement packet and a sink location query packet are sent along two respectively on the sink node location query route and sink location announcement route, so two routes are required in the query. In the first stage, Flooding is applied in order to make each wireless sensor node obtain anchor data, which results in the higher number of controlled packet. Once the anchors are failed, the sink location query route or the sink location announcement route need to be recreated. Since there is no repair mechanism, an enormous amount of controlled packet is seen when the anchor break down.

In the paper, ANS packet produced by the Sink Nodes will be transmitted along the border nodes and send back the anchor data along the same route. In this way, Flooding is not applied; instead, a substitute mechanism is available and substitute node is expected to be found once the route node is dead or breaking down. There is no need to rebuild the route, reducing the number of controlled packet compared with other routing mechanisms.

In addition to the comparison of the number of controlled packets, the relation between energy consumption and time of the wireless sensor node is stimulated and analyzed. Figure 10 shows the diagram of time and energy consumption of wireless sensor nodes. The stimulating time ranges from 20 seconds to 200 seconds.

From the diagram, it is revealed that both of the approach in this paper and that in "Anchor Node Based Sink Location Dissemination Scheme for Geographic Routing" share higher level of energy consumption in the beginning compared to other three routing since both approaches require more energy in initialization which involves the coordinates of a Pre_Node, a Next_Node and the coordinates of four border nodes.

However, GPS devices is not necessary in this paper, therefore it causes lesser energy consumption than that in "Anchor Node Based Sink Location Dissemination Scheme for Geographic Routing". Hence, the approach in the paper is expected to increase the lifetime of the system, extend life cycle, provide more processing time for wireless sensor network, and increase the efficiency of the mechanism.
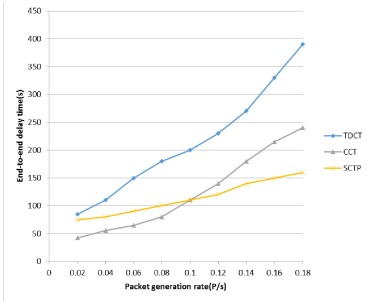

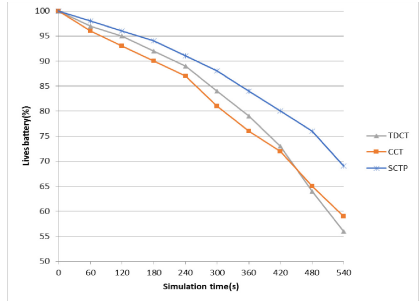**Fig. 9.** Diagram of the relation of end-to-end delay time and packet generation rate

Fig. 10. Diagram of relation of time and lives battery

Finally, success rates of detecting and sending data to the sink node per unit time is also compared. Figure 15 shows the comparison of success rates of each mechanism in detecting event and sending data to Sink Node. The success rate of data transfer in Flooding decreases with the most rapid speed since the collision may cause failure in data transmission to Sink Node and the loss of data. As for TTDD, the grid is sufficient to separate the risk; however, only four routes reach the Sink Node. Furthermore, every time an event is detected, a new grid is needed to build, which might be contributable in the increase of the possibility of collision when the grid is under construction together with the increase of the number of event. For LEACH, since it need cluster head to transmit data back to Sink Node; however, serious collision might be likely to occur while an enormous amount of data flood into the cluster head. When different types of event take place in "Anchor Node Based Sink Location Dissemination Scheme for Geographic Routing", it will recreate sink location query route and sink location announcement route to separate date. In this paper, instead, the node will search for the nearest wireless sensor node to send data when an event is detected. Meanwhile, the virtual anchors and the route nodes can send data back even though no data has been sent to them. Therefore, the loss of data is unlikely to happen.

From three sets of stimulating experiment, it is apparent that "A Routing Mechanism Using Virtual Coordination Anchor Node" in this paper is a sufficient routing which can save energy, expand lifetime of the wireless sensor network, and reach a high delivering rate. Finally, the purpose of the research is achieved.

5   Conclusion and Future Research

In the mechanism proposed in this paper, GPS is not applied in the process so as to save the consumption of energy and cost. With the application of virtual coordination anchor, the function of the mechanism is expected to be sufficient. From the result of

the stimulation, it is obvious that the number of controlled packets as well as energy consumption is reduced under the routing protocol even the density of transferring packets is high; and the success rate of date transmission is raised as the survival rate of each node rises, which further ensures the function of the network and expand the lifetime of wireless sensor network. Meanwhile, power-saving mechanism is anticipated to be involved in the future research to lower the energy consumption in the initializing level of the network. In addition, cluster routing or tree routing protocol are wished to join the working of the network so the efficiency of the Internet could be maximized.

References

1. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. Computer Networks 38, 393–422 (2002)
2. Liu, P., Zhang, X., Tian, S., Zhao, Z., Sun, P.: A Novel Virtual Anchor Node-Based Localization Algorithm for Wireless Sensor Networks. In: Proceedings of the Sixth International Conference on Networking, p. 9 (2007)
3. Fucai, Y., Younghwan, C., Sang-Ha, K., Euisin, L.: Anchor Node Based Sink Location Dissemination Scheme for Geographic Routing. In: Proceedings of IEEE Vehicular Technology Conference, VTC 2008, pp. 2451–2455 (May 2008)
4. Fucai, Y., Younghwan, C., Soochang, P., Euisin, L., Ye, T., Minsuk, J., Sang-Ha, K.: Anchor Node Based Virtual Modeling of Holes in Wireless Sensor Networks. In: Proceedings of IEEE International Conference on Communications, ICC 2008, pp. 3120–3124 (May 2008)
5. Shankarananda, B.M., Saxena, A.: Energy efficient localized routing algorithm for Wireless Sensor Networks. In: Proceedings of 3rd International Conference on Electronics Computer Technology, ICECT 2011, pp. 72–75 (April 2011)
6. Intanagonwiwat, C., Govindan, R., Estrin, D.: Directed diffusion: a scalable and robust communication paradigm for sensor networks. In: Proceedings of the 6th Annual International Conference on Mobile Computing and Networking, Boston, Massachusetts, United States, pp. 56–67 (2000)
7. Dressler, F., Awad, A., Gerla, M.: Inter-Domain Routing and Data Replication in Virtual Coordinate Based Networks. In: Proceedings of IEEE International Conference on Communications, ICC 2010, pp. 1–5 (May 2010)
8. Yaling, T., Yongbing, Z.: Hierarchical flow balancing protocol for data aggregation in wireless sensor networks. In: Proceedings of Computing, Communications and Applications Conference 2012, January 11-13, pp. 7–12 (2012)
9. Kuo, T.-W., Tsai, M.-J.: On the construction of data aggregation tree with minimum energy cost in wireless sensor networks: NP-completeness and approximation algorithms. In: Proceedings of 31st Annual IEEE International Conference on Computer Communications, INFOCOM 2012, pp. 2591–2595 (March 2012)
10. Heinzelman, W.B., Chandrakasan, A.P., Balakrishnan, H.: An application-specific protocol architecture for wireless microsensor networks. IEEE Transactions on Wireless Communications 1, 660–670 (2002)
11. Luo, H., Ye, F., Cheng, J., Lu, S., Zhang, L.: TTDD: two-tier data dissemination in large-scale wireless sensor networks. Wirel. Netw. 11, 161–175 (2005)

The Study of Using Game Theory for Live Migration Prediction over Cloud Computing

Yen-Liang Chen, Yao-Chiang Yang, and Wei-Tsong Lee

Department of Electrical Engineering, Tamkang University,
Taiwan, R.O.C.
{brian566716,ychiangy}@yahoo.com.tw, wtlee@mail.tku.edu.tw

Abstract. Cloud computing was a technology in recent years which had been concerned. More and more network applications provided client a more convenient experience for use on the cloud computing service. Cloud computing is using virtualization technology. It can not only improve the performance on the server, but including a characteristic dynamic data assignment. Additionally, any server with fault, over loading or maintenance…etc. which need to be stopped, the user is not aware that the service has interrupted, that is because the technology of live migration will quickly backup the remaining data from original server to another server. The study [1] used Gilbert-Elliot model has a capability to predict the probability on dirty page until performing 10 times iteration. From this study, using Game Theory model of reducing predicted number effectively can early determine whether to go the stop-and-copy phase. That saves the time on live migration.

Keywords: Game Theory, Gilbert-Elliot, Live Migration, Pre-Copy.

1 Introduction

Cloud computing can integrate the operation or idle servers into a single system which provides the necessary terminal equipment service at any time, any place.
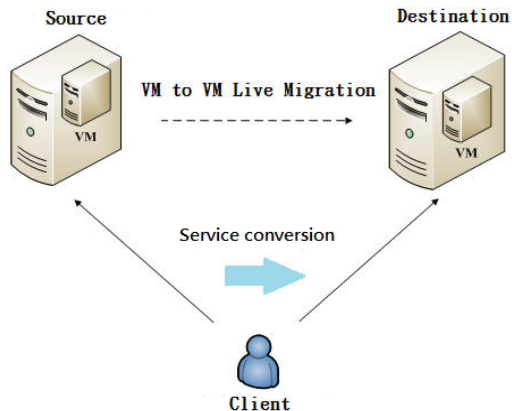

Fig. 1. Live migration [1]

Through the Internet, cloud computing provides resources and information of hardware and software according to the user's demand.

The operating systems of virtual machine are built on the existing server system. Users can create multiple systems through virtualization technology on single server. As figure 1, the operating system of virtual machine migrate the data from source server to destination and to ensure all programs can be working correctly on the destination server [2].

2   Background and Related Work

2.1   Overview of Live Migration

The data migration on virtual machines can be divided into non-live migration and live migration which using the two indicators, total migration time and downtime ([3], [4]), to demonstrate the performance of migration technology. Total migration time is the duration when the migration begins from source server until all transmission completed and stopped. Downtime is the duration when the service on source server suspended until the service resumes on destination server. An excellent migration tool can reduce total migration time and downtime, and minimize the impact of service on destination server.

Live Migration is called dynamic migration or online migration as well. Live migration can guarantee the service is always existed when virtual machine migrates to another server. The live migration process can be divided into three phases [4]: Push Phase, Stop-and-copy Phase and Pull Phase, shown as figure 2, the data migration can be done by these three phases. Pre-Copy is a combination of the first two, and Post-Copy is a combination of the latter two ([1], [5]).
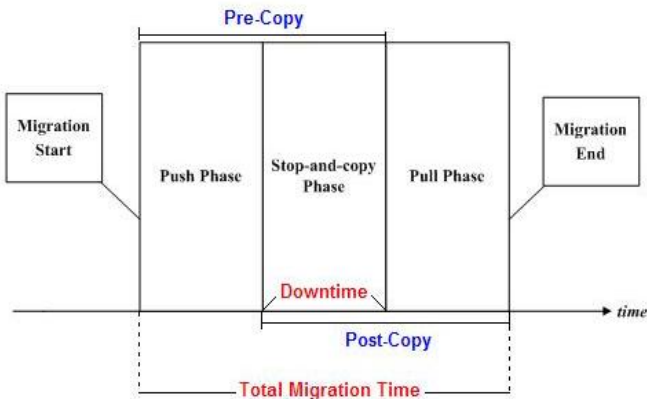

Fig. 2. The phase of live migration

Push phase means when the operating system begins the migration, the data on virtual machine will migrate from source server to destination. Data will be sent to destination of virtual machine within the operating time. In order to maintain the data consistency in both sides, data need to be re-sent if the data has been modified during the operation. The stop-and-copy phase means the source side of virtual machine stops working. The destination of virtual machine will resume operating system until data is fully transferred. The pull phase means when the information on the destination of virtual machine has not been migrated. The page fault occurs. The missing information will be re-sent from source of virtual machine [1].

2.2   The Prediction of Gilbert-Elliot Model

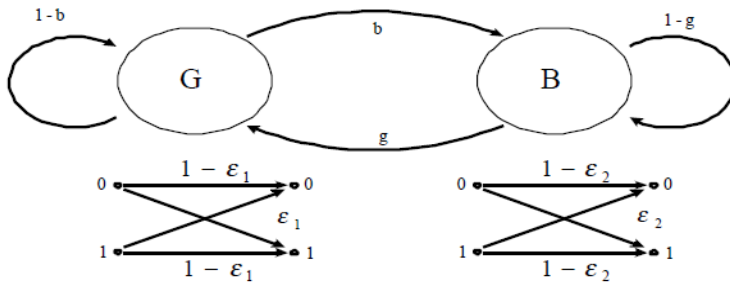The prediction of Gilbert-Elliot Model is shown as below.

Fig. 3. The gilbert-elliot model [1]

Figure 3 shows two states of Markov Chain. The state "G" indicates the good condition in communication, and the corresponding error bit is ε_1. The state "B" indicates the bad condition in communication, and the corresponding error bit is ε_2. The $\varepsilon_1 << \varepsilon_2$.

The probability of transition states are:

$P_{GG} = P(G \mid G) = P[(i+1)^{th} \text{ state of G} \mid i^{th} \text{ state of G}] = 1\text{-}b$

$P_{BG} = P(B \mid G) = P[(i+1)^{th} \text{ state of B} \mid i^{th} \text{ state of G}] = g$

$P_{GB} = P(G \mid B) = P[(i+1)^{th} \text{ state of G} \mid i^{th} \text{ state of B}] = b$

$P_{BB} = P(B \mid B) = P[(i+1)^{th} \text{ state of B} \mid i^{th} \text{ state of B}] = 1\text{-}g$

Therefore, the state matrix is (1) [1].

$$S = \begin{bmatrix} P_{GG} & P_{GB} \\ P_{BG} & P_{BB} \end{bmatrix} = \begin{bmatrix} 1-b & b \\ g & 1-g \end{bmatrix} \tag{1}$$

The time proportional of Markov model in state "G" or state "B" will be close to a constant in the long time. This constant is steady state probability. P(G) is good condition of steady state probability, P(B) is bad condition of steady state probability. It can get (2) and (3) [1].

$$P(G) = \frac{g}{b+g} \tag{2}$$

$$P(B) = \frac{b}{b+g} \tag{3}$$

According to Gilbert-Elliot model can know each bit error probability is (4) [1].

$$Ave\_BER = P(G) \times \varepsilon_1 + P(B) \times \varepsilon_2 \quad = \quad \frac{g}{b+g} \times \varepsilon_1 + \frac{b}{b+g} \times \varepsilon_2 \tag{4}$$

2.3   Overview of Game Theory

Game Theory was created by John von Neumann in 1944. It was only as a part of applied mathematics in the beginning. Today, the Game Theory has been used in all applied sciences. Game theory usually analyzes interactive strategy selection for rational person. Game Theory considers the rule and relative payoff of participants to predict the best strategy and relative reactivity ([6], [7], [8]). The major research of Game theory is to study the best strategy from each party's conflict of interests and to ensure the best interests of them.

Assuming the payoff matrix of two players R and C is

$$M = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \qquad\qquad ; a + d \neq b + c$$

R and C mixed strategies are

$$P = \begin{bmatrix} p & 1-p \end{bmatrix} \tag{5}$$

$$Q = \begin{bmatrix} q \\ 1-q \end{bmatrix} \tag{6}$$

The expected value of R is

$$E(P, Q) = PMQ = \begin{bmatrix} p & 1-p \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} q \\ 1-q \end{bmatrix} \tag{7}$$

3   Problem Statement and Proposed Method

3.1   The Problem Statement

The study [5] proposed utilizing Markov model to observe variation of dirty page to predict the probability of memory modification.

The study [1] quoted Gilbert-Elliot model to consider two states of Markov model. The variation of dirty page will be more accurate than [5], however study [1] has to assemble the probability of memory modification until 10th iteration, If the dirty page

less or none in previous iterations, the live migration still can't go into stop-and-copy phase to complete the job earlier. That will waste more time for estimation and being unnecessary.

3.2   The Prediction of Game Theory

The study that using Game Theory model to compare the probability which got from last time with this time to predict the next probability of dirty page. The advantage is that we just need to calculate the probability of last and this time. The bellows are our parameters definition on Game Theory:

P: The strategy for average probability of dirty page on the Nth iteration.
M: The strategy matrix of Nth and $(N-1)^{th}$ iterations.
Q: The strategy for average probability of dirty page on the $(N-1)^{th}$ iteration.
E: The strategy for expected value in this game. It is also a predicted number for the average probability of dirty page on $(N+1)^{th}$ iteration.

The table 1 below is the payoff matrix of Pre-Copy estimation. Through this table we can estimate whether we should perform the Pre-Copy or not, the first column shows if want to perform Pre-Copy at the Nth iteration. The first row shows $(N-1)^{th}$ iteration.

Table 1. The payoff matrix of $(N+1)^{th}$ iteration

N \ N-1	Pre-Copy	Non Pre-Copy
Pre-Copy	a	b
Non Pre-Copy	c	d

We set number "a"=1 for explaining both N^{th} and $(N-1)^{th}$ determine to perform Pre-Copy, that is because we don't need to concern about the case for both their determinations are the same. Based on this thinking, we can set number "d"=1 as well.

Regarding number "b", that will be calculated as the average probability of dirty page on the Nth iteration, so the $(N-1)^{th}$ iteration will be also through the calculation to get the average probability of dirty page for the number "c". Both number "a" and "c" are calculated by the Gilbert-Elliot model. Finally, the table will come out an expected value $E_{(N+1)^{th}}$ for the best strategy, where $E_{(N+1)^{th}}$ =PMQ. The expected value $E_{(N+1)^{th}}$ will be a predicted number for the $(N+1)^{th}$ iteration.

Using the same way, we could get the expected value as $E_{N^{th}}$. In (8) below, the threshold value is determined as we should go Pre-Copy or go into stop-and-copy

phase immediately. Putting expected value with $E_{(N+1)}{}^{th}$ and $E_N{}^{th}$ from payoff matrix will obtain the threshold value.

$$Threshold\ value = \mid E_{(N+1)}{}^{th} - E_N{}^{th} \mid \qquad (8)$$

We set the threshold value as 10%, and start to monitor from U1 which means $E_{(N+1)}{}^{th}$ is from U2. If the threshold value is large than 10%, which means the server could go Pre-Copy, however if the threshold value is less than 10%, which means the dirty page variation is much lower. That means there is no any contribution for doing the Pre-Copy, the server should go into the stop-and-copy phase, rather than spend a lot of time on Pre-Copy. It will save totally migration time.

Table 2. The list of dirty page

Update / Page Number	U0	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
1	1	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	0	0	0
3	1	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0
5	1	1	0	0	0	0	0	0	0	0	0
6	1	1	0	0	0	0	0	0	0	0	0
7	1	1	1	1	1	0	0	0	0	0	0
8	1	1	0	0	0	1	1	1	0	0	0
9	1	1	0	0	0	0	0	0	1	1	1
10	1	1	0	0	0	0	0	0	0	0	0

Table 3. The average probability of dirty page on U1's update

Page Number	U1 Average Prob.	P_{GG}	P_{BG}	P_{GB}	P_{BB}
1	10%	0%	0%	100%	0%
2	10%	0%	0%	100%	0%
3	10%	0%	0%	100%	0%
4	10%	0%	0%	100%	0%
5	0%	0%	0%	0%	100%
6	0%	0%	0%	0%	100%
7	0%	0%	0%	0%	100%
8	0%	0%	0%	0%	100%
9	0%	0%	0%	0%	100%
10	0%	0%	0%	0%	100%
Average	10%	0%	0%	40%	60%

Following this way to calculate the average probability of dirty page by   Gilbert-Elliot model for U0 to U10 is shown as table 4.

Table 4. The average probability for U0 to U10

Update	U0	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
G-E.	90%	10%	10%	0%	0%	18%	0%	0%	18%	0%	0%

Then using Game Theory model to predict the average probability of dirty page by every iterations. The table 5 is just a payoff matrix for predicting average probability on U2. The number "c" is a number for average probability on U1 as 10%. The number "b" is a number for the average probability on U0 as 90%.

Table 5. The payoff matrix of 2nd iteration

U1 \ U0	Pre-Copy	Non Pre-Copy
Pre-Copy	1	b (90%)
Non Pre-Copy	c (10%)	1

If competitor's strategy is first column, the expected payoff is

$$E_{1st} = (1)p + (0.1)(1-p) = 0.9p + 0.1 \tag{9}$$

If competitor's strategy is second column, the expected payoff is

$$E_{1st} = (0.9)p + (1)(1-p) = -0.1p + 1 \tag{10}$$

To solve (9) and (10)

$$0.9p + 0.1 = -0.1p + 1$$
$$p = 0.9$$

So the best strategy for U1 is

$$P = \begin{bmatrix} 0.9 & 0.1 \end{bmatrix}$$

The best strategy for U0 is

$$Q = \begin{bmatrix} q \\ 1-q \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.9 \end{bmatrix}$$

At last, the expected value in this game is calculated as follows:

$$E_{2nd} = PMQ = \begin{bmatrix} 0.9 & 0.1 \end{bmatrix} \begin{bmatrix} 1 & 0.9 \\ 0.1 & 1 \end{bmatrix} \begin{bmatrix} 0.1 \\ 0.9 \end{bmatrix} = 0.91 = 91\%$$

4   Results and Analysis

Following the way in previous section, we used Game Theory model to simulate the situation with locality.

4.1   The Predicted Probability of Locality State

Table 6. The locality state

Update Page Number	U0	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
1	1	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	0	0	0	1	1	1
3	1	1	0	0	0	0	0	0	0	0	0
4	1	1	1	0	0	0	0	0	0	0	0
5	1	1	1	1	1	1	0	0	0	0	0
6	1	0	0	0	0	0	0	0	1	1	1
7	1	1	1	1	1	1	1	0	0	0	0
8	1	1	1	1	0	0	0	0	1	1	1
9	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	0	1	1	1

Using Gilbert-Elliot model to calculate the rate of dirty page for all iterations is shown as the table 7.

Table 7. The rate calculation of locality state

Update	U0	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
G-E	90%	10%	10%	10%	10%	0%	10%	10%	90%	0%	0%

Then to calculate the predicted value by Game Theory model for the $(N+1)^{th}$ iteration from U1 to U10 as table 8.

Table 8. The rate calculation by game theory model.

Update	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10
Game Theory	91%	91%	55%	55%	55%	53%	53%	55%	91%	91%

The simulation result is showed as figure 4. We could predict the 3rd iteration is 55% through Game Theory model, and the 4th predicted iteration is 55% as well. The threshold value is 55%- 55% = 0%. So we can consider going into the stop-and-copy phase at 3rd iteration.
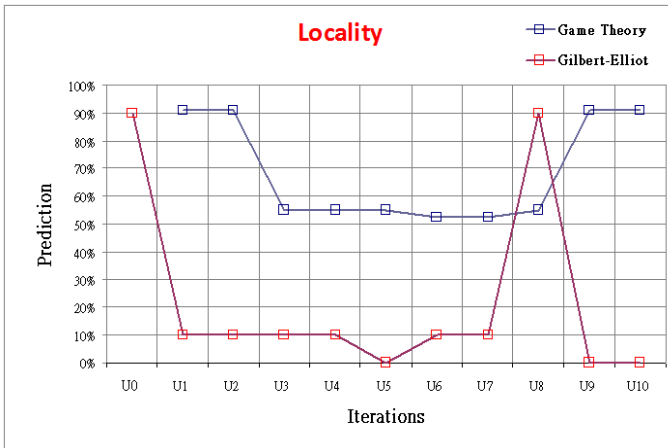
Fig. 4. The simulation result of locality state

5   Conclusions

Based on the simulation results, using Game Theory model to predict average probability of dirty page for all iterations will be better than only using Gilbert-Elliot model. When the threshold value is less than 10%, the Pre-Copy will not reduce the total migration time. It will be better to go into the stop-and-copy phase immediately. The unnecessary iteration can be saved.

References

1. Lin, Y.-W.: The Improvement of Predicted Probability with Cloud Computing Memory Live Migration, Department of Electrical Engineering, Tamkang University, Taiwan, R.O.C. (2013)
2. Chen, C.-C.: The effect of Cache-Improving of Live Migration for Virtual Machines. Department of Electrical Engineering, Tamkang University, Taiwan, R.O.C. (2013)
3. Anderson, E., Hobbs, M., Keeton, K., Spence, S., Uysal, M., Veitch, A.: Hippodrome: running circles around storage administration. In: Proceedings of the First Usenix Conference on File and Storage Technologies, FAST (2002)
4. Milojicic, D., Douglis, F., Paindaveine, Y., Wheeler, R., Zhou, S.: Process migration. ACM Computing Surveys 32(3), 241299 (2000)
5. Huang, J.-S.: Related Dirty Memory Prediction Mechanism for Live Migration in Cloud Computing Systems. Department of Electrical Engineering, Tamkang University (2012)
6. Wang, T.-M.: A Novel P2P Sharing Mechanism based on Social Network and Game Theory., Department of Electrical Engineering, Tamkang University, Taiwan, R.O.C. (2012)
7. Feng, H., Zhang, S., Liu, C., Yan, J., Zhang, M.: P2P Incentive Model On Evolutionary Game Theory. In: 4th International Conference on Wireless Communications, Networking and Mobile Computing, pp. 1–4 (2008)
8. Ouyang, J.-C., Wang, Y.-B., Hu, X.-H., Lin, Y.-P.: An Incentive Mechanism Using Game Theory for P2P Networks. In: NSWCTC 2009, International Conference on Networks Security, Wireless Communications and Trusted Computing, vol. 2, pp. 715–718 (2009)

A Greedy-Based Supply Partner Selection
for Side-by-Side 3D Streaming on P2P Environment

Yu-Jhih Wang[1], Hsin-Hung Cho[2], Wei-Chung Liu[3],
Han-Chieh Chao[1,3], and Timothy K. Shih[2]

[1] Department of Electrical Engineering, National Dong Hwa University, Hualien, Taiwan,
Republic of China
[2] Department of Computer Science and Information Engineering, National Central University,
Taoyuan, Taiwan, Republic of China
[2] Department of Computer Science and Information Engineering, National Ilan University,
Ilan, Taiwan, Republic of China
{hen19891015,wjll792,timothykshih}@gmail.com,
hsin-hung@ieee.org, hcc@niu.edu.tw

Abstract. In the world of today, the mobile device technologies are advancing with each passing day, the humans use them to watch videos and deal business anywhere. However, to support the huge stream so that the bandwidth will face enormous challenges. Because such technology is easy to lead the network congestion and longer waiting time. In this paper, we use the front camera of a mobile device to track user's viewing angle then to calculate the current the 3D stream of current needs then find the most suitable peer for source supplying. This way is consider to the maximum available upload bandwidth to chosen supply partners to relieve the network additional burden.

Keywords: P2P, 3D video, gaze tracking.

1   Introduction

In recent years, the related applications of P2P (Peer-to-Peer) system have has a lot of success stories [9]. Since each P2P user can simultaneously serve as P2P provider and receiver, so it also reduce the data acquisition time thus reduce network congestion and reduce maintenance cost from ISP (Internet Service Provider). Such observation shows that the P2P system has more advantages than the Client-Server architecture that include the more improvement of scalability, low latency and low cost.

At present, more and more humans are start to follow with interest the 3D technology since it presents a more realistic scenes than the 2D video thus allowing users to get the better viewing quality. The 3D somatosensory game, 3D TV and 3D smart phones are indeed reflecting the demand of the 3D technologies.

Streaming technology is very suitable for the P2P applications that use distributed transmission to relieve the 3D stream traffic. The humans can share information by the mobile devices with each other thus decreasing the downloading burden from stream server to the users thus reducing the playback delay and interruption. However, for

mobile devices, the video streaming occupied bandwidth is still a heavy burden especially 3D streaming, which will take up twice the theoretical bandwidth, computing resources.

We investigated a lot of literatures about 3D streaming. We found that the considerations of current mechanism are only simple focus on the factors influence of network environment. We observed that the humans using mobile devices to watch video in different habits, and the use of mobile devices to watch video usually with fixed viewing angle since not everyone likes the same posture. Therefore, we boldly hypothesize that since some particular sub-streams may unable enter the eyes successfully, and such sub-streams are superfluous we can delete safely. If this assumption is correct, we should be used the routing approach to reduce the network traffic. In this paper, we must detect the user's perspective viz the front camera of mobile devices. Then devices will calculate the viewing angle and distance between the screen and user by the results from the front camera of mobile device. The second needed calculation is the quantity of right flow and left flow which are users be able to see. Finally, we proposed GSR (Greedy-based Streaming Reduction) algorithm to discover the routing path and find out the suitable 3D streaming source node.

This paper is organized as follows: In section 2, we introduce relevant research on eye tracking and P2P 3D streaming. In section 3, we use linear programming to define our traffic reduction problem. In section 4, we introduce our proposed GSR algorithm in detail. Section 5 is the simulation results. Finally we will summarize the contribution of this paper, and simply distribute the future works.

2   Related Work

Some eye tracking studies have been proposed [1][2][3]. W. Waizenegger et al. are focus on the real-time business video conferencing. They use the front camera to tracks the viewer's eyes perspective that be presented eyes perspective correctly in the case of continuous movement. It can makes perfect connectivity between participants in 3D video conferencing [1]. In order to estimate gaze direction, Reale, M. et al. mapping the two iris centers and iris contour of the eyes 3D sphere so that the human eye can be clearly estimate the line of sight. . In this work, the authors prove that using a mobile device's front camera for face recognition and gaze estimation are feasible [2]. We will use above works to find the central point of eyes and the pupil position, then we can found the user's viewing angle thus to calculate the quantity of needed left and right stream respectively. Our goal is that we want to reduce the computational complexity of mobile devices to reduce the network burden.

Many recent studies have been dedicated to P2P 3D streaming to find available supply node [4][5][6][7]. And they have proved their methods able to increase the transmission performance. They also using AoI (Area of Interest) to avoid unnecessary message transmission based on the position message detection mechanism, that the main aims is to provide the needed information more quickly thus to minimize network bandwidth consumption and avoid the flooding problem of network congestion. However, due to the AoI's restrictions are easy to make a substantial reduction in the scope of the search to limit the scope of exploration. From the above related works, we can learned that the biggest challenge in finding suppliers

node of P2P 3D streaming is how to find out the correct contents on mobile devices and any nodes with the largest available upload bandwidth. Therefore, we will take a selection of the maximum bandwidth to achieve their goals.

3   Problem Definition

Our environment is built on the unstructured P2P networks. Any mobile devices of the network are regard as the peers, and they have the ability to store and forward the video stream. We assume the 3D video files are distributed to some peers randomly. Fig. 1 shows the overall architecture diagram. We can see that when the user starts the selected 3D video content via mobile devices, the mobile device's front camera will start to detect the human's face and eye model in a limited wide-angle range. Next, the periodic task is to find the user's perspective, then to consider the relationship between the user's angle and distance.
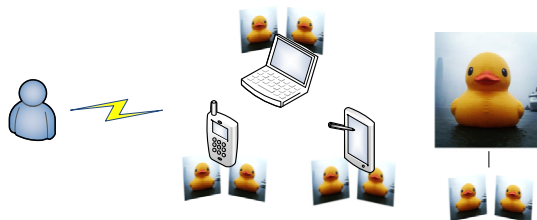

Fig. 1. Architecture of P2P environment

If calculated out the correct message of the angle of views, then to convert them into the quantity of needed left and right stream. Finally, we want to find out the suitable supply nodes and complete a thrifty routing protocol. In simple terms, the original routing problem will divided into two sub-problems according to the different needed quantity of both eyes stream. Furthermore, two sub-problems may affects with each other. Hence, the problem will become more difficulty due to the interaction of different requirements. Our ultimate goal is to maximize available bandwidth size of the overall network as shown in Table 1.

At here, we define the available upload bandwidth of the candidate nodes and relay nodes to find out the source nodes that meets be consumed the bandwidth of left flow and right flow. Next, we subject to n_i less than N_p. Namely, we wanted to find out the candidate nodes with video content from all nodes of P2P network. And the quantity of streaming of candidate nodes will greater than 0% that represent each candidate nodes has different quantity of streaming in the same video content.

4   Proposed Mechanisms

4.1   Views Detection

The front camera usually set above the screen. Then we detect the position of the user's face and eyes. If the eyes cannot be found, then redetect human face again. Next, we

estimate the iris position to find the user's line of sight thus be able to calculate the current viewing angle and viewing distance. If views angle calculation in error, it will return to iris detection, and calculate the correct viewing angle again. The results will convert into the quantity of needed stream then put this result to GSR. Then we can find out most suitable source node for forwarding. Details as Fig.2.

Table 1. Symbol defnitions and ILP

N_p: Number of peer in P2P system, $\forall p=1,2,\ldots,P$.

n_i: Number of peer has the required stream content,
    $\forall i=1,2,\ldots,j$.

CL: List of peer has the required stream content.

V_i^r: Right stream flow of n_i.

V_i^l: Left stream flow of n_i.

B_i: Available upload bandwidth of n_i.

B_R: Available upload bandwidth of V_R needed.

B_L: Available upload bandwidth of V_L needed.

B_y: Available upload bandwidth of relay node.

Maximize

$$\sum_{i=1}^{j} B_i$$

Subject to

$$n_i \leq N_p \ , \forall i$$
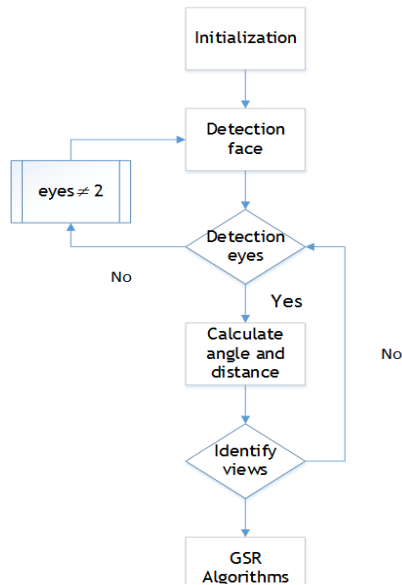$$V_i^r, V_i^l > 0 \ , \forall i$$

Fig. 2. Detection views process

When the eyeball and the pupil center is calculated, we can use line of sight to find out the user viewing angle and start to calculate current views of the user. We consider that the viewing angle and distance of the user will affect the quality of viewing. Therefore, we must to analyze the relationship between distance and angle of the viewing:

$$h = \frac{w}{tan\theta} \text{ ,} \tag{1}$$

The above formula can be proved in Fig. 3. If the viewing distance h is increased from h_1 to h_2 then viewing angle θ is reduced from θ_1 to θ_2, In other words, the shorter distance between the user and screen, front camera to detect the greater viewing angle, otherwise, the longer distance the screen, front camera to detect smaller the viewing angle.
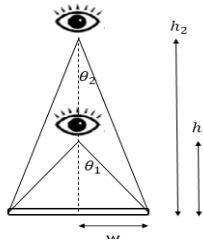

Fig. 3. Users viewing screen of mobile devices

When a user with too long or too short visible range of the distance and deviation left or deviation right to watch 3D videos, that will causing blurred pictures because the deviation focal length of the eyes, it must first determine the distance before the viewing angle calculation and the exclude some case such as that it cannot get the left and right eyes sight and the clear image of showing circumstances: 1) Some one or both eyes pupil over wide angle range of front camera in the image area. 2) Focal length of the user's line of sight beyond the screen. When these situations occur will cause the camera cannot correctly determine the line of sight and video stream of required to rendering, so these cases should be excluded from the above two conditions before detect.
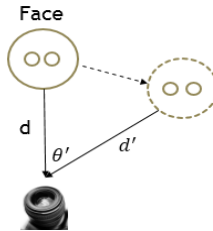

Fig. 4. Detection views process

In the angle detection work, we will divide the plane of the screen into 90 degrees for left and right, respectively. It represents that the users possible viewing area. The plus sign (+) denote that the user of mobile devices to watch on the left side. The negative sign (-) denote users of mobile devices to watch on the right side. If users with optimum angle equal to 0 ($\theta = 0$) to watch the screen as shown in Fig. 4, we can get the initial distance d [8] that will be able to obtain the trigonometric formulas to estimate the angle and distance:

$$d' = \frac{d}{cos\theta'} , \tag{2}$$

θ' and d' are respectively represents that the viewing angle and distance of the user movement. In order to express the interaction of every metric, we make a normalization of viewing angle and distance:

$$\alpha \frac{|\theta_{max}-\theta'|}{\theta_{max}} - \beta \frac{|d'-D|}{D} , \tag{3}$$

The α and β denote the angle ratio and the distance ratio of impact. They may not same, because their setting may different according to the different devices type. If $\frac{|\theta_{max}-\theta'|}{\theta_{max}} = 1$ denote the user watching the screen based on optimum angle that can experience the highest quality of 3D video. Otherwise, $\frac{|\theta_{max}-\theta'|}{\theta_{max}} = 0$ represents the user has the worst angle ($\theta = 90°$) so that the screen cannot be watched. And $\frac{|d'-D|}{D}$ approaches 1 represents that the resolution of screen is blurring. Otherwise, if this value approaches 0 denote that the video quality is better. Note that since the user's viewing distance is unlimited, so user may in a overlong distance which is greater than the optimum distance D. According to this consideration, this case will let this value as 1.

We will analyze to the proportion of left and right eye views thus to obtain percentage of binocular streams:

$$V_L = X \left[\alpha \frac{|\theta_{max}-\theta'_l|}{\theta_{max}} - \beta \frac{|d'_l-D|}{D}\right] \times 100\% , \tag{4}$$

$$V_R = X \left[\alpha \frac{|\theta_{max}-\theta'_r|}{\theta_{max}} - \beta \frac{|d'_r-D|}{D}\right] \times 100\% , \tag{5}$$

$X = \frac{\theta'}{|\theta'|}$ denotes the user watching with different directions, The following equation shows that a boundary with mean that $\theta=90°$ divided into left and right line of sight:

$$X = \begin{cases} < 0, left\ line\ of\ sigt \\ > 0, right\ line\ of\ sigt \end{cases} , \tag{6}$$

V_L and V_R represent left and right eyes viewing quality of the user. We firstly consider the percentage of the current views, and then calculate the proportion of distance between eyes and the screen. The vary in distance will lead views quality decreasing, so it is necessary to balance the relationship of a variable, then subtracting the proportion of perspective and distance to get the left and right ideal streams which are

limited to the user's current views quality. According to this method, we can reduce the traffic load to avoid the additional bandwidth consumption.

4.2   Greedy-Based Streaming Reduction

When we get V_L and V_R, we can obtain the percentage of stream of video content which is user needed to watch . Putting these results to GSR (Greedy-based Streaming Reduction) algorithm then it starts to look for a suitable supply source. We pick out a candidate node which has suitable steam flow. Next, we consider to its available upload bandwidth. And we will select one which has the higher the priority.

Table 2 is GSR algorithm process, the requester first sends the flooding to obtain coordinates of N_p, and set TTL (Time To Live) less than 5 that to avoid the messages are passed between nodes unlimitedly. When a node received the query packet, it will check the own local directly or cache directly whether have available content. If nodes have, return packet and start to download. If nodes haven't, they will continue to look for available videos content from neighbor list.

Table 2. GSR algorithm

Input:V_L and V_R
Output: supply partner of maximize bandwidth
1.   flooding neighbor and return
2.   find all neighbors of N_p
3.   broadcast V_L, V_R
4.   if $V_i^l < V_L$ or $V_i^r < V_R$ then
5.   return line 2 else
6.   if $B_i < B_L$ and $B_i < B_R$ then
7.   return line 2 else
8.   join to CL
9.   end if
10.   end if
11.   if $y \geq 1$
12.   if $B_y < B_L$ or $B_y < B_R$
13.   choose neighbor with B_y bigger then B_L or B_R
14.   end if
15.   choose $\max(b_i^l)$,$\max(b_i^r)$ to be supply partner
16.   end if

Before the source node transmitting streaming, it first checking to the original path through the relay node, and the available upload bandwidth is lower than the source node. If the bandwidth of relay node is lower than bandwidth of source node then source node will return the streams. There is no sufficient upload bandwidth to support such stream, and it may cause transmission rate reducing even increasing the probability of packet loss. So the source node will finding out the relay nodes with available upload bandwidth greater than needed bandwidth of V_L or V_R from neighbor list of relay node, then to continue transmitting the stream so that such principle of greedy recursive execution will get maximum residual bandwidth for the overall network.

5   Simulation Results

We design a simulation scenario to highlight the advantages of viewing angle detection with GSR algorithm. We set 50 nodes randomly distributed in P2P network. Each node randomly assigned 3D stream data, as well as the available upload bandwidth from 0.5Mbps to 1.5Mbps. Our method GSR compared with original method and observed the quantity of streaming, the quality of the user's viewing, and the network load balancing.
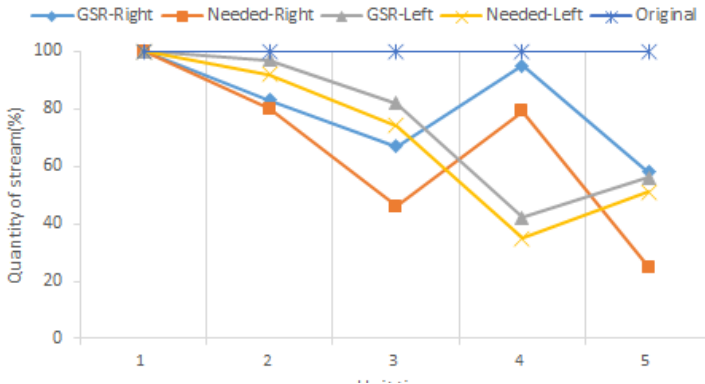

Fig. 5. Users viewing quantity of stream

Fig. 5 shows that the users viewing quantity of left stream and right stream in a case of user to see a video stream. 'Needed' represents that the low bound of required resource. We can see the line of resource consumption of Original is the highest. Because Original not used front camera to analyzing current quantity of required stream to rendering, so always maintain 100% of flow with maximum resource. GSR can determines Need with minimum resource and found the source node meet quantity of required stream.

Fig. 6 represents that the size of left stream and right stream of supplier partners have available upload bandwidth. Original's available upload bandwidth almost less than GSR, because it was not selected the largest of available upload bandwidth for the overall network nodes. When the supply partners be selected and start streaming, if the current available upload bandwidth of supply partners is not sufficient to support the bandwidth of needed   streaming, it would seriously affect the quality of the user's viewing. Although the available upload bandwidth of supply partner by GSR less than the Original for right stream in unit time 1, but the impact is not great for the overall stream.

When the available upload bandwidth of GSR is larger the Original's will increase upload bandwidth of remaining after the stream transmitting. It will significantly reduce the probability of the node be paralyzed and increase the efficiency of overall network.
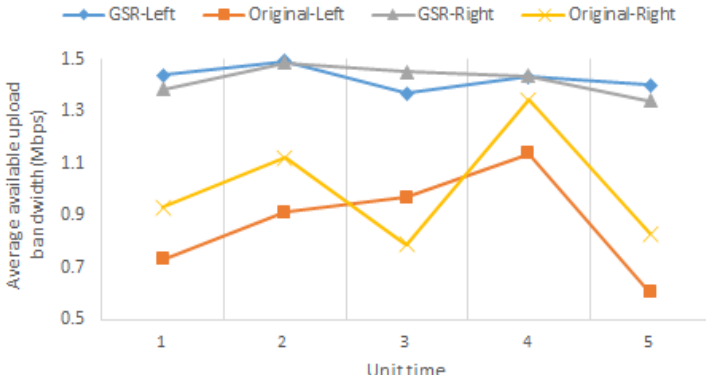
Fig. 6. Average available upload bandwidth of supply partners

Fig. 7 shows that the loading of P2P overall network. We can see the loading of GSR lower than Original because the quantity of needed streams is reduced that leads to the source nodes bandwidth consumption relative to be reduced, and improve overall network load.
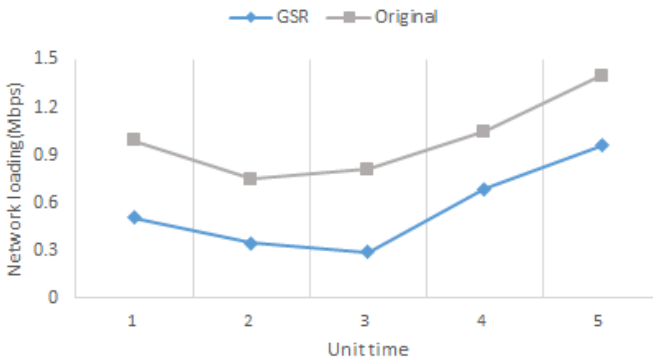

Fig. 7. Network Load Balance

6   Conclusions and Future Works

Due to the side-by-side 3D may encounter the blurred images cause from the extreme viewing angle and extreme distance. So we use sight technology of the front camera tracking line   that can calculated the viewing angle and distance to changed quantity of streaming   without affecting the quality of the user viewing. In this way, we can reduce unnecessary traffic thus to avoid additional network burden. The simulation results illustrate that GSR can reduce quantity of streaming effectively, because the worst case of this study is just a subset of the old method. As long as the user viewing angle is not perpendicular to screen and the quantity of streaming achieves conditions of the quality of user needs so that the redundant stream will be deleted.

In our future works, we will express a more realistic scenario thus design a more objective mechanism with multi-objective direction. For example, both of the mobility, interference, electricity and security can effects the entire situation.   Therefore, to solve these issues are the important efforts.

References

1. Waizenegger, W., Atzpadin, N., Schreer, O., Feldmann, I., Eisert, P.: Model based 3D gaze estimation for provision of virtual eye contact. In: 2012 19th IEEE International Conference on Image Processing (ICIP), September 30-October 3, pp. 1973–1976 (2012)
2. Reale, M., Hung, T., Yin, L.: Viewing direction estimation based on 3D eyeball construction for HRI. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 13-18, pp. 24–31 (2010)
3. Hu, P., Shen, G., Li, L., Lu, D.: ViRi: view it right. In: 2013 ACM MobiSys 2013, Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services, pp. 277–290 (2013)
4. Hu, S.-Y., Huang, T.-H., Chang, S.-C., Sung, W.-L., Jiang, J.-R., Chen, B.-Y.: FLoD: A Framework for Peer-to-Peer 3D Streaming. In: IEEE INFOCOM 2008, The 27th Conference on Computer Communications, pp. 13–18 (April 2008)
5. Maamar, H.R., Alonso, G.R., Boukerche, A., Petriu, E.: Energy-aware analysis for supplying partner selection in mobile P2P 3D streaming. In: 2011 IEEE Symposium on Computers and Communications (ISCC), June 28-July 1, pp. 74–79 (2011)
6. Maamar, H.R., Pazzi, R.W., Boukerche, A., Petriu, E.: A supplying partner strategy for mobile networks-based 3D streaming - proof of concept. In: 2010 IEEE International Symposium on Parallel & Distributed Processing, Workshops and Phd Forum (IPDPSW), April 19-23, pp. 1–6 (2010)
7. Chien, C.-H., Hu, S.-Y., Jiang, J.-R.: Bandwidth-aware Peer-to-Peer 3D streaming. In: 2009 8th Annual Workshop on Network and Systems Support for Games (NetGames), November 23-24, pp. 1–6 (2009)
8. Rahman, K.A., Hossain, M.S., Bhuiyan, M.A., Zhang, T., Hasanuzzaman, M., Ueno, H.: Person to Camera Distance Measurement Based on Eye-Distance. In: 2009 Third International Conference on Multimedia and Ubiquitous Engineering, June 4-6, pp. 137–141 (2009)
9. Lai, C.-F., Huang, Y.-M., Chao, H.-C.: DLNA-Based Multimedia Sharing System for OSGI Framework With Extension to P2P Network. IEEE Systems Journal 4(2), 262–270 (2010)

A SIP/IMS Platform for Internet of Things in WLAN-3GPP Integration Networks

Whai-En Chen and Shih-Yuan Cheng

Department of Computer Science and Information Engineering,
National Ilan University, I-Lan, Taiwan
`wechen@niu.edu.tw`

Abstract. With the growth of internet access technologies, sensors/machines performing environment sensing and control can connect to Internet anytime and anywhere. However, there is no framework for users to integrate the sensors/machines and the server. The *IP Multimedia Subsystem* (IMS) based on *Session Initiation Protocol* (SIP) and all-IP architecture has been proposed as a common platform for *Next Generation Network* (NGN). This paper proposes a SIP/IMS platform for *Internet of Things* (IoT) in WLAN-3GPP integrated networks. This IMS platform includes *Instant Message and Presence Service* (IMPS) *application servers* (ASs). The *Call Session Control Functions* (CSCFs) perform the session setup and termination for a long-term communications. The IMPS ASs handles message exchange for short-term communications. Finally, this paper analyzes the efficiency of difference query/report methods and gives a conclusion.

Keywords: Call Session Control Functions (CSCFs), Internet of Things (IoT), Instant Message (IM), IP Multimedia Subsystem (IMS), Session Initiation Protocol (SIP), Presence Service (PS).

1 Introduction

With growth of Internet, the *Internet of Things* (IoT) [1] has been rapidly and widely developed. However, there is no standard framework for integrating the IoT devices and server. The *Session Initiation Protocol* (SIP) [6] and *IP Multimedia Subsystem* (IMS) [4] are two standards for the *all-IP Next-Generation Network* (NGN). In this paper, based on SIP and IMS, we propose an integrated platform referred to as SIP/IMS platform for supporting large number of IoT devices in Wireless LAN-3GPP integrated networks. The SIP/IMS platform provides several useful functions such as registration function, transmission function and notification function. With these functions, the IoT devices can easily interact with the IoT server and the subscribers effectively retrieve the IoT information from the IoT server. The proposed SIP/IMS platform is shown in Fig. 1 and elaborated as follows.
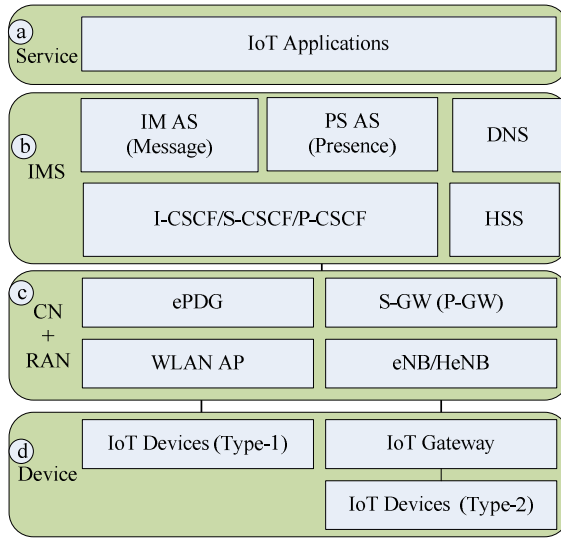
Fig. 1. The proposed SIP/IMS Platform for IoT

This platform consists of four layers that are service layer, the IMS layer, the *Core Network* (CN) and *Radio Access Network* (RAN) layer of the *3rd Generation Partnership Project* (3GPP) *Long Term Evolution* (LTE) and the device layer. In the service layer, IoT server provides application and service functions to users (i.e., device). The IMS layer is an overlay network based on the all-IP architecture of the 3GPP CN. The CN and RAN layer provides the packet transmission functions between the server and the device. The device layer consists of the IoT devices and the gateway. The components are elaborated as follows.

1.1   The IMS Layer

IMS layer contains *Call Session Control Functions* (CSCFs), *Home Subscriber Server* (HSS), *Domain Name Service* (DNS), and *Application Servers* (ASs).

The major functions in the proposed architecture are CSCFs. A CSCF can be a *Proxy-CSCF* (P-CSCF), an *Interrogating-CSCF* (I-CSCF), or a *Serving-CSCF* (S-CSCF). The CSCFs forward the SIP messages to establish the session, terminate the session and trigger the appropriate AS. In this paper, we employ two ASs that are the *Instant Message* (IM) AS [6] and the *Presence Service* (PS) AS [2][3]. The IM AS processes the instant messages between IoT server and device. The instant messages carry the commands from the IoT server to the IoT device and the reports from the IoT device to the IoT server. The PS AS handles the subscriptions from the IoT server and notifies the status to the IoT server. With the PS AS, the IoT device can save lots of energy. The IoT device sends a report once and the PS AS stores the report in its storage. Upon receipt of the subscription, the PS AS can retrieve the report from its storage without query the IoT device again. The detail message flows will be elaborated later.

1.2 The CN and RAN Layer

The environment in this paper includes a *Wireless Local Area Network* (WLAN) and a 3GPP mobile network. The WLAN *Access Point* (AP) connects to the mobile operator's core network through the *evolved Packet Data Gateway* (ePDG), which handles the secure access for WLAN. Similarly, the *Serving Gateway* (S-GW) is the anchor point for the data forwarding between the *eNB* (evolved NodeB) and the *Home eNB* (HeNB). Based on 3GPP 23.888 [5], the IP addresses of the IoT devices can be assigned by different ways. In this paper, we assume that the IP addresses are assigned by the *Packet Data Network Gateway* (P-GW).

1.3 The Device Layer

There are two types of IoT device. The first type (i.e., Type-1) device equipped with larger computing resources (e.g., CPU and memory). Type-1 device utilizes IP to connect to the IoT server through the 802.11 or LTE interface. The second type (i.e., Type-2) device equips with Zigbee or Bluetooth interface and does not have IP protocol. The type-2 device can upload their data to the IoT server through the assistance of the IoT Gateway. The IoT Gateway is a device with powerful resources (e.g., CPU and more memory). The IoT Gateway collects the data from the IoT devices and then transmits the data to the IoT server. In the proposed platform, the Type-1 device and IoT Gateway play the role as the SIP *User Agents* (UAs).

1.4 SIP Interfaces

The SIP interfaces are illustrated in Fig. 2. The CSCFs inter-connect with the IoT server, IMPS ASs, the type-1 device and the IoT gateway through SIP interfaces. For the long-term communications (e.g., video streaming), the **SIP INVITE transaction** is used to establish the *Real-time Transport Protocol* (RTP) session, and the **BYE transaction** is used to terminate the RTP session.
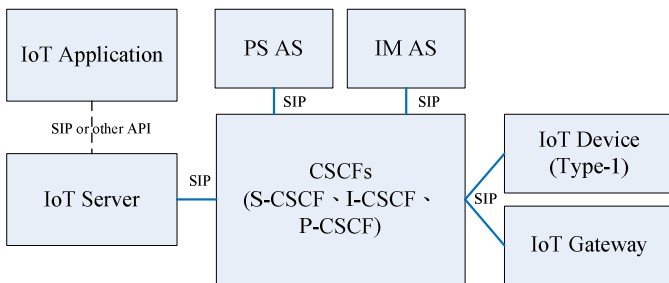

Fig. 2. SIP Interfaces between SIP Components

For the short-term communication (e.g., messaging), the instant messages utilize the **SIP MESSAGE** message to deliver the command and the report. In addition, The IoT subscribers can retrieve the information from the IoT server through the **SIP MESSAGE** message and **SIP SUBSCRIBE** message. When the **SUBSCRIBE** message is used and the status is changed, the IoT server will actively send the report to the subscriber through **SIP NOTIFY** message.

Note that the interface between the IoT application and IoT server is using SIP. It can also deploy other *Application Programming Interface* (API), which is not defined in this paper.

This rest of the paper is organized as follows. Section 2 presents the IoT functions including the registration function, the transmission function for streaming data, the transmission function for short request/report and the notification function. Section 3 analyzes the efficiency of different query methods. Finally, the conclusion is given in section 5.

2   IoT Functions

In this section, we utilize several message flows as examples to present the IoT functions provided by the proposed SIP/IMS platform.

2.1   Naming and Registration

The form of the SIP *Uniform Resource Identifier* (URI) for identifying the IoT device is defined as **<device_name>@<server_name>**. Both the IoT server and the IoT device should register to the CSCFs with the SIP URIs and their IP addresses. Fig. 3 shows the IoT device (video camera) and the IoT server registers to the CSCFs.
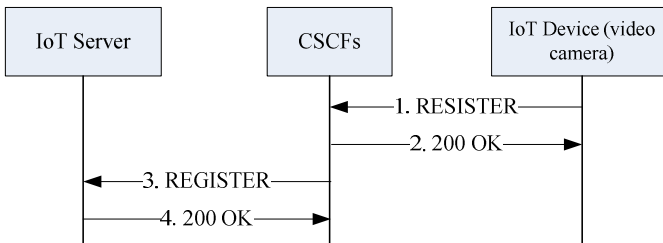

Fig. 3. SIP Registration Flow

After the registration, the CSCFs obtain the URI-to-IP mappings of the IoT device and the IoT server.

2.2   Streaming Data Transmission

Assume that the IoT device attempts to utilize the RTP to send the streaming video to the IoT server. The IoT device should use **SIP INVITE** transaction to establish an RTP session. The message flow is demonstrated in Fig. 4 and the detail steps are elaborated as follows.
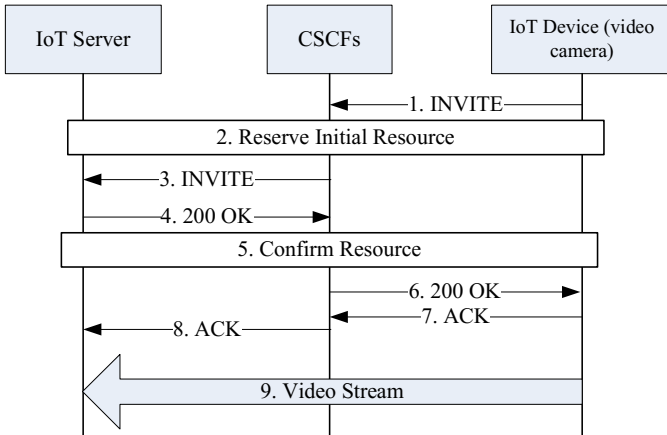

Fig. 4. Video Stream Establishment Flow

Step 1. To create a session, the IoT device sends an **INVITE** message to the IoT server. The **INVITE** message includes the QoS requirements in the *Session Description Protocol* (SDP). The **INVITE** message is sent to the CSCFs.

Step 2. Upon receipt the **INVITE** message, the CSCFs retrieve the QoS requirements from the SIP header fields and SDP fields.

Step 3. The CSCFs then perform the initial resource reservation and forward the **INVITE** message to the IoT server.

Step 4. The IoT server replies a **200 OK** message to confirm the IP address and port number for receiving the video stream.

Step 5. Upon receipt of the **200 OK** message, the CSCFs confirm the reserved resource.

Step 6. ~ Step 8. The CSCFs forward the **200 OK** message to the IoT device, and the IoT device replies an **ACK** message to the server through the CSCFs.

Step 9. The RTP session is established and the IoT device starts to send the video stream to the IoT server.

1. Note that through the above steps, the session for the video stream is set up and the resource is reserved.

2.3   Short Message Transmission

The session created by the **SIP INVITE** transaction is suitable for the long-term communication. However, there are many IoT devices such as light and thermometer,

which transmit, or receive short messages only. For such devices, to set up a session is not effective. Therefore, we utilize instant messages to transmit this kind of data. Specifically, the **SIP MESSAGE** message is used in this case. Fig. 5 demonstrates an example for transmitting the request (Case A) and the report (Case B).
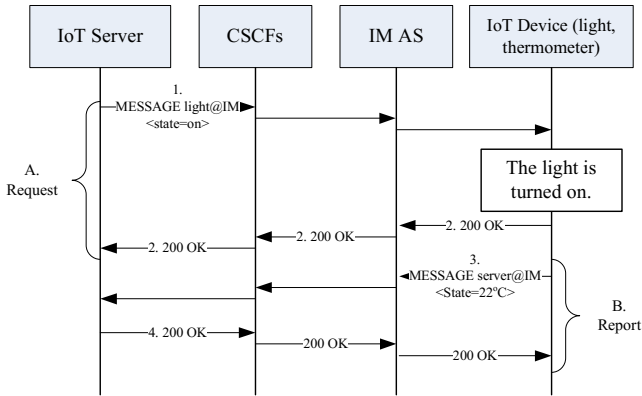

Fig. 5. Request and Report Message Flow

Case A. IoT server turns on the light.

Step 1. The IoT server sends a **MESSAGE** to the IoT device (i.e., a light) with the command <state=on> stored in the SIP message body.
Step 2. Upon receipt the command, the light is turned on and the IoT device replies a **200 OK message** to the server.

Case B. The Thermometer reports its status.

Step 1. In this case, the IoT device equips a thermometer and reports the status to the server. The device sends a **MESSAGE** with the report <state=22°C> to the server.
Step 2. Upon receipt the message, the server stores the record and replies a **200 OK message**.

   Note that in **Case B**, the IoT device communicates with the IoT server directly.

2.4   The Subscription and Notification Functions

The user (i.e., subscriber) can retrieves the information of the IoT devices from the IoT server. In the SIP/IMS platform, we provide two ways to retrieve the information. The subscriber can utilize the **MESSAGE** to query the IoT server, or the subscriber can utilize **SUBSCRIBE** and **NOTIFY** to subscribe the information. The **SUBSCRIBE/ NOTIFY** example shown in Fig. 6 is described as follows.
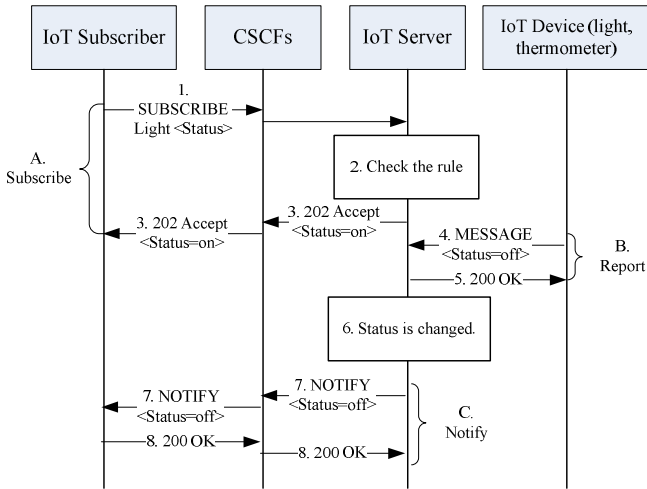
Fig. 6. Subscription and Notification Flow

Step 1. The IoT subscriber requests the status of the IoT device to the IoT server.

Step 2. Upon receipt the **SUBSCRIBE** message, the IoT server checks the rule to see whether the subscription can be accepted.

Step 3. The IoT server replies a **202 Accept** message to IoT server with the <Status=on>.

Step 4. The IoT device reports its status by using the **SIP MESSAGE** message with the <state=off>. The **MESSAGE** message is sent to the IoT server.

Step 5. The IoT server stores the status and replies a **200 OK** message.

Step 6. The IoT server finds the status of the IoT device is changed. The IoT server will check its storage to find the subscribers of this record.

Step 7. ~ Steps 8. The IoT server sends the notification to the IoT server actively by using **SIP NOTIFY** message.

In this design, the IoT device can report to the IoT server, and the IoT server can duplicate and dispatch the report to multiple subscribers. In addition, the subscriber can set the notification conditions or rules in the **SUBSCRIBE** messages. The notifications are sent to the subscriber only when the conditions/rules are matched.

3   Efficiency of Query/Report

Based on the description in Section 2, the subscriber can initialize the query by using **MESSAGE** method or subscribe the event through **SUBSCRIBE/NOTIFY** method. In this section, we attempt to conduct the analytic models to analyze the query and report activities. Since the IoT server will actively inform the subscriber in the **SUBSCRIBE/NOTIFY** method (i.e., the efficiency is 100%), this section only analyzes the **MESSAGE** method. Fig. 7 shows the timing diagram example for the report and query activities on an IoT server. At the time t_0, t_2, and t_4, the IoT device issues

report requests to the IoT server since the IoT device detects the state change or the timer expires. The inter-report intervals $t_2 - t_0$, $t_4 - t_2$, etc., are denoted by a random variable t_r. In this figure, subscriber's queries are performed at t_1 and t_3, where the query interval is represented by a random variable t_q.
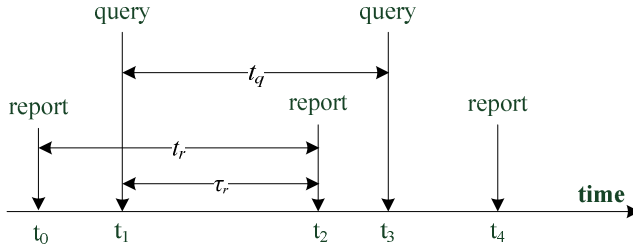

Fig. 7. Timing diagram for report and querying

The interval τ_r between a query and the next report (e.g., $t_2 - t_1$) is the *excess life* of the inter-report interval. At a report, the IoT server stores the IoT information into its database. At a query, the IoT server retrieves the information and sends it to the subscriber. Let p_e be the probability that the report record is successfully retrieved by a query, then

$$p_e = \Pr\left[t_q > \tau_r\right].$$

Note that the higher p value means that the query is more effective. Usually, the "fixed query" and "exponential query" intervals are considered in the analytic model [8]. The "fixed query" means that the subscriber performs queries with a fixed interval $1/\mu$. In the "exponential query", the inter-query interval follows the exponential distribution with mean $1/\mu$. Assume that the inter-report intervals t_r follow an exponential distribution with mean $1/\lambda$. In other words, the report process creates a *Poisson* process. That is, for an arbitrary time interval T, the number X of reports occurring in this period has a Poisson distribution [7]. That is

$$\Pr[X = x, T = t] = \left[\frac{(\lambda t)^x}{x!}\right] e^{-\lambda t} \tag{1}$$

and

$$\Pr[t_c > \tau_r] = 1 - \Pr[X = 0, T = t_c]. \tag{2}$$

From (2), the probability p_e for fixed query is presented as

$$p_e = 1 - e^{-\frac{\lambda}{\mu}}. \tag{3}$$

Similarly, the probability p_e for exponential query is expressed as

$$p_e = 1 - \int_{t_c=0}^{\infty} e^{-\lambda t_c} \mu e^{-\mu t_c} dt_c = \frac{\lambda}{\lambda + \mu}. \tag{4}$$

Fig. 8 plots p_e for "fixed query" and "exponential query" based on (3) and (4).
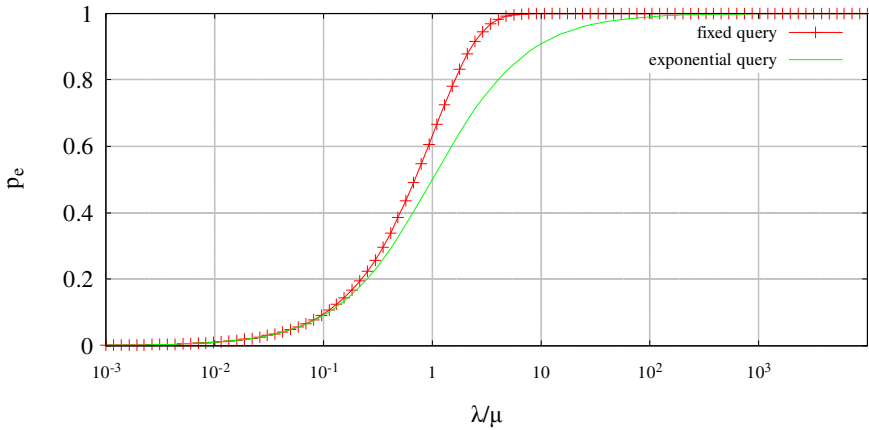
Fig. 8. Comparing fixed and exponential query.

The figure indicates that p_e for fixed query is larger than that for exponential query. That is the fixed query is more effective than the exponential query. In the rest of this paper, we only consider the exponential query to further study the report activity. General conclusions drawn from this paper also apply to fixed query. Consider the inter-report interval random variable t_r with mean $1/\lambda$, density function $f(\cdot)$, and Laplace transform $f^*(s)$. The excess life τ_r has a distribution function $R(\cdot)$, density function $r(\cdot)$, and Laplace transform $r^*(s)$. Since exponential query is a Poisson process, t_1 in Fig. 7 is a random observer of the t_r intervals. From the excess life theorem [7]

$$\gamma^*(s) = \left(\frac{\lambda}{s}\right)[1 - f^*(s)]. \tag{5}$$

Based on (5), we derive p_e as

$$p_e = \Pr[t_q > \tau_r]$$

$$= \int_{t_c=0}^{\infty} \int_{\tau_r=0}^{t_q} \gamma(\tau_r)\, \mu e^{-\mu t_q}\, d\tau_r\, dt_q$$

$$= \int_{t_c=0}^{\infty} R(t_q)\mu e^{-\mu t_q}\, dt_q$$

$$= \left.\frac{\mu \gamma^*(s)}{s}\right|_{s=\mu}$$

$$= \left(\frac{\lambda}{\mu}\right)[1 - f^*(\mu)]. \tag{6}$$

Assume that t_r is a Gamma random variable with mean $1/\lambda$, variance V, and Laplace transform

$$f^*(s) = \left(\frac{1}{V\lambda s+1}\right)^{\frac{1}{V\lambda^2}}. \tag{7}$$

Based on (7), (6) is rewritten as

$$p_e = \left(\frac{\lambda}{\mu}\right)\left[1 - \left(\frac{1}{V\lambda\mu+1}\right)^{\frac{1}{V\lambda^2}}\right]. \tag{8}$$

We can verify (8) by setting t_r. When t_r is exponentially distributed (i.e., $V = 1/\lambda^2$), (8) is rewritten as $p_e = \lambda/(\lambda + \mu)$. The result of (8) is the same as (4).

Fig. 9 plots p_e for Gamma inter-report intervals with different variance values. The figure indicates that p_e decreases as the variance V increases. In other words, when the report activities become more irregular (i.e., V is larger), the queries with more than one reports and more queries without any report are observed. Therefore, lower p_e is observed. Then, the query efficiency is worse when the report activities become more irregular.
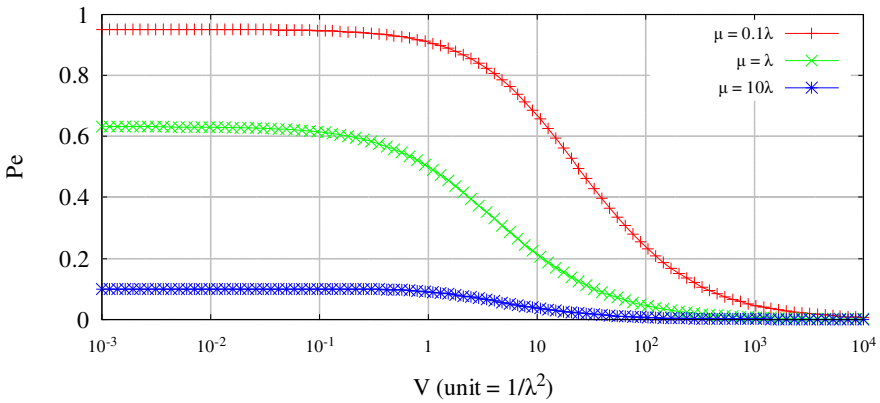

Fig. 9. Effects of the variance of the inter-report intervals on p_u.

4   Conclusion

In this paper, a SIP/IMS platform is proposed to support Internet of Things in WLAN-3GPP integrated networks. The SIP/IMS platform provides the naming of SIP URI for various IoT devices, flexible communication types (i.e., long-term and short-term communications), and query/report methods for the IoT subscribers. After the IoT functions are elaborated, we conduct the analytic models to analyze the efficiency of the query/report methods. Our study indicates that the fixed query outperforms the exponential query and the efficiency decreases as the variance of the report increases.

References

1. Commission of the European Communities, Internet of Things — An Action Plan for Europe, (June 2009)
2. Chen, W.-E., Lin, Y.-B., Liou, R.-H.: A weakly consistent scheme for IMS presence service. IEEE Transactions on Wireless Communications 8(7), 3815–3821 (2009)
3. Quan, W., Wu, J., Zhan, X., Huang, X., Ma, Y.: Research of presence service testbed on cloud-computing environment. In: 2010 3rd IEEE International Conference on Broadband Network and Multimedia Technology (IC-BNMT), October 26-28, pp. 865–869 (2010)
4. 3GPP TS 23.228 v12.3.0, IP Multimedia Subsystem (IMS); Stage 2 (Release 12), (December 2013)
5. 3GPP TR 23.888 v11.0.0, System Improvements for Machine-Type Communications, Release 11 (September 2012)
6. Rosenberg, J.: SIMPLE Made Simple: An Overview of the IETF Specifications for Instant Messaging and Presence Using the Session Initiation Protocol (SIP). IETF. RFC 6914 (April 2013) (retrieved September 24, 2013)
7. Ross, S.M.: Introduction to Probability Models. Academic, New York (1985)
8. Lin, Y.-B., Tsai, M.-H.: Caching in I-CSCF of UMTS IP multimedia subsystem. IEEE Transactions on Wireless Communications 5(1), 186–192 (2006)

Globally Optimized Cooperative Game for Interference Avoidance in WBAN

Wen-Kai Liu and Tin-Yu Wu

Department of Computer Science and Information Engineering,
National I-Lan University,
I-Lan, Taiwan, R.O.C.
kevin1990115@gmail.com, tyw@niu.edu.tw

Abstract. Wireless Sensor Networks (WSNs) have been developed for collecting and monitoring environmental data over the years. Recently, based on the idea of WSN, researchers have begun to monitor the human or animal body by placing sensor nodes on the skin or inside the body. The wearable relay nodes then collect biosignals from the sensor nodes and send the collected data to a sink node for data storage. However, the coverage is a typical problem in WSNs, which may not only generate interference but also affect system reliability. While dealing with very important bioinformation, supposing the system reliability is decreased due to high delay or packet loss, important bioinformation might be lost and could be life-threatening when, for example, a patient's heart stops beating but medical personnel are not warned. Therefore, this paper proposes a non-zero-sum cooperative game model to control the transmission power of the system for reducing the interference level between simultaneous transmissions and solving contention between different messages.

Keywords: WSN, WBAN, game theory, global optimization.

1 Introduction

Over the past few decades, researchers have been working on the development of wireless technologies that started from wireless transmission, Wireless Local Area Network (WLAN) that enables users to connect to Access Points (APs) via computers or cell phones for wireless data exchange, to Wireless Sensor Network (WSNs), in which sensors are densely deployed for data collection and the data collected by the wireless gateway are sent to a sink via wireless communication for research purposes. In recent years, because of the wide use of medical information, researchers integrated the concept of WSN and WLAN and proposed Wireless Body Area Network (WBAN). WSN and WBAN operate similarly but the biggest difference is that sensors are placed on or implanted inside the human or animal body. Due to the limitations of sensor placement, researchers found that sensor material and transmission methods of WBAN differ greatly from WSN and WLAN. Therefore, in the WBAN studies, much more attention has been paid to the influence of transmission signals on the human body. Since WBAN are often used for medical purposes, many researchers also have been focusing on issues about interference, latency and power.
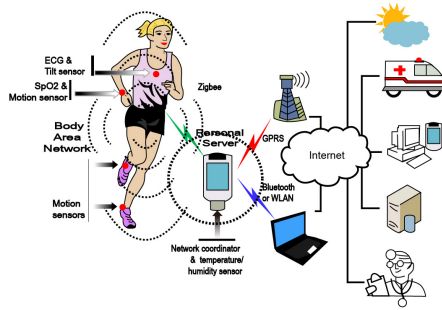
Fig. 1. Typical WBAN structure

Fig. 1. shows a typical WBAN structure, in which sensors are deployed in, on or around the human body to collect bioinformation, like heartbeat, blood pressure, SpO2 and so on. Gathered by a wearable network coordinator, the bioinformation is then transmitted to the database via Bluetooth or WLAN.

Table 1. WBAN Specifications

Attribute	Value
Distance	2 meter standard. 5meter special case
Start up time	< 100 ns
Network Setup time	< 1 sec/device
Power consumption	~ 1mW / Mbps
Network density	2-4 Nets / m^2
Latency (end to end)	10 ms
Network size	Max: 100 devices / Network

According to the bioinformation from sensors, the healthcare team or medical service team can provide the most appropriate treatment for each patient. When emergencies occur, the WBAN system send timely alerts, substantially reducing the postponement of hospitalization.

Table 1.   lists the WBAN specifications. Because current WBAN applications are related to medical care and patient monitoring, in case that sensors stop working in emergency situations, attention has been paid to several specific issues, including power consumption, transmission distance and latency. Basic requirements of a WBAN are listed below:

- High efficiency (Low power consumption)
- Scalable duty cycle and sensor nodes
- Strict QoS
- High transmission reliability
- Guaranteed low data latency
- Guaranteed availability of bandwidth
- Path loss, especially in human body related applications
- Coexistence with other WBANs
- Security

The rest of this paper is structured as follows. Section II offers an overview of related work and the technologies adopted in this paper. Section III described our proposed scheme, its operation flow and principles. Section IV provides the simulation results to compare our scheme with others. Section 5 concludes this paper and gives the future objectives.

2   Related Work

Transmission quality and operation time of wireless systems are often affected by factors, like interference coverage and contention between transmissions of different nodes, especially in wearable and implantable WBANs. As shown in Fig. 2. , T_{11} and T_{22} refer to normal transmission paths. When T_{12} and T_{21} are used for transmission, interference and latency problems may occur, affecting the real-time message delivery and even delaying the emergency medical treatment. Therefore, latency is a very important issue in WBANs for medical purposes.
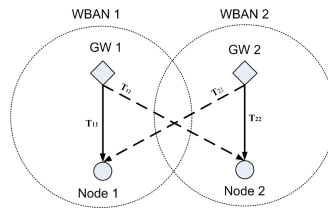

Fig. 2. Inter-network interference model in WBANs

Game theory was a branch of applied mathematics but now has been widely used in a variety of fields. Game theory is used to analyze the behaviors of rational players in a game and predict the strategies they will choose next. Assuming every possible situation, each player has various strategies to choose and will receive a payoff according to the choice of strategies. Whether a strategy is good or not can be determined by the payoff. Rewards are provided for well-behavior players while punishments are imposed on misbehavior players.

Games can be classified in many different ways and some of the most common examples are listed below:

- Cooperative and non-cooperative games:
  In a cooperative game, players work together to beat the game and vice versa. However, cooperative games are not absolutely beneficial under some special conditions.
- Self-interested and non self-interested games:
  In a self-interested game, players choose selfish strategies. Supposing all players choose selfish strategies in the game and no player can receive the payoff, this set of strategies is called a Nash equilibrium.
- Zero-sum and non-zero-sum games:
  When a player's reward is balanced by the losses of the other players, i.e. Player A's losses equals to Player B's reward, this is a zero-sum game.

Usually used for conflict avoidance and strategy-making, Game theory also can assess possible strategies to achieve the optimal outcome. For this reason, this paper

uses game theory for global optimization in WBAN environment. Some studies have been made to solve contention in wireless environments by game theory. In [1], *Zhi-mi Cheng et al.* used the energy consumption and transmission power as parameters, controlled the transmission power in simulated cooperative, non-cooperative and grand coalition games, and found the optimal way to minimize the interference produced by sources and guarantee energy efficiency. *Fan-Hsun Tseng et al.* [2] observed the duty cycle in a WSN and found that sensors generate different power consumption in different situations. Motivated by this, a non-zero-sum duty-cycle game was proposed to find the most efficient power conservation scheme to achieve the longest network lifetime. In [3], *Ramtin Kazemi et al.* investigated the cross-interference between links in different WBANs and proposed a non-cooperative power control game for WBAN. Also, an adaptive pricing mechanism was presented to dynamically adjust the tradeoff between utility and power consumption based on the channel gains and users' power budget whenever new nodes join the system. The researchers (*Weidong Su et al.*, 2013) have also developed a Static Bayesian Game Modeling for quality of service (QoS) improvement in overlapped WBAN environment.

3   Proposed Scheme

According to the above-mentioned literatures and after the environmental assessment process, we found that when multiple WBANs coexist in an area, the nodes in the overlapping part may simultaneously exist in several games. As shown in Fig. 3. , the nodes in red color in the overlapping area are the players in both Game 1 and 2. Based on game theory, the nodes in red color must be considered in Game 1 or 2separately.

When such a situation occurs, special considerations must be made to global optimization. Global optimization means that all nodes must work simultaneously. Supposing one of the nodes stops working, it could be life-threatening. For example, when the heartbeat sensor stops sending messages because the battery has run down, the emergency medical staff will not receive alerts and may lead to delays in the best rescue time.
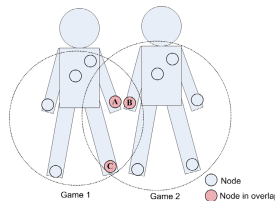

Fig. 3. The nodes in overlapping area

To achieve global optimization, this paper proposed a non-zero-sum cooperative game model, in which nodes within the transmission range can exchange information, like residue power or signal strength. After two nodes exchange information, they work together as a team and propose the most beneficial strategy to beat the game. As displayed in Fig. 4., Node A and B exchange information with each other and choose the optimal strategy together to beat Node C.
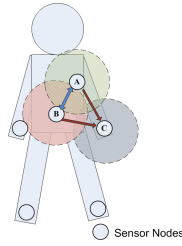
Fig. 4. Proposed cooperative game model

Fig. 5. shows the process of our proposed game model. In the Initialization stage, the system measures the parameters defined in Table 2. and checks whether the values conform to the constraints in the table. If yes, move to the Detection stage. In this stage, Node A uses Equation (1) to calculate the interference level $I_i(t)$, where η_i denotes the interference parameter of Player i and $P_i(t)$ denotes the transmission power of Player i at time t, which can be measured by instruments.

According to the calculation results, Node A can judge whether there is a neighboring node generating interference or not. Supposing the number of neighboring nodes generating interference is zero, it means that Node A is not interfered by other nodes currently. If not, we use Equation (2) to calculate the instantaneous channel gain between nodes and Equation (3) to calculate the cost needed to reduce the interference, where $\overline{P_i}$ denotes the upper bound of transmission power consumption, β_i denotes the channel gain of transmission, $X(t)$ denotes the interference at time t and when the cost of Player i paying for abating interference is $C_i(t)$, and $I_{-i}(t)$ denotes the interference level of the players excluding Player i. The calculation process of the non-zero-sum game will be discussed next.
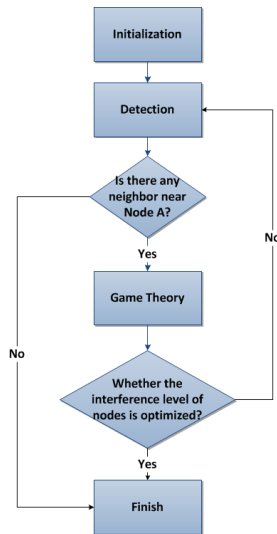

Fig. 5. Process of proposed game model

$$I_i(t) = \eta_i P_i(t) \tag{1}$$

$$P_i(t)\left[\overline{P_i} - \frac{1}{2}P_i(t)\right] - \beta_i[X(t) - \eta_i P_i(t)] \tag{2}$$

$$C_i(t) = \beta_i I_{-i}(t) = \beta_i[X(t) - \eta_i P_i(t)] \tag{3}$$

Table 2. Symbol definitions and ILP

$P_i(t)$: The transmission power of player i at time t.

η_i : The interference parameter of player i.

X(t) : The stock of interference in the current network by time t and
$\quad\quad$ $C_i(t)$ be the cost of player i paying for abating interference.

δ : A constant natural decay rate of the interference.

$W_X(N, X, t)$: A continuously differentiable function.

β_i : The channel gain of transmission.

Minimize

$$\sum_{i=1}^{n} X^i(t)$$

Subject to

$0 < \eta_i < 1,$

$0 < t < \infty,$

The key elements of the proposed game model include:

- Players:
  In the proposed game model, each node in the WBAN is a player.
- Strategies of each player:
  In order to win a game, every player chooses a strategy that gives himself the highest payoff. In the proposed game model, the strategy Player i chooses at time t, $\psi_i^{N}(t)$, can be denoted by Equation (4), where $W_X(N, X, t) : R^m \to R$ is a continuable differentiable equation.

$$\psi_i^{N}(t) = \overline{P_i} + \beta_i \eta_i + \eta_i W_X(N, X, t) \tag{4}$$

- The payoff a player receives for playing a particular strategy:
  The payoff means the outcome after Player i chooses a particular strategy, usually used to determine whether this is the optimal strategy. In the proposed game model, Equation (5) can be used to calculate the payoff Player i receives at time t, $R_i(P_i(t))$.

$$R_i(P_i(t)) = P_i(t)\left[\overline{P_i} - \frac{1}{2}P_i(t)\right] \tag{5}$$

After some calculations, we use Equation (6) to obtain the optimal result of the node, $X^N(t)$. m_i^{N} denotes the strategy currently chosen by the player and $X^N(t)$ denotes the outcome of the game but this does not mean that this node has been globally optimized in the system.

$$X^N(t) = ex\,p(-\delta t)\,X_0 + \frac{1}{\delta}\left\{\left[\sum_{i=1}^{N} m_i{}^N\right][1 - ex\,p(-\delta t)]\right\} + RW \tag{6}$$

Now, we run into the second "whether" in the flowchart. In this paper, we add a RW value to judge whether the interference level of nodes is optimized at the end of the game and whether the system has reached steady-state. The RW value can be calculated by Equation (7). Since the optimization result must be compared with other nodes, β_{rand}, η_{rand}, $\overline{P_{rand}}$ and $P_{rand}(t)$ in the equation refer to a parameter set randomly selected from the rest nodes. Using the parameter set for calculation, we can find the value of RW.

$$RW = \frac{\beta_i[X(t) - \eta_i P_i(t)]}{P_i(t)\left[\overline{P_i} - \frac{1}{2}P_i(t)\right]} - \frac{\beta_{rand}[X(t) - \eta_{rand}P_{rand}(t)]}{P_{rand}(t)\left[\overline{P_{rand}} - \frac{1}{2}P_{rand}(t)\right]} \tag{7}$$

Global optimization is not limited in the WBAN where the node is and must be compared with other nodes. Therefore, in addition to the parameters of the original node, those of neighboring nodes or one random nodes in other WBANs are retrieved for calculation. When the RW achieves the optimal value, it means the comparison between the node and other nodes are also in the optimal state and the node can be judged as a globally optimized node and will not be included as a player in the next game. After playing a game several times, supposing all nodes have been optimized, the system reaches the Finish stage in Fig. 5. . If not, the game will be continued until all nodes have been globally optimized.

4   Simulation Analysis

In this section, we present an example to explain our proposed globally optimized non-zero-sum cooperative game. 50 source nodes, s_0, s_1, $s_2 \ldots s_{50}$, are taken into consideration to control the transmission power and coordinate with one another's interference. In brief, supposing the sensors are the same, according to the definitions of parameters η and β, the optimal transmission power, the interference signal level and interference gain will be inversely proportional to η and β. For a clearer simulation, the optimal transmission power, the optimal interference signal level, the maximum gain and the maximum payoff consider only the influence of η. Table 3. lists the simulation parameters and uses MATLAB 2008a to simulate the theoretic results of the parameters.

Table 3. Simulation parameters

$\overline{P_i}$	η	β	γ	δ	t
5	[0, 0.14]	0.05	0.05	0.02	5

4.1   Optimal Transmission Power

Fig. 6.   depicts the relationship between the interference parameter η and the optimal transmission power. The increase of the interference parameter η results in the decrease of the transmission power. Equation (5) reveals that the increase of the interference parameter η denotes the increase of the interference from other nodes. Therefore, the node has to reduce the transmission power to prevent more interference and additonal cost. In a cooperative game, the node makes the decision based on the strategies of other nodes.
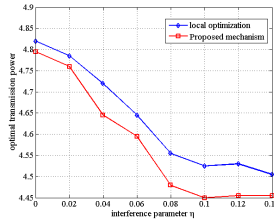

Fig. 6. Optimal transmission power

4.2   Optimal Trajectory of the Interference Signal Level

Fig. 7. shows that in a cooperative game, the interference parameter η is inversely proportional to the interference signal level. When the interference parameter η increases, the interference signal level decreases and vice versa. The reason for such a phenomenon is the transmission power. The increase of the interference parameter η denotes the more cross interference between nodes. Thus, the node has to reduce the transmission power to solve the problem. However, with the reduction of the transmission power, the interference signal level reduces also.
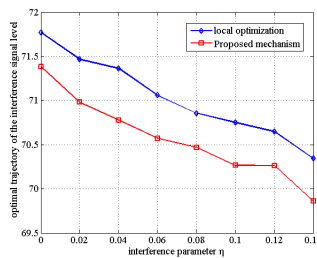

Fig. 7. Optimal trajectory of the interference signal level

5   Conclusion and Future Objectives

In this paper, we proposed a non-zero-sum cooperative game model that controls power to minimize the interference generated by nodes and guarantee more power saved. The simulation results prove that our proposed scheme can reach the balance

between interference and energy efficiency. Moreover, our scheme enables the system to reach global optimization and steady-state for WBAN that focuses on power-related issues. Therefore, the proposed scheme can not only improve transmission reliability and node lifetime, but also solve the loss of emergency information transmission. In this way, the possibility of postponed hospitalization can be reduced and great contributions be made to medical related applications.

Our future objective is to implement the proposed scheme in real WBAN environment and to observe and gather related information for parameters and structure adjustment.

References

1. Cheng, Z.-M., Zhou, X.-W., Ding, Y., Miao, X.-N.: A Cooperative Differential Game of Transmission Power Control in Wireless Networks. Wireless Personal Communications 72(4), 2015–2027 (2013)
2. Tseng, F.-H., Cho, H.-H., Chou, L.-D., Shih, T.K., Chao, H.-C.: Efficient Power Conservation Scheme in Non-Zero-Sum Duty-Cycle Game for Wireless Sensor Networks. Accepted for publication in International Journal of Sensor Networks (IJSNet)
3. Kazemi, R., Vesilo, R., Dutkiewicz, E., Fang, G.: Inter-network interference mitigation in Wireless Body Area Networks using power control games. In: Communications and Information Technologies (ISCIT 2010), October 26-29, pp. 81–86 (2010)
4. Su, W., Jim, Z., Cho, J.: A Static Bayesian Game Modeling for QoS Support in Overlapped WBAN Environment
5. Gharehshiran, O.N., Attar, A., Krishnamurthy, V.: Collaborative Sub-Channel Allocation in Cognitive LTE Femto-Cells: A Cooperative Game-Theoretic Approach. IEEE Transactions on Communications 61(1), 325–334 (2013)
6. Wu, G., Ren, J., Xia, F., Yao, L., Xu, Z., Shang, P.: A Game Theoretic Approach for Interuser Interference Reduction in Body Sensor Networks. International Journal of Distributed Sensor Networks 2011, 12–30 (2011)
7. Shin, S., Weidong, S., Cho, J.: A Game Theory Model to Support QoS in Overlapped WBAN Environment. In: International Conference on Ubiquitous Information Management and Communication (ICUIMC 2012), Article No. 47 (February 20, 2012)
8. Fang, G., Dutkiewicz, E., Yu, K., Vesilo, R., Yu, Y.: Distributed Inter-Network Interference Coordination for Wireless Body Area Networks. In: Global Telecommunications Conference (GLOBECOM 2010), December 6-10, pp. 1–5 (2010)
9. Larsson, E.G., Jorswieck, E.A., Lindblom, J., Mochaourab, R.: Game Theory and the Flat-Fading Gaussian Interference Channel. Signal Processing Magazine 26(5), 18–27 (2009)
10. Leshem, A., Zehavi, E.: Cooperative game theory and the Gaussian interference channel. IEEE Journal on Selected Areas in Communications 26(7) (September 2008)
11. Suris, J.E., DaSilva, L.A., Han, Z., MacKenzie, A.B.: Cooperative Game Theory for Distributed Spectrum Sharing. In: IEEE International Conference on Communications, June 24-28, pp. 5282–5287 (2007)

A Study of Random Neural Network Performance for Supervised Learning Tasks in CUDA

Sebastián Basterrech, Jan Janoušek, and Vaclav Snášel

IT4Innovation
VŠB–Technical University of Ostrava,
Czech Republic
{Sebastian.Basterrech.Tiscordio,Jan.Janousek,Vaclav.Snasel}@vsb.cz

Abstract. The *Graphics Processing Unit (GPU)* have been used for accelerating graphic calculations as well as for developing more general devices. One of the most used parallel platform is *Compute Unified Device Architecture (CUDA)*. This one allows to implement in parallel multiple GPU obtaining a high computational performance. Over the last years, CUDA has been used for the implementation of several parallel distributed systems. At the end of the 80s, it was introduced a stochastic neural network named *Random Neural Networks (RNN)*. The method have been successfully used in the Machine Learning community for solving many learning tasks. In this paper we present the gradient descent algorithm for the RNN model in CUDA. We evaluate the performance of the algorithm on two real benchmark problems about energy sources, and we compare it with the obtained using a classic implementation in C.

Keywords: Random Neural Network, Parallel Computing, CUDA, Gradient Descent Algorithm.

1 Introduction

The *Compute Unified Device Architecture (CUDA)* is a parallel computing platform, which uses the *Graphics Processing Units (GPUs)* for increasing the computing performance. Parallel computing using CUDA has been proven to be an powerful tool for implementing distributed systems using big data. Recently, this platform has been used for the implementation of several meta-heuristic and parallel bio-inspired systems [1–4]. At the end of the 80s a stochastic connectionist model called *Random Neural Network (RNN)* was introduced in the *Machine Learning (ML)* community [5]. The model belongs to the domain of Neural Network methods and the Queueing models. In the queueing theory context, the RNN model is often called *G-networks*. A RNN is a distributed system where the nodes exchange positive and negative signals among according to a probabilistic routing. The signals are modeled using Poisson processes. The neurons are excited receiving signals from their neighborhoods and from outside (the environment). Each neuron processes the input information using a special counter

function as its transfer function. A *gradient descent (GD)* algorithm for RNN model was introduced by Erol Gelenbe in 1993 [6]. The RNN have been used with a relative success in supervised learning tasks. Surveys about the model and their applications are provided in [7, 8]. In this work, we present the GD algorithm for the RNN model in supervised learning implemented in CUDA. We test the implementation performance with two kinds of ML problems. The first one, is a traditional supervised learning task where the data is independent of the time. We use a data set about energy efficiency in buildings from the UCI Machine Learning Repository [9, 10]. The second one, is a supervised temporal learning task, in this case we use a problem about power solar estimation [11]. We compare the performance of the results with an implementation in C. Additionally, we analyze the impact of the density in the graph of node connections in the accuracy of the model. We present in this article another GD algorithm variation, which is called *GD with a stochastic weight updating.* This consists in updating each weight according to some probability.

The remainder of this article is organized as follows. In Section 2, we introduce some related work on parallel computing and CUDA. Section 3 presents the RNN model. The GD algorithm for RNN is presented in Subsection 3.1. In Section 4, we describe the methodology used for the parallel computing experiments. Next, we present the benchmark problems used and the empirical results obtained. Finally, the last Section concludes the article.

2   Parallel Implementation

In 1965, Gordon Earle Moore published a paper predicting that the number of transistors packaging in integrated circuits doubles every year [12]. This relationship is often referred as *Moore's law.* Even though this assumption was corrected from a year to 18 months, the growing up number of transistors per circuit is a trend that has continued for the last 30 years. Nowadays, a common solution to increase the number of microprocessors is to use graphic cards. The *Graphics Processing Unit (GPU)* was originally designed in order to accelerate graphic calculations. Although, the community of developers have started to use GPUs for more general computing devices. In 2006, *NVIDIA Corporation* presented the new parallel platform *Compute Unified Device Architecture (CUDA)*. This platform has been used for the implementation of several types of Neural Networks, such as Self-Organizing Maps [1], Hpfield Networks [2], Back-Propagation algorithm for Neural Networks [3] and Fuzzy Neural Networks [4]. The software development kit (SDK) of CUDA includes extensions for *C*, *C++*, *Fortran*, *Python* and other languages. This advantage allows for relatively easy programming without the knowledge of CUDA assembler. The main advantage of CUDA platform is the low price of devices a high computational performance with the possibility to compute in a distributed parallel system multiple GPUs.

3   Description of the Random Neural Network Model

A Random Neural Network is a specific type of spiking neuron network. Each neuron receives signals from outside and from its neighborhoods. There are two disjoint types of signals: *inhibitory* and *exhibitory* which are also called *negative* and *positive*, respectively. Each neuron has associated an integer variable called *potential*. After numbering the neurons in an arbitrary order, let's denote by $S_i(t)$ the potential of neuron i at time t. When the neuron's potential is strictly positive, we say that the neuron is *excited* or *active*. Only active neurons can fire positive and negative signals. The potential of a neuron is increased by 1 when a positive signal arrive. In the case that a neuron with strictly positive potential receives a negative signal its potential decreases by 1, another case (potential equal to 0) its potential remains null. An active neuron i sends signals according to a Poisson process of rate $r_i > 0$. This parameter is often called *service rate*. The pattern of routing is given by the following rules: d_i is the probability that a neuron i fires spikes to outside, $p_{i,j}^+$ is the probability that a neuron i spikes a positive signal to its neighborhood j and $p_{i,j}^-$ is the probability that a neuron i sends a negative signal to neuron j. The arrival signals from outside of positive spikes to a neuron i is Poisson with rate λ_i^+. A Poisson process with rate λ_i^- models the arrival signals from outside to a neuron i. Let N be the number of neuron in the system. In order to avoid the trivial case where the network is inactive, usually is assumed that $\sum_{i=1}^N \lambda_i^+ > 0$.

At every time t the network is characterized by the potential of its neurons, let $\mathbf{S}(t) = S_1(t), \cdots, S_N(t))$ be the state of the network at time t. Note that $\mathbf{S}$ is a continuous time Markov process over the state space $\mathbb{N}^N$. In [5,13] Gelenbe shows that, in an equilibrium situation and under certain algebraic hypothesis, the charges of the neurons ϱ_i satisfies the following non-linear system of equations:

$$\varrho_i = \frac{T_i^+}{r_i + T_i^-}, \qquad T_i^+ = \lambda_i^+ + \sum_{j=1}^N \varrho_j r_j p_{j,i}^+, \qquad \text{and} \qquad T_i^- = \lambda_i^- + \sum_{j=1}^N \varrho_j r_j p_{j,i}^-,$$

(1)

with the supplementary condition that, for every neuron i, we have $\varrho_i < 1$. Furthermore, if a unique non negative solution of equations 1 exists such that ϱ_i for all i, then the stationary distribution of S is given by a product form of the marginal probabilities of the neuron's potential. For more details and proofs, see [5, 14].

3.1   Random Neural Networks as Supervised Learning Tool

A supervised learning problem is defining as follows. Given a collection of K pairs $\mathcal{L} = \{(\mathbf{a}^{(k)}, \mathbf{b}^{(k)}) : \mathbf{a}^{(k)} \in \mathbb{R}^{N_a}, \mathbf{b}^{(k)} \in \mathbb{R}^{N_b}, k = 1, \ldots, K\}$, of some unknown function $f : [0,1]^{N_a} \to [0,1]^{N_b}$, the tasks consists in learning a parametric mapping $\phi(\mathbf{a}, \mathcal{L})$ between the input pattern $\mathbf{a}$ and $\mathbf{b}$, such that certain distance between $\phi(\mathbf{a}^{(k)}, \mathcal{L})$ and $\mathbf{b}^{(k)}$ is minimized for all k. The first learning algorithm for RNN was introduced by Erol Gelenbe in [6]. This procedure was based in

the gradient descent method [15]. Additionally, Quasi-Newton methods for using RNN to solve supervised learning problems were introduced in [16, 17].

The adjustable parameters of the model are the *weight connections*, which are defined by $w_{i,j}^+ = r_i p_{i,j}^+$ and $w_{i,j}^- = r_i p_{i,j}^-$, for every neurons i and j. Let $\mathcal{I}$, $\mathcal{H}$ and $\mathcal{O}$ be the set of input, hidden and output neurons, respectively. The relationship between the service rate and the weights should satisfy:

$$r_i = \frac{1}{1 - d_i} \sum_{j=1}^{N} w_{i,j}^+ + w_{i,j}^-,$$ (2)

for all $i \in \mathcal{I} \cup \mathcal{H}$. Note that, r_i is a free-parameter when $i \in \mathcal{O}$. When the model is used for solving learning tasks the input patterns are the rates of positive signal arrivals ($\lambda_i^+ = a_i, \forall i \in \mathcal{I}$). The rate of negative signal arrivals is set to zero ($\lambda_i^- = 0$ forall input neuron i). Each output neuron i has $d_i = 1$, for other neurons $d_i = 0$. The output of the model are the loads of output neurons ϱ_i for all $i \in \mathcal{O}$. All other model parameters are fixed *a priori*.The model in a supervised learning tasks consists of finding the weights $\mathbf{w}^+$ and $\mathbf{w}^-$ which minimize the function:

$$E(\mathbf{w}) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i=1}^{N} c_i \big(\varrho_i^{(k)} - b_i^{(k)}\big)^2.$$ (3)

where the factor c_i gives different degrees of importance among output variables, it is 0 when $i \notin \mathcal{O}$ and $c_i > 0$ when $i \in \mathcal{O}$.

3.2   The Gradient Descent Method for Random Neural Networks

The algorithm described in this Section was presented in [6]. The procedure is iterative, at the τth iteration a pattern $\big(\mathbf{a}^{(k)}, \mathbf{b}^{(k)}\big)$, $k = 1, \ldots, K$, is presented to the network. The relation between the iteration $\tau \geq 1$ and the index of training example k is: $k = [(\tau - 1) \mod K] + 1$, where *mod* is the modulo operation. The weights are initialized following some arbitrary criterion. The topology of the network is free of assumptions. We denote by $\mathbf{w}_{u,v}^*$ both positive and negative weight between u and v. The update of each weight at the τth iteration is proportional to the partial derivative of the error measure with respect to the weight. For each network weight $w_{u,v}^*$ the update rule is:

$$w_{u,v}^{*\,(\tau)} = w_{u,v}^{*\,(\tau-1)} - \eta^{(\tau-1)} \left[\frac{\partial}{\partial w_{u,v}^*} \sum_{i=1}^{N} \big(\varrho_i^{(k)} - b_i^{(k)}\big)^2 \right]^{(\tau)},$$ (4)

where η is a parameter in $[0, 1]$ called *learning factor* used to tune the convergence speed of the algorithm. Consider the auxiliary matrix $\boldsymbol{\Omega}$ of size $N \times N$ such that its (i, j) element is:

$$\Omega_{j,i} = \frac{w_{j,i}^+ - w_{j,i}^- \varrho_i}{r_i + T_i^-},$$ (5)

and the row vectors $\boldsymbol{\gamma}_{u,v}^{+}$ and $\boldsymbol{\gamma}_{u,v}^{-}$ of dimension N whose elements in the ith position are $\gamma_{u,v;i}^{+}$ and $\gamma_{u,v;i}^{-}$, defined as:

$$\gamma_{u,v;i}^{+} = \begin{cases} -\dfrac{1}{r_i + T_i^{-}}, & \text{if } u = i, \quad v \neq i, \\ \dfrac{1}{r_i + T_i^{-}}, & \text{if } u \neq i, \quad v = i, \\ 0, & \text{otherwise;} \end{cases} \qquad \gamma_{u,v;i}^{-} = \begin{cases} -\dfrac{1 + \varrho_i}{r_i + T_i^{-}}, & \text{if } u = i, \quad v = i, \\ -\dfrac{1}{r_i + T_i^{-}}, & \text{if } u = i, \quad v \neq i, \\ -\dfrac{\varrho_i}{r_i + T_i^{-}}, & \text{if } u \neq i, \quad v = i, \\ 0, & \text{otherwise.} \end{cases}$$

As a consequence, using a vector notation:

$$\frac{\partial \boldsymbol{\varrho}}{\partial w_{u,v}^{+}} = \boldsymbol{\gamma}_{u,v}^{+} \varrho_u \left[\mathbf{I} - \boldsymbol{\Omega} \right]^{-1}, \qquad \frac{\partial \boldsymbol{\varrho}}{\partial w_{u,v}^{-}} = \boldsymbol{\gamma}_{u,v}^{-} \varrho_u \left[\mathbf{I} - \boldsymbol{\Omega} \right]^{-1}, \tag{6}$$

where $\mathbf{I}$ is the $N \times N$ identity matrix.

We assume that there are not feedback connections from the output neurons to other neurons, that is: $d_i = 1$, for all $i \in \mathcal{O}$. Hence, for each input pattern presented at the τth-iteration, the update rule for the parameters is:

$$w_{u,v}^{*(\tau)} = w_{u,v}^{*(\tau-1)} - \eta \varrho_u^{(\tau)} \boldsymbol{\gamma}_{u,v}^{*(\tau)} \left[\mathbf{I} - \boldsymbol{\Omega} \right]^{-1} \boldsymbol{e}^{(\tau)}, \tag{7}$$

where $\left[\mathbf{I} - \boldsymbol{\Omega} \right]^{-1}$ is computed using the current input $(\mathbf{a}, \mathbf{b})$ and $\boldsymbol{e} = \varrho - \mathbf{b}$.

4   Methodology

The computation in each step can be implemented as computation of matrices, this is suitable for the implementation in parallel. The computation of r_i for each intern neuron i can be performed using a parallel reduction. For the parallel reduction, we used the function $\_shfl\_xor$, which is available on devices with computing capability 3.0 and higher. Using this function, the reduction is faster than the data exchange via shared memory. Additionally, at the same time it reduces the amount of shared memory required in the algorithm. In the case of recurrent networks, when one data pattern is presented to update each adjustable parameter in the network the main computational effort consists of computing $\left[\mathbf{I} - \boldsymbol{\Omega} \right]^{-1}$ using the expression (6). This effort has $O(N^3)$ time complexity. Although, when is used a m-step relaxation method the time complexity becomes $O(mN)$ [6]. Below, the general scheme of the algorithm can be adapted when the network is a multilayer feed-forward. In this case, the matrix $(\mathbf{I} - \boldsymbol{\Omega})$ becomes triangular, so the computational complexity decreases to $O(N^2)$. According to our experiments, in a parallel implementation of the GD algorithm for RNN, the computational effort for computing the inverse of $(\mathbf{I} - \boldsymbol{\Omega})$ is around the 90% of the total algorithm time. For the inverse matrix computation was used the CULA library for CUDA, this library provides an implementation of LAPACK library for GPU. Specifically, it provides the method $SGESV$ for the LU decomposition.

5   Empirical Results

5.1   Benchmark Description

In this paper, we use a data set from the UCI Repository Machine Learning concerning the energy sources in building. Specifically, the heating load and cooling load requirements, which are dependent variables of the energy efficiency of the buildings. This data set was previously studied in [10] and it is available in [9]. The data set contains information about 12 different building shapes simulated in *Ecotect*. The input variables of the buildings are the glazing area, the glazing area distribution, the orientation as well as other parameters, such as surface, wall and roof area. There are 768 samples and 8 feature in the data set. The output variables are the heating and cooling loads. The second data set presents serial order, thus the model must learn also the temporal dependency of the data. The data was collected from a power plant in Czech Republic between 2010-July-1 to 2011-March-31. The goal is to predict the power solar at current moment using solar irradiance measures from the past. The data set was collected every 10 minutes. The data was described and analyzed in [11, 18]. We consider a sliding windows as input information which contains 36 measures of solar irradiance. The input pattern contains 6 measures taken each 10 minutes, 6 measures taken each hour, 6 measures taken each 3 hours, 6 measures taken each 6 hours, 6 measures taken each 12 hours and 6 measures taken each 24 hours. Given this input time lag the goal is to predict the real value response. The pre-processing of the data for all cases was only the normalization in [0, 1].

5.2   Results

The accuracy of the implementation in CUDA of the GD for RNN using the energy sources data set is presented in Table 1. The error order of the accuracy when the method is implemented in CUDA is the same than when the algorithm is implemented in C. Although, when the neural network is large enough than the execution time of C code is substantially longer comparing to the execution time of the CUDA implementation. We can see that the algorithm implemented in CUDA using the first data set with 60 neurons is 1.25 times faster than the one implemented in C. More interesting are the results for larger networks. For instance, a network with 410 units and using the first data set, we obtain that the implementation with GPU is 37 times faster than the another one. Figure 1 illustrates the execution time for both implementations according the number of neurons in the network. The figures 2 and 3 show the estimation of the model in respect of the target outputs. This kind of visualization is usually presented in regression problems. For the second data set, we test the algorithm using GD with stochastic weight update. This means, the algorithm is on-line and each weight is updated (or not) according some probability. Table 2 presents the MSE for the second data set in respect of different probabilities for updating the weights. The best accuracy is presented when the probability to update a specific weight is 0.5. This result coincides with the empirical experiences using
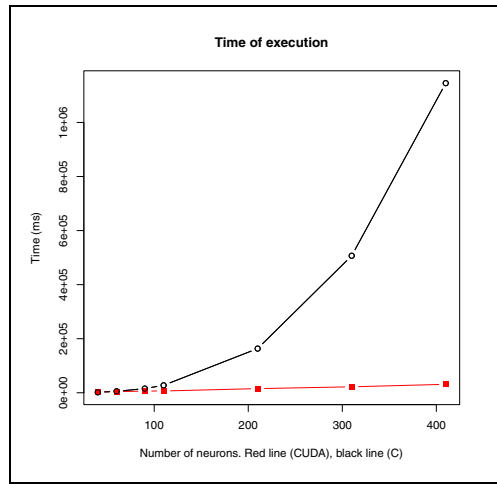
Fig. 1. Execution time of both implementations C and CUDA according the number of units in the network. The red line with square dots corresponds the time using CUDA and the black line with circle dots refers to the time using C.

Table 1. Accuracy of the GD for RNN implemented in CUDA using the first benchmark problem. Table shows the accuracy according to the density of the weight matrix connections in the recurrent hidden layer. The first column shows the sparsity of the weight matrix, for instance 0.2 corresponds to 20% of non-zero weight connections. The second column shows the MSE error when there are 40 hidden units, respectively.

Sparsity	40 hidden neurons	80 hidden neurons
0.2	0.030	0.0228
0.4	0.051	0.0105

Table 2. Accuracy of the GD for RNN implemented in CUDA using the second benchmark problem. We tested a variation of the GD method where each weight is selected to be updated according a probability. The first column shows the probability to update each weight. The second column shows the accuracy obtained. The network has 36 input neurons, 370 hidden neurons and one output. The hidden layer presents circuits and a sparsity of the 50%.

p	0.2	0.5	0.8	1
MSE	0.002790	0.002753	0.002774	0.002790

GD for the Neural Networks with sigmoid neurons. The second benchmark is a temporal learning tasks, the strategy for training a RNN was the common approach presented in the literature. We consider few delayed instances from the past as pattern inputs. Even though, the accuracy of the model for the first data set was promising, the accuracy obtained for the second data set is less
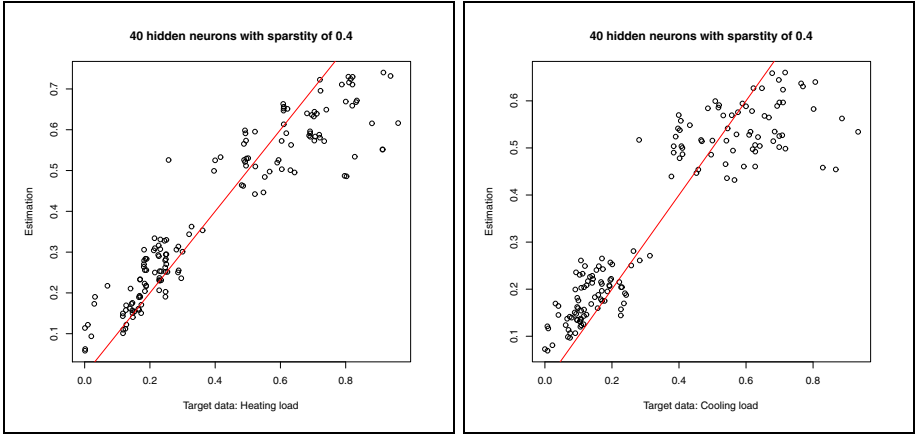
Fig. 2. RNN prediction for the energy source data using 40 hidden neurons with a matrix of hidden connections with density of 40%. The left figure shows the estimation for the variable *heating load* and the right figure illustrates the estimation of the model for the output variable *cooling load*. The red line shows the identity function.
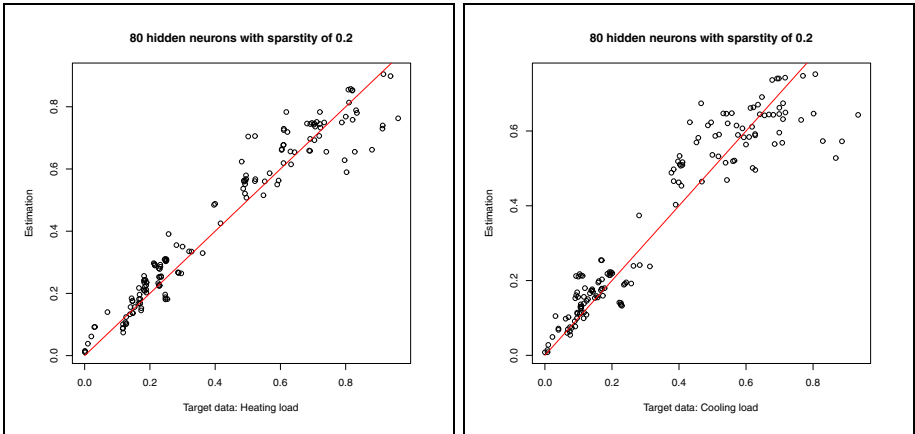

Fig. 3. RNN prediction for the energy source data using 80 hidden neurons with a matrix of hidden connections with density of 20%. The left figure shows the estimation for the variable *heating load* and the right figure illustrates the estimation of the model for the output variable *cooling load*. The red line shows the identity function.

interesting. This occurs due to the fact that the GD algorithm method for a NN with recurrences presents often convergence problems [19].

6   Conclusions and Future Work

We presented the Random Neural Network model and the procedure to use it in supervised learning problems. We described the algorithm of gradient type

which was introduced in 1993 by Erol Gelenbe [6]. The goal of this paper was analyzing the accuracy of the model in two real benchmark problem using parallel computing. We tested the performance of the model in two real problems about energy sources. We study the performance in respect of execution time and accuracy for the implementation in C and CUDA. Moreover, we analyzed the impact on the model accuracy of the density in the network graph. Below, we tested the GD with a *stochastic weight updating* performance. For the classic NN it has been shown that to update only the 50% of the weights (selected in a random way) for each input pattern can obtain better results than to update all network weights. Even though the RNN model has been proven to be a powerful tool for supervised machine learning, in practice it is hard to train it. This also occurs with the classical neural networks with recurrences [19]. The main problem that can appears in a recurrent networks is the vanishing and exploding gradient problem during the training process [19]. In future works, we will apply this approach in larger data set. For the second benchmark, we will use a larger sliding window as input information containing data from several days. As a consequence, the network topology will be larger. Additionally, we will implement in CUDA other numerical algorithms for RNN (for instance Quasi-Newton techniques) in the context of supervised learning.

References

1. Snášel, V., Klement, P., Gajdoš, P., Abraham, A.: Self Organising Maps on Compute Unified Device Architecture for the Performance Monitoring of Emergency Call-Taking Centre. In: Gavrilova, M.L., Tan, C.J.K., Abraham, A. (eds.) Transactions on Computational Science XXI. LNCS, vol. 8160, pp. 339–366. Springer, Heidelberg (2013)
2. Liang, L.: Parallel Implementations of Hopfield Neural Networks on GPU (2011)
3. Sierra-Canto, X., Madera-Ramirez, F., Uc-Cetina, V.: Parallel training of a back-propagation neural network using cuda. In: 2010 Ninth International Conference on Machine Learning and Applications (ICMLA), pp. 307–312 (December 2010)
4. Juang, C.-F., Chen, T.-C., Cheng, W.-Y.: Speedup of implementing fuzzy neural networks with high-dimensional inputs through parallel processing on graphic processing units. IEEE Transactions on Fuzzy Systems 19(4), 717–728 (2011)

5. Gelenbe, E.: Random Neural Networks with Negative and Positive Signals and Product Form Solution. Neural Computation 1(4), 502–510 (1989)
6. Gelenbe, E.: Learning in the Recurrent Random Neural Network. Neural Computation 5(1), 154–511 (1993)
7. Bakircioğlu, H., Koçak, T.: Survey of Random Neural Network applications. European Journal of Operational Research 126(2), 319–330 (2000)
8. Timotheou, S.: The Random Neural Network: A Survey. The Computer Journal 53(3), 251–267 (2010)
9. Center for Machine Learning and Intelligent Systems. UCI Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets/Energy+efficiency (date of access: February 6, 2014)
10. Tsanas, A., Xifara, A.: Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools. Energy and Buildings 49, 560–567 (2012)
11. Prokop, L., Misak, S., Snasel, V., Platos, J., Kroemer, P.: Supervised learning of photovoltaic power plant output prediction models. Neural Network World 23(4), 321–338 (2013)
12. Moore, G.E.: Cramming more components onto integrated circuits. Electronics, 114–117 (April 1965)
13. Gelenbe, E.: Product-Form Queueing Networks with Negative and Positive Customers. Journal of Applied Probability 28(3), 656–663 (1991)
14. Gelenbe, E.: The Spiked Random Neural Network: Nonlinearity, Learning and Approximation. In: Proc. Fifth IEEE International Workshop on Cellular Neural Networks and Their Applications, London, England, pp. 14–19 (April 1998)
15. Rumelhart, D.E., Hinton, G.E., McClelland, J.L.: A general framework for parallel distributed processing. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Computational Models of Cognition and Perception, ch. 2, vol. 1, pp. 45–76. MIT Press, Cambridge (1986)
16. Basterrech, S., Mohamed, S., Rubino, G., Soliman, M.: Levenberg-Marquardt Training Algorithms for Random Neural Networks. Computer Journal 54(1), 125–135 (2011)
17. Likas, A., Stafylopatis, A.: Training the Random Neural Network using Quasi-Newton methods. European Journal of Operational Research 126(2), 331–339 (2000)
18. Basterrech, S., Zjavka, L., Prokop, L., Mišák, S.: Irradiance prediction using echo state queueing networks and differential polynomial neural networks. In: IEEE Intelligent Systems Design and Applications (ISDA), Hanoi, Veit Nam (December 2013)
19. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks 5(2), 157–166 (1994)

Part VIII

**Multimedia Signal Processing
and Classification**

Feature Line-Based Local Discriminant Analysis for Image Feature Extraction

Jeng-Shyang Pan[1], Shu-Chuan Chu[2], and Lijun Yan[1],[*]

[1] Harbin Institute of Technology Shenzhen Graduate School
Xili University Town, NanShan, Shenzhen, China
[2] School of Computer Science,Engineering and Mathematics
Flinders University of South Australia
GPO Box 2100, Adelaide, South Australia 5001, Australia
yanlijun@126.com

Abstract. In this paper, a novel image feature extraction algorithm, entitled Feature Line-based Local Discriminant Analysis (FLLDA), is proposed. FLLDA is a subspace learning algorithm based on Feature Line (FL) metric. FL metric is used for the evaluation of the local within-class scatter and local between class scatter in the proposed FLLDA approach. The Experimental results on COIL20 image database confirm the effectiveness of the proposed algorithm.

Keywords: Feature Extraction, Image Classification, Nearest Feature Line.

1 Introduction

Feature extraction is an important step of image classification, including face recognition [6], palmprint recognition [3] and so on, becomes more and more popular. Feature extraction procedure can improve the effectiveness and efficiency of the classification systems. If the classification procedure is performed on the images directly, the recognition accuracy rate will be low due to the some small interclass scatter and the large intraclass scatter. The feature extraction procedure is in order to depress the intraclass scatter and increase the interclass scatter, and thus improves the effectiveness of the systems. Some dimensionality reduction based feature extraction algorithms also reduce the dimensionality of the samples, so it also can improve the efficiency of the classification system for the classification procedure.

Feature extraction procedure also can improve the robustness of the classifier for small sample size (SSS) classification task. Images usually reside in a space with thousands dimension. Nevertheless, the number of the prototype image samples from each class is commonly smaller than 100, and even only one in some special cases. This leads to the famous SSS problem. That the classifier uses so few sample for the high dimensional image samples for training may result

---

[*] Corresponding author.

in the system can not generalize well and lack of the robustness. Dimensionality reduction based feature extraction algorithms can reduce the dimensionality of the image samples significantly and handle these problems.

In recent years, many popular feature extraction methods based on dimensionality reduction are proposed. This kind of approaches consider the m by D image samples as a point in $m \times D$ dimensional space, and find the feature extraction algorithm for these point. Dimensionality reduction based feature extraction methods can extract more discriminant features of the image samples and get higher classification accuracy rate. What is more, dimensionality reduction based feature extraction methods can reduce the dimensionality of the image samples, so it is more efficient for classification.

Subspace learning algorithms are a popular dimensionality reduction based feature extraction algorithms. A lot of different subspace learning methods are proposed for feature extraction. Principal Component Analysis (PCA)[13, 12], Linear Discriminant Analysis (LDA)[1, 9] are two most well-known subspace learning-based dimensionality reduction algorithms. PCA projects the original samples to a low dimensional subspace, which is generated by the eigenvectors corresponding to the largest eigenvalues of the covariance matrix of all training samples. PCA aims at minimizing the mean squared error. However, PCA is an unsupervised algorithm, which may reduce the efficiency of feature extraction. LDA is to find a optimal transformation matrix U that linearly projects the high-dimensional sample $x \in R^D$ to low-dimension space by $y = U^T x \in R^d$, where $d \ll D$. LDA can compute an optimal discriminant projection by maximizing the ratio of the trace of the between-class scatter matrix to that of the within-class scatter matrix. LDA uses the labels of the prototype samples during the training and improves the discriminant ability. However, LDA has to suffer from the famous SSS problem. Many effective algorithms have been introduced to solve the problem. Local Preserving Projection (LPP) [2] is to preserve the neighborhood of the samples. Some nonlinear extensions using kernel trick of these algorithms are proposed in the recent 10 years[15, 4, 5]. Besides, there are some works on feature fusion presented[14] to improve the performance of feature extraction.

The above mentioned algorithms are all based on Euclidean distance. Recently, many algorithms based on Nearest Feature Line (NFL) [7] are proposed. NFL is a classifier proposed by Li et. al in 1999. Its kernel is Feature Line (FL) metric. Zheng et al. proposed a nearest neighbour line nonparametric discriminant analysis (NNL-NDA)[17] algorithm, Pang et al. presented a nearest feature line-based space (NFLS)[11] method, and Lu et al. put forward an uncorrelated discriminant nearest feature line analysis (UDNFLA)[8]. Neighborhood discriminant nearest feature line analysis (NDNFLA)[16] is proposed to extract the local discriminant features of prototype samples. In these methods, FL metric is used to evaluate the scatter of the samples. In this paper, A novel Feature Line-based Local Discriminant Analysis (FLLDA) is proposed for image feature extraction.

The paper is organized as follows: Section II gives some preliminaries. The new subspace learning method algorithm is proposed in Section III with the various experimental results in Section IV. Lastly, Section V concludes the paper.

2   Preliminaries

2.1   LDA

Suppose there are c pattern classes. N is the total number of training samples, and N_i is the number of ith class. In the class, the jth training image is denoted by $X_j^i \in R^D$. $\bar{X}^i$ is the mean matrix of training samples of the ith class. And $\bar{X}$ is the mean matrix of all training samples.

The between-class scatter matrix S_b and within-class scatter matrix S_w can be constructed by

$$S_b = \frac{1}{N} \sum_{i=1}^{c} N_i (\bar{X}^i - \bar{X})(\bar{X}^i - \bar{X})^T, \tag{1}$$

and

$$S_w = \frac{1}{N} \sum_{i=1}^{c} \sum_{j=1}^{N_i} (X_j^i - \bar{X}^i)(X_j^i - \bar{X}^i)^T. \tag{2}$$

By LDA, sample I will be projected to the space R^d with $d \ll D$. The criterion to select the most discriminative features can be defined by

$$\max J(u) = \left(u^T S_b u\right) / \left(u^T S_w u\right). \tag{3}$$

In fact, it is clear that the optimal projection matrix U is a set of generalized eigenvectors of S_b and S_w corresponding to the first d biggest generalized eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$. Let $Y = IU$. The resulting matrix Y is called the feature matrix of image I and used to represent X for classification.

2.2   NFL

Nearest feature line is a classifier. It is first presented by Stan Z. Li and Juwei Lu. Given a training samples set, $X = \{x_n \in R^M : n = 1, 2, \cdots, N\}$, denote the class label of x_i by $l(x_i)$, the training samples sharing the same class label with x_i by $P(i)$, and the training samples with different label with x_i by $R(i)$. NFL generalizes each pair of prototype feature points belonging to the same class: $\{x_m, x_n\}$ by a linear function $L_{m,n}$, which is called the feature line. The line $L_{m,n}$ is expressed by the span $L_{m,n} = sp(x_m, x_n)$. The query x_i is projected onto $L_{m,n}$ as a point $x_{m.n}^i$. This projection can be computed as

$$x_{m,n}^i = x_m + t(x_n - x_m) \tag{4}$$

where $t = [(x_i - x_n)(x_m - x_n)]/[(x_m - x_n)^T(x_m - x_n)]$.

The Euclidean distance of x_i and $x_{m,n}^i$ is termed as FL distance. The less the FL distance is, the more probability that x_i belongs to the same class as x_m and x_n. Fig. 1 shows a sample of FL distance. In Fig. 1, the distance between y_p and the feature line $L_{1,2}$ equals to the distance between y_q and y_p, where y_p is the projection point of y_q to the feature line $L_{1,2}$.
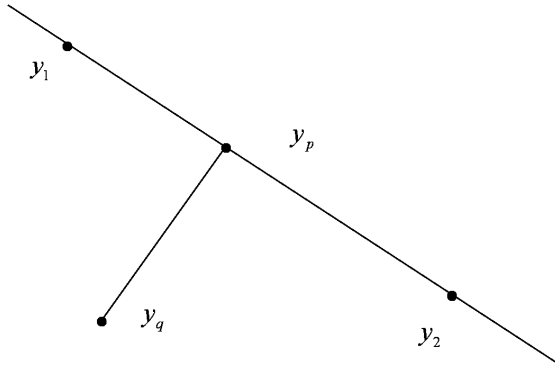

Fig. 1. NFL metric

2.3   LPP

LPP is a popular and effective manifold learning algorithm. It can preserve the locality feature of the sample structure. Given a training sample set $X = x_1, x_2, ..., x_n$, LPP aims to find an optimal transform W by solving the following optimization problem:

$$
\begin{aligned}
W_t &= \arg\min_W \sum_{i \neq j} \|y_i - y_j\|^2 S_{i,j} \\
&= \arg\min_W \sum_{i \neq j} \|W^T y_i - W^T y_j\|^2 S_{i,j} \\
&= \arg\min_W tr(W^T X L X^T W),
\end{aligned}
\tag{5}
$$

where $L = D - S$ is a Laplacian matrix, D is a Diagonal matrix in which its elements on the diagonal are the column sums of the similarity matrix S.

3   The Proposed Algorithm

Given a prototype samples set $X = x_1, x_2, ..., x_n$, let's give two definitions firstly:

Definition 1 k-Nearest Homogeneous Feature Lines: For a sample x_i, its k-Nearest Homogeneous Feature Lines N_i^o is the set of k nearest feature lines which are in the same class with x_i.

Definition 2 k-Nearest Heterogeneous Feature Lines: For a sample x_i, its k-Nearest Heterogeneous Feature Lines N_i^e is the set of k nearest feature lines which are not in the same class with x_i.

In this paper, a novel local feature extraction algorithm named Feature Line-based Local Discriminant Analysis (FLLDA) is proposed. FLLDA aims to preserve the local structure of the samples after dimensionality reduction. The optimization function is as follows:

$$\max J(W) = \frac{\sum\limits_{i=1}^{N} \frac{1}{Nk} \sum\limits_{L_{m,n} \in N_i^e} \left\| W^T x_i - W^T x_{m,n}^i \right\|^2}{\sum\limits_{i=1}^{N} \frac{1}{Nk} \sum\limits_{L_{m,n} \in N_i^o} \left\| W^T x_i - W^T x_{m,n}^i \right\|^2} \tag{6}$$

Using matrix computation,

$$
\begin{aligned}
&\sum_{i=1}^{N} \frac{1}{Nk} \sum_{L_{m,n} \in N_i^e} \left\| W^T x_i - W^T x_{m,n}^i \right\|^2 \\
&= \sum_{i=1}^{N} \frac{1}{Nk} \sum_{L_{m,n} \in N_i^e} \mathrm{tr}[W^T (x_i - x_{m,n}^i)(x_i - x_{m,n}^i)^T W] \\
&= \mathrm{tr}\{W^T \sum_{i=1}^{N} \frac{1}{Nk} \sum_{L_{m,n} \in N_i^e} [(x_i - x_{m,n}^i)(x_i - x_{m,n}^i)^T]W\}
\end{aligned}
\tag{7}
$$

where tr denotes the trace of a matrix. Similar with the above,

$$
\begin{aligned}
&\sum_{i=1}^{N} \frac{1}{Nk} \sum_{L_{m,n} \in N_i^o} \left\| W^T x_i - W^T x_{m,n}^i \right\|^2 \\
&= \mathrm{tr}\{W^T \sum_{i=1}^{N} \frac{1}{Nk} \sum_{L_{m,n} \in N_i^o} [(x_i - x_{m,n}^i)(x_i - x_{m,n}^i)^T]W\}
\end{aligned}
\tag{8}
$$

Then the problem becomes

$$\max J(W) = \mathrm{tr}[W^T \mathrm{tr}(A)/\mathrm{tr}(B)W] \tag{9}$$

where

$$A = \sum_{i=1}^{N} \frac{1}{Nk} \sum_{L_{m,n} \in N_i^e} [(x_i - x_{m,n}^i)(x_i - x_{m,n}^i)^T] \tag{10}$$

$$B = \sum_{i=1}^{N} \frac{1}{Nk} \sum_{L_{m,n} \in N_i^o} [(x_i - x_{m,n}^i)(x_i - x_{m,n}^i)^T] \tag{11}$$

A length constraint $w^T w = 1$ is imposed on the proposed FLLDA. Then, the optimal projection W of FLLDA can be obtained by solving the following generalized eigenvalue problem.

$$Aw = \lambda Bw \qquad (12)$$

Let $w_1, w_2, \cdots, w_q$ be the eigenvectors of formula(12) corresponding to the first q biggest eigenvalues ordered according to $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q$. An $M \times q$ transformation matrix $W = [w_1, w_2, \cdots, w_q]$ can be obtained to project each sample $M \times 1$ x_i into a feature vector $q \times 1$ y_i as follows:

$$y_i = W^T x_i, \qquad i = 1, 2, \cdots, N \qquad (13)$$

In the proposed algorithm, matrix A can be viewed as local between-class scatter matrix and matrix A can be viewed as local within-class scatter matrix. Both of scatters are evaluated by FL metric. For classification task, we hope get a smaller within-class scatter and a bigger between-class scatter after the feature extraction.

4 Experimental Results

To evaluate the effectiveness of the proposed FLLDA method, we compare it with PCA, LDA, LPP, UDNFLA, and NFLS on the COIL20 [10] database. COIL20 database is the one of most popular image databases. There are 1400 image samples of 20 individuals in the COIL20 database. Each individual has 72 image samples with the size of 200×200. To reduce the computation complexity, all the image samples in COIL database are cropped to 32×32. PCA is performed on the samples firstly. 97% energy is preserved. During this experiment, 10 images per class are selected randomly for training and the others are for test. Nearest Neighbor classifier is applied for classification. The system ran ten times in order to reduce the variation on each database. Therefore, the average recognition rate (ARR) is used to evaluate the performance of different algorithms. Label 1 gives the ARR of different algorithm on COIL. From the label, we can get the proposed algorithm has the higher ARR than the other classical algorithms.

Table 1. ARR of different algorithms on COIL20 database

Approaches	ARR	Feature dimension
PCA	0.6521	75
LDA	0.6986	19
LPP	0.6758	45
NFLS	0.7016	100
UDNFLA	0.7856	80
The Proposed FLLDA	0.8284	19

Fig. 2. Some image samples in COIL database [10]

5 Conclusion

This paper has proposed a new subspace learning algorithm called Feature Line-based Local Discriminant Analysis (FLLDA) for image feature extraction. FLLDA aims to maximize the local between class scatter and minimize the local within class scatter. Both of the scatters are evaluated by Feature Line metric. In the experiments, the proposed algorithm has been compared with some popular methods. The experimental results on COIL20 database show the effectiveness of GGDA.

References

[1] Belhumenur, P., Hepanha, J., Kriegman, D.: Eigenfaces vs. Fisherface: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis Machine and Intelligence 19(7), 711–720 (1997)

[2] He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.: Face recognition using laplacianface. IEEE Transactions on Pattern Analysis Machine and Intelligence 27(3), 328–340 (2005)

[3] Kong, A., Zhang, D., Kamel, M.: A survey of palmprint recognition. Pattern Recognition 42(7), 1408–1418 (2009)

[4] Li, J., Gao, H.: Sparse data-dependent kernel principal component analysis based on least squares support vector machine for feature extraction and recognition. Neural Computing and Applications 21(8), 1971–1980 (2012)

[5] Li, J., Pan, J., Chu, S.: Kernel class-wise locality preserving projection. Information Sciences 178(7), 1825–1835 (2008)

[6] Li, J.B., Chu, S.C., Pan, J.S., Jain, L.C.: Multiple viewpoints based overview for face recognition. Journal of Information Hiding and Multimedia Signal Processing 3(4), 352–369 (2012)

[7] Li, S., Lu, J.: Face recognition using the nearest feature line method. IEEE Transactions on Neural Networks 10(2), 439–443 (1999)

[8] Lu, J., Tan, Y.P.: Uncorrelated discriminant nearest feature line analysis for face recognition. IEEE Signal Processing Letters 17(2), 185–188 (2010)

[9] Martinez, A.M., Kak, A.: Pca versus lda. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(2), 228–233 (2001)

[10] Nene, S.A., Nayar, S.K., Murase, H.: Columbia object image library (coil-20). In: Technical Report CUCS-005-96 (1996)

[11] Pang, Y., Yuan, Y., Li, X.: Nearest neighbour line nonparametric discriminant analysis for feature extraction. Electron. Lett. 42(12), 679–680 (2006)

[12] Roweis, S.: Em algorithms for pca and spca. In: Advances in Neural Information Processing Systems, pp. 626–632. MIT Press (1998)

[13] Turk, M., Pentland, A.: Eigenfaces for recognition. J. Cogn. Neurosci. 3(1), 71–86 (1991)

[14] Xu, Y., Zhang, D.: Represent and fuse bimodal biometric images at the feature fevel: Complex-matrix-based fusion scheme. Optical Engineering 49(3), 037002–037002-6(2010)

[15] Xu, Y., Zhang, D., Jin, Z., Li, M., Yang, J.: A fast kernel-based nonlinear discriminant analysis for multi-class problems. Pattern Recognition 39(6), 1026–1033 (2006)

[16] Yan, L., Zheng, W., Chu, S., Roddick, J.: Neighborhood discriminant nearest feature line analysis for face recognition. Journal of Internet Technology 14(1), 344–347 (2013)

[17] Zheng, Y., Yang, J., Yang, J., Wu, X., Jin, Z.: Nearest neighbour line nonparametric discriminant analysis for feature extraction. Electron. Lett. 42(12), 679–680 (2006)

Research and Design of Campus Network VOD System

Kuiliang Xia, Xiaoming Song, and Xiangfeng Suo

Heihe University, Heihe
Heilongjiang, China, 164399
250515679@qq.com

Abstract. VOD (Video on Demand) is referred to as video-on-demand technology, also known as interactive television on-demand system, meaning that the corresponding video playback program according to the user's needs. If the broadband network were highway, then VOD is the most eye-catching car on the road. VOD originally appeared is to better meet user's demand for autonomy to watch video programs, but with the continuous progress of VOD technology, the widely use have produced a strong impact on mass culture and business models, and it has got currently being widespread attention and application of the education sector. In order to meet students' needs and aspirations of real-time learning, in order to better carry out two-way multimedia school teaching, in order to achieve the network management of educational resources, it is imperative to build a campus network-based video-on-demand system.

Keywords: Campus network, video on demand, streaming media.

1   Video on Demand System

1.1   The Work Process of VOD System

When the user starts play request on the client, the request is sent through the network to reach and receipted by the server card to send to the server. After the request is validated, the server prepares the accessible program name in the storage subsystem, so that users can browse the favorite program list [1]. After the user selects a program, the program content is brought from the storage subsystem by the server and transmitted to the client player [2].

1.2   The Characteristics of VOD

(1)It has changed the long-standing one-way propagation characteristics of radio and television, achieved the needs and aspirations of users to actively select the program and control the playback.

   (2) VOD uses MPEG, H.26x and other video compression standards, so that the system has a high data compression ratio while obtaining high-quality picture reproduction characteristics.

(3) The video information after digital processing stored on VOD server will not change the watching effect because of repeat play or passage of time.

(4) In VOD system, it can be achieved that multiple users simultaneously on-demand same program, and independently of each other, not only people can watch for the same program at the same time, but also everyone can control the playback the progress according to their own circumstances.

1.3   The Importance and Significance of the Construction of the Campus Network-Based VOD System

Construction of Campus Network VOD system can improve the school's two-way multimedia teaching system, speeding up the process of network of educational resources, thereby enabling individualized instruction to develop. Campus Network VOD system can retrieve the necessary audio-visual learning material, make it easy to store and manage digital information; schools can conduct video -on-demand network distance education, which can bring more economic and social benefits to schools; video on demand can make the national, provincial quality courses get good utilization and advocacy; and video -on-demand can also be applied to the meeting and academic reports organized by the school, and the important meetings and directives of country on campus webcast . In summary, the construction of VOD system based on the campus network has great significance for the development either of the school, or our society as a whole. A good available campus network VOD system allows school teaching resources and applications to combine, and to make the teaching content fuller, the students' interests in learning more intense, to stimulate the spirit of exploration and innovation of learning, play a good role in promoting teaching and conducting research on the work of the school[3].

2   The Design of Campus Network VOD System

Internet of information age has brought to mankind a great convenience, which has become the main way for people to obtain knowledge and information. Under the influence and lead by such grand environment, to build a campus network VOD system can not only meet the teaching needs, and also achieve the idealized entertaining educational purpose, so that students learn in entertainment, get self-improvement in learning. This is not only needed by the two-way multimedia teaching, but also required by network of educational resources and personalized teaching.

2.1   The Design Requirements of Campus Networks VOD System

First, from the standpoint of safety, we must guarantee the normal operation of the video-on-demand system and make a backup, resolutely obey and implement national

laws, regulations and policies. The contents of pornography and violence are allowed to exist. With considerations of copyright protection, set download limit for the contents in the system. Consideration from the functional aspect of the system, it is required to achieve "full intelligent" on the system management, and achieve online search, on-demand, publishing information and videos and other functions in the system, response is fast, picture is smooth and clear, and the system should be easy to operate, manage and maintain. From the investment considerations, requires the system to minimize investment in human and capital [4].

2.2   The General Design of Campus Network VOD System

Video -on-demand system usually consists of a streaming media server, web server and database server three parts. Streaming media server which is the core server of the entire video -on-demand system is the key platform for operator offering video services to users. Its main function is to collect; cache, schedule and transmission play for media contents. The performance of streaming media application system mainly depends on the performance and quality of the media server. The main function of the streaming media server is transferring the video files as streaming protocol (RTP / RTSP, MMS, RTMP, etc.) to the client for users to watch online; and can also receive live video stream from video capture and compression software, and then live to the client as streaming protocol. Compared to the network player which is in form of watching after downloading, the typical characteristic of streaming media is putting the continuous compression of audio and video information onto a network server, users watch while downloading, without having to wait for the entire file is downloaded. Web server is a type of program residing on the internet. Web server can not only store information, but also to run scripts and programs on the basis of information provided by users on the web browser. When a web browser (client) connected to the server and requests for files, the server processes the request and feedback the file to the browser, with attached information will tell the browser how to view the file (namely file type). The server uses HTTP (Hypertext Transfer Protocol) to exchange information with client browser. The main role of the database server is to complete handling the commands processing of the data, cache, conversion and other work, on the query, adding, deleting and modification sent from web server[5].

The VOD system based on Browser / server architecture, successively to build streaming media server, web server and database server these three servers, and then as the basis to build the entire video-on-demand system, and ultimately achieve the construction of the campus network of video-on-demand system.

2.3   Function Design

The system is divided into two modules: the video on demand and browser subsystem and the video on demand management subsystem. Which consists in the browser

subsystem including video search, video category display, hit movies list, classic movies list, premiere movie listings, bulletin board, video display, video playback and other sub-modules; the management subsystem is divided into video management and exit two sub-modules. Figure 1 shows the specific contents:
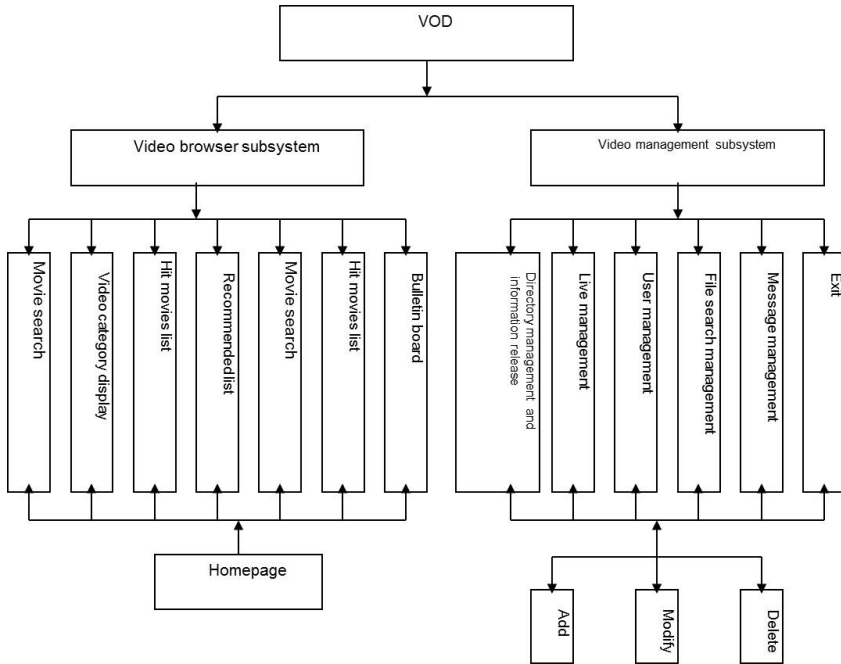

Fig. 1. System function figure

2.4   Database Design

The database design of this system is mainly divided into 5 steps as following.

(1) Overall design database, establish new database, get required data be grouped according to a certain rules and stored in the new database;

(2) To build lists, group according to the type of data, store the data meet the same rule into the same list.

(3) Field establishment. After create lists, then to determine the required fields for each list according to the classification of each data.

(4) Field has been established, to establish relationships between lists based on the primary key in the list.

(5) Finally, in order to facilitate users to quickly find the required data, to optimize the structure of the database.

According to the above steps, first create a database named Movie in SQL Server; create video list, category list and episode list and other lists in the database. Wherein store the desired video data in the video list, shown in Figure 2.

列名	数据类型	长度	允许空
id	int	4	
title	nvarchar	255	✓
player	nvarchar	255	✓
filmlevel	nvarchar	255	✓
runsystem	nvarchar	255	✓
images	nvarchar	255	✓
content	ntext	16	✓
dateandtime	smalldatetime	4	✓
hits	int	4	✓
typeid	int	4	✓
review	ntext	16	✓
reviewcount	int	4	✓
best	bit	1	
[no]	nvarchar	50	✓

Fig. 2. Video list

The data after store videos in the category list in accordance with a certain rule shown in Figure 3

列名	数据类型	长度	允许空
typeid	int	4	
type	nvarchar	50	✓

Fig. 3. category list

The data store the episode number in the episode list shown as Figure 4.

列名	数据类型	长度	允许空
id	int	4	
[no]	nvarchar	244	✓
url	nvarchar	255	✓
filmid	int	4	✓
title	nvarchar	255	✓

Fig. 4. episode list

3   Implementation of Campus Network VOD System

3.1   Three-Tier Architecture System Design

The system uses a three-tier structure for design, namely the presentation layer, business layer and persistence layer.

（1） Presentation layer: the interactive interface of users' operations. The design of this layer plays a vital for the success of the overall system design. The main function of this layer is to receive the user's request and displays the results of the query. As the campus network video-on-demand system is a web site, so the presentation layer is a dynamic page.

（2） Business layer: also known as the business logic layer. Receiving the request sent from the presentation layer, and it is the core structure of the system. The main role of this layer is to achieve the main function of the system, perform the logic processing to the user's request and returns the results to the presentation layer. Apart from the events that the presentation layer can handle by itself, other requests will call the business layer to deal with.

（3） Persistence layer: its main function providing the function that it can store the system data into "permanently no lost" devices and any re-use at any time. In this layer, the code is generated by the NHibemate Synchronizer, basic database operations can be performed directly.

3.2   Implementation of Video-On-Demand Browser Subsystem Function

VOD browser subsystem refers to the interface provided to client for video-on-demand, and users can perform the VOD operation through this interface for information required, to receive and experience the real-time video playback services from streaming media server. Typically, it is only required the client has internet access and multimedia service functions, and download and install the appropriate browser and player then the video-on-demand can be achieved.

(1)Homepage

In this interface, the client can search many different types of classification index information. Users select the video-on-demand content through the provided category, or through the hit movie, premiere movies, classic movies to choose their favorite videos, and users can also extract desired videos directly according to their own interests and hobbies, and can also directly search for videos which are wanted to watch. The videos and classification on its pages is managed by the management tool admin.

(2)Video classification display page

Users choose any one video category on the home page, click on this option, the system upload this request to the business logic layer, then the business logic layer to search the selected category of video, and feed back to the presentation layer, and presentation layer will present the results. This query process shown in Figure 5:
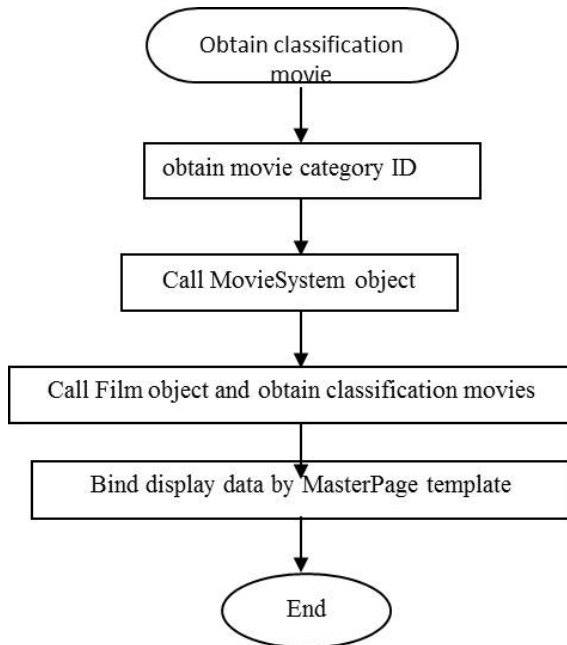
Fig. 5. Video Classification Display Pages Query Process

(3)Video information display page

Users choose one video, click this option, the system uploads this request to the business logic layer, then the business logic layer to search the selected category of video, and feed back to the presentation layer, and presentation layer will present the results.

(4)Video playback page

Streaming media server is responsible for the video playback work, the protocol used is RSTP.

(5)Video search page

Users input the keywords needed to search into the search of the page, and click "Search" option button. First the presentation layer check if the information input by user is complete or not, then the presentation layer uploads the checked complete information to the business logic layer, and business logic layer to find related videos, and feedback the results to the presentation layer, and presentation layer will present the results to the user.

3.3   The Realization of VOD Management Subsystem Functions

The management subsystem includes video information management and exit subsystem in VOD management subsystem. There are three modules in video information management subsystem: add, modify and delete. You can click on the appropriate option to add, modify and delete. Specific methods of operations are shown in Figure 6.
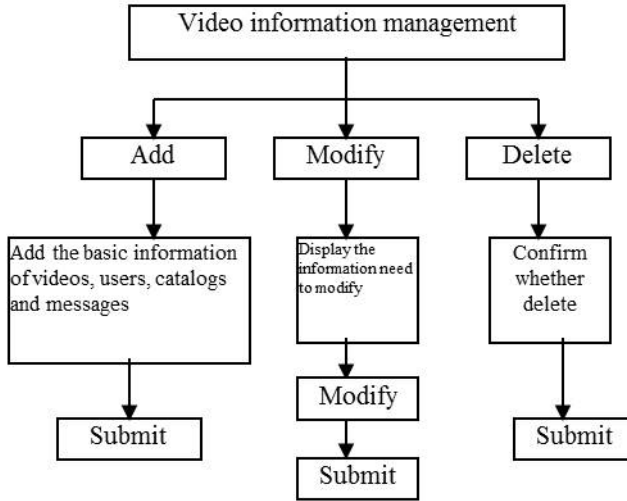
Fig. 6. The video management flowchart

4   System Test

In testing, the current system only considers bandwidth utilization, expand offerings and load, thus simplifying the peripheral part out, but only test play performance. Use eight pcs to build a test environment, place the Web server, database server, and the center on the same machine, but the actual application can be independent. This paper also start three nodes in the same network segment with the hosts in 100M and analog bandwidth of more than 1.5M-demand requests, record the maximum number of requests that the server can handle. This article only lists the average value of several tests of the system performance, as shown in Table 1.

Table 1. The average value in the test

Test Method	maximum number of requests
All the requests for the same programme	1124
All the requests for different programme	32
Request randomly in the 240 programmes	38
90% request concentrate on 20 programme	60

As can be seen from the test results, the bottleneck of disk bandwidth has been significantly improved. In the test of the same program is played for all the requests, the performance of the system increases by times, which is based on the benefits of IP Multicast batch algorithm.

5   Conclusion

Building a reasonable and feasible VOD campus network is a prerequisite and an important guarantee for the construction of school education information. It has a good acceleration on digital teaching mode promotion and distance education launch. The interactive, real-time and dynamic features of VOD bring convenience for students' learning and improve student learning initiative which truly embodies the principle of individualized education, laying the foundation for the development of the real quality-oriented education.

References

1. Viswanathan, S., Imielinski, T.: Metropolitan area video-on-demand service using pramid Broadcasting. Multimedia System 4(4), 179–208 (1996)
2. Cai, Y., Hua, K.A.: Sharing Multicast Videos using Patching Streams. Multimedia Tools and Application 21(2), 125–146 (2003)
3. Casalicchio, E., Tucci, S.: Static and Dynamic Scheduling Algorithms for Scalable Web Server Farm. In: Proc. of 9th Euro micro Workshop on Parallel and Distributed Processing, PDP 2001, Mantova, Italy, vol. (2), pp. 369–376. IEEE Computer Society (2001)
4. Dan, A., Sitaram, D., Shahabuddin, P.: Dynamic Batching Policies for all On-Demand Video Serve. In: Proceedings of ACM Multimedia System, vol. (3), pp. 112–121 (1996)
5. Cai, Y., Hua, K.A.: Optimizing Patching Performance. In: Proceedings of SPIE/ACM Conference on Multimedia Computing and Networking (MMCN 1999), San Jose, CA, vol. (2), pp. 251–260 (1999)

Prediction of Chaotic Time Series of RBF Neural Network Based on Particle Swarm Optimization

Baoxiang Du[1], Wei Xu[2], Bingbing Song[1],
Qun Ding[1,*], and Shu-Chuan Chu[3]

[1] School of Electronic Engineering, Heilongjiang University,
Harbin 150080, China
[2] School of Computer Science and Technology, Heilongjiang University,
Harbin 150080, China
[3] School of Computer Science, Engineering and Mathematics Flinders University,
Australia
dubaoxiang@sina.com, 4788216@qq.com, songbingbing1988@163.com,
qunding@aliyun.com

Abstract. Radial basis function (RBF) neural network has very good performance on prediction of chaotic time series, but the precision of prediction is great affected by embedding dimension and delay time of phase-space reconstruction in the process of predicting. Based on hereinbefore problems, we comprehensive optimize embedding dimension and delay time by particle swarm optimization, to get the optimal values of embedding dimension and delay time in RBF single-step and multi-step prediction models. In addition, we made single step and multi-step prediction to the Lorenz system by this method, the results show that the prediction accuracy of optimized prediction model is obvious improved.

Keywords: radial basis function (RBF), neural network, particle swarm optimization, prediction.

1   Introduction

Chaos phenomenon is a kind of complex behavior generated by deterministic nonlinear dynamics system and widely exists in natural system and social system. It is a irregular movement between deterministic and random, and extreme sensitivity to initial condition. Based on such special characteristic mentioned above, chaos has been widely applied in the secure communications. Because

the chaotic series stems from to deterministic system, it can be predicted in the short-term. But the chaotic series are unpredictable in the long-term for its behavior extremely sensitive to the initial condition which small changes can cause a huge difference to the results. The chaotic series can be predicted in the short-term brought hidden danger to the chaotic encryption systems, while it is very vital significance for the research of chaotic prediction. In recent years, the domestic and foreign scholars have made many achievements in terms of chaotic time series prediction [1,2,3,4,5]. Chaotic time series analysis and prediction are carried out in the phase space reconstructed, so the phase space reconstructed is crucial to chaotic time series analysis and forecasting. The accuracy of prediction is directly influenced by stand or fall of refactoring. Phase space reconstruction method is proposed by Takens [6], the selection of the delay time τ and the embedding dimension m is the key to the phase space reconstruction. At present, there are two primary methods for selection of the two parameters. One think the two parameters are unrelated and the delay time τ and embedding dimension m should be chosen independently. Such as the literature [7], the author adopt the method of mutual information to choose delay time τ, and then using the method of false nearest neighbor domain to determine the embedding dimension m. The other one think delay time τ and embedding dimension m are related. For example, the time window method [8] and $C - C$ method[9] proposed that the two kinds of parameters are integer multiplication relationship. However, the delay time τ and the embedding dimension m are not a simple multiplication relations, there are a lot of inner connection between them which can not be expressed by a simple formula, thus the time window method and $C - C$ method can not effectively predict the time series. Because the particle swarm algorithm can effectively optimize the parameters, this paper put forward joint optimization the delay time τ and embedding dimension m after phase space reconstruction by particle swarm optimization algorithm, then made single step and multi-step prediction to the Lorenz system by means of these optimized parameters.

2   Phase Space Reconstruction Theory

The phase space reconstruction theory was proposed by Packard et al.[10] in 1980, which constitute the theoretical basis of chaotic time series prediction. This paper take the prediction error of chaotic time series as objective optimization function, optimize the delay time τ and embedding dimension m at the same time. The purpose of phase space reconstruction is obtain geometric features in phase space of nonlinear dynamic systems, and recovery the chaotic attractors in the high dimensional phase space. Packard and Takens et al.[6] put forward made phase space reconstruction to one-dimensional chaotic time series

$x(i), i = 1, 2, \cdots, n$ by the delay-coordinate method. The m-dimensional recon-struction state vector can be represented as

$$X(i) = (x(i), x(i+\tau), x(i+2\tau) \cdots, x(i+(m-1)\tau))^T, i = 1, 2, \cdots, N \quad (1)$$

where $N = n - (m-1)\tau$ is the number of phase points in reconstructed phase space, τ is the delay time and m is embedding dimension.

Takens theorem has proved that if the embedding dimension $m \geq 2d + 1$ and d is the dimension of system dynamics, the phase space which composed of primitive state variables is equivalent to the dynamics behavior in the phase space reconstruction of one-dimensional observations. So we can obtain the state of the next moment according to the current state of the system, and then obtain the predicted value of time series in the next moment. The study provide the theory basis for prediction of chaotic time series.

3   RBF Neural Network

Radial basis function (RBF) neural network is a novel effective feed forward neural network. It has the performance of the best approximation and the global optimal , training method is rapid and feasible at the same time and there is no local optimum problem, all these advantages make the RBF network has been widely used in prediction of nonlinear time series.

The RBF network is a kind of three layer forward network: the first layer is input layer which is made up of source nodes. The second layer is hidden layer, and the transformation function of the hidden unit is a kind of local distribution of nonnegative and nonlinear function which radial symmetry and attenuation to the center point. The number of units in hidden layer are determine by the re-quirement of the described problems. The third layer is output layer, the output of the network are the linear weighted of hidden unit output. It is non-linear trans-form from the input space of RBF network to the space of hidden layer, but the transform from the space of hidden layer to the space of output layer is linear.

This paper proposed achieve the mapping of m dimension to one dimensional by two layer before of the RBF neural network, thus establish the prediction model. The model as shown in Fig. 1.The formula is defined as

$$x(n+1) = f(x(n)) = \sum_{i=1}^{m} \omega_i \varphi_i(\|x(n) - c_i\|) = W^T \varphi \quad (2)$$

$$\varphi_i(\|x(n) - c_i\|) = \exp(-\frac{\|x(n) - c_i\|^2}{2\sigma_i^2}), i = 1, 2, \cdots, m \quad (3)$$

where $x(n) \in R^m$ are the input of the network, $f \in R^1$ are the output of the network, $\varphi_i(\bullet)$ is the Gauss function, $\varphi = [\varphi_1, \varphi_2, \cdots \varphi_m]^T$ are the output vector of hidden layer, m are the unit numbers of hidden layer, the c_i and σ_i are center and width of Gauss function respectively, The $W = [\omega_1, \omega_2, \cdots \omega_m]^T$ are weight vector of the output layer.
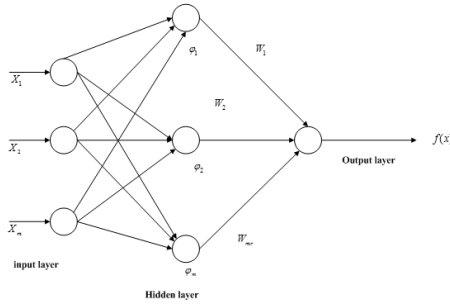
Fig. 1. The structure of RBF network

The Takens' theorem shows that the dynamic system after the phase space reconstruction is topological equivalence to original system on the basis of differential homeomorphism by choose the appropriate delay time and embedding dimension m. There is a smooth nonlinear mapping $f : R^m \rightarrow R^1$ in the refactoring phase space, that is

$$x(i + \tau) = f[x(i)] \tag{4}$$

The chose of delay time τ and embedding dimension m is particularly important to achieve efficient approximation of function for the mapping. We optimization function by take the prediction precision as the objective, and comprehensive optimization τ and m by particle swarm optimization algorithm.

4   Particle Swarm Optimization Algorithm of RBF Neural Network

Particle swarm optimization (PSO) is proposed an adaptive evolutionary that based on population search technique by Kennedy and Eberhart et al in 1995, this algorithm was originally inspired by birds and fish group activities regularity, it use organizational social behavior instead of the evolutionary algorithm mechanism of natural selection, and through the population between individual collaboration to achieve the optimal solution to the problem of search. The PSO algorithm generated an initial swarm and give each particle a random speed. The speed of the particles adjust dynamically by its own and companion's flight experience, the whole swarm have the ability to fly to the better search area[11,12,13,14]. Since the concept of PSO algorithm is simple and easy to achieve, it obtained good performance in aspects such as function optimization and neural network weights training.

4.1   The Basic Principles of PSO Algorithm

Define the global optimization problems (P):$\min\{f(\mathbf{x}) : \mathbf{x} \in \Omega \subset R^n\}, f : \Omega \subset R^n \rightarrow R$, a set of multiple feasible solutions of the problem (P) are known as a

swarm, each element (feasible solution) in a swarm is referred to as a particle. The number of particles is the size of the swarm. We use $\mathbf{X_i} = (x_{i1}, x_{i2}, \cdots, x_{in})^T \in \Omega$ to indicate the position of the ith particle and $\mathbf{V_i} = (v_{i1}, v_{i2}, \cdots, v_{in})^T \in \Omega$ to represent the speed of the ith particle. The $\mathbf{P_i} = (p_{i1}, p_{i2}, \cdots, p_{in})^T$ is the best position of particles have gone through in search space, it is also called pbest. In swarm, we use index symbol g to represent all the best position of particles have gone through, that is $\mathbf{P_g}$ which also called gbest.

Therefore, the velocity of particles in each generation and the location of the evaluation function can be calculated by the following two formulas:

$$v_{id}(t+1) = w \times v_{id}(t) + c_1 \times r_1 \times (p_{id}(t) - x_{id}(t)) + c_2 \times r_2 \times (p_{gd}(t) - x_{id}(t)) \quad (5)$$

$$x_{id}(t+1) = x_{id}(t) + v_{id}(t+1) \quad (6)$$

Where w is the inertia weight, c_1 and c_2 are acceleration constants, r_1 and r_2 are random variables obey uniform distribution within the scope of [0, 1]. The first part of the formula (5) is called momentum part which not only represent the trust of the current own movement state and the ability of space exploration, but also provide a necessary momentum to particle's made inertia motion according to its own speed. The second part is cognitive part which represent the particle's own thinking behavior and encourage particles fly to the best place used to find. The third part is society part which represent sharing information and mutual cooperation between particles and lead particles fly to the best position in particle swarm. The two parts of the last of the formula represent the exploration ability of spatial solution. The main performance of the algorithm are decided by the balance and restrict each other between the three parts.

4.2 The Prediction Algorithm of RBF Neural Network Based on PSO

The basic procedures are as follows:

Step 1: Initialization the parameters such as the size of swarm, the number of iterations, learning factor, the value range of the position and speed. Thereinto the particles and particle velocity are initialized by giving random values to the the position of the particles and the particle velocity.

Step 2: Set the values of the input neurons and the hidden neurons is m which equal to the embedding dimensions of chaotic time series in phase space, set τ is the delay time of phase space, generated a group of chaotic time series randomly and reconstructed the phase space, the initial value of reconstruction are $m = 3$ and $\tau = 1$ respectively.

Step 3: Determine the evaluation parameters of particles, the individual fitness value of algorithm is associated with the prediction ability of model, define the prediction mean square error of the jth parameters as follow

$$MSE(j) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \overset{\wedge}{y_i})^2 \quad (7)$$

where n are sample size of chaotic time series' testing set, y_i and $\overset{\wedge}{y_i}$ are the real and forecast value of chaotic time series respectively. Define the jth fitness for individuals as follow

$$f(j) = \frac{1}{MSE(j)} \tag{8}$$

Step 4: Update the populations constantly, produce the next generation of population if not meet the end conditions, otherwise, output the value of parameter m and τ corresponding to the condition.

Step 5: Substituting the parameter values output by step 4 into RBF neural network to predict chaotic time series.

5   The Experimental Simulation and Result Analysis

5.1   The Experimental Simulation

To illustrate the effectiveness of the algorithm, this paper made the single step and multi-step prediction to Lorenz chaotic time series, and then compared the results between before optimization and after optimization by particle swarm optimization.

The nonlinear Lorenz chaotic system is defined as follow

$$\begin{cases} \frac{dx}{dt} = -a(x-y) \\ \frac{dy}{dt} = -xz + cx - y \\ \frac{dz}{dt} = xy - bz \end{cases} \tag{9}$$

The system (9) in a state of chaos when $c \geq 28$,we chose the parameters $a = 16, b = 4, c = 45.92$ to generated 2000 data of chaotic time series, select the top 1000 data as the training sample and the latter1000 data as test samples used to test the validity of the model. Where the population size of particle swarm are 30, evolutionary algebra are 300,accelerating factor are $c_1 = c_2 = 1.49445$.

This experiment using relative error to evaluate the efficacy of the experiment. The relative error shown as follows:

$$perr = \frac{\sum\limits_{i=1}^{n}(x_i - \overset{\wedge}{x_i})^2}{\sum\limits_{i=1}^{n}(x_i)^2} \tag{10}$$

The experimental data were changed to the time series which the mean is zero and the amplitude is one according to the formula (11).

$$y_i = \frac{x_i - \frac{1}{n}\sum\limits_{i=1}^{n} x_i}{\max(x_i) - \min(x_i)} \tag{11}$$

5.2 The Result Analysis

This paper carried on the single-step prediction to Lorenz chaotic time series by RBF neural network, the prediction results are shown in Fig. 2:
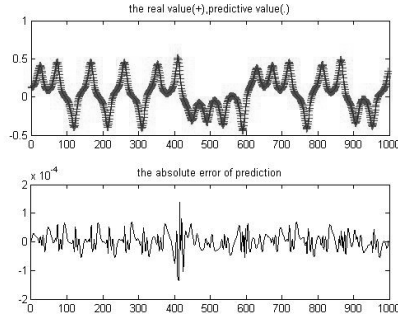

Fig. 2. The single-step predict results of neural network model

According to particle swarm optimization, we made the single-step prediction of particle swarm optimization RBF neural network by chose the parameters $\tau = 4$ and $m = 5$, the results as shown in Fig. 3:
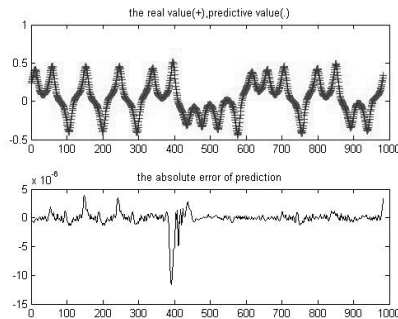

Fig. 3. The single-step prediction of RBF neural network model based on PSO

Fig 3 The single-step prediction of RBF neural network model based on PSO Then we carried on the multi-step prediction to Lorenz chaotic time series by RBF neural network, the 300 prediction results are shown in Fig. 4.

According to particle swarm optimization, we made the 300 multi-step prediction of particle swarm optimization RBF neural network model by chose the parameters $\tau = 4$ and $m = 5$, the results as shown in Fig. 5.
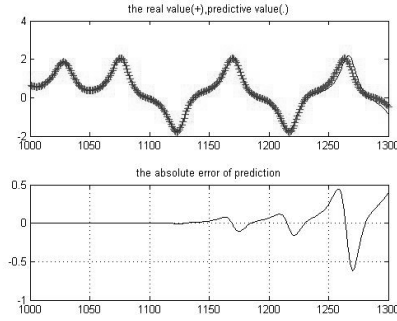
Fig. 4. The 300 prediction results of RBF neural network model
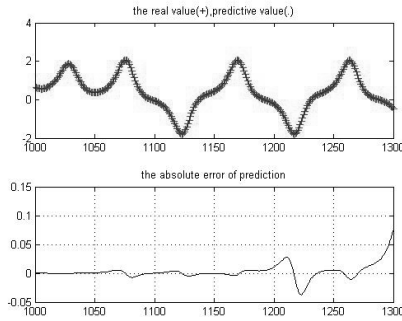

Fig. 5. The 300 multi-step prediction results of RBF neural network model based on PSO

The relative error of single-step and multi-step prediction as shown in table 1:

Table 1. The prediction results of Lorenz chaotic time series by different methods

prediction-step		single-step	multi-step
perr	RBF	7.248E-008	9.9817E-007
	PSORBF	3.4697E-011	6.918E-008

It can be seen from the Fig. 2 to 5 and table 1,the Lorenz chaotic time series can be predicted by RBF neural network forecasting model and RBF neural network based on PSO prediction model. The prediction accuracy of single-prediction-step model is significantly higher than multiple-prediction-step model., which verified the chaotic series can be predicted in the short-term but can not be predicted in the long-term. Compared with RBF neural network prediction model, the prediction accuracy of RBF neural network model based on PSO is significantly higher whether it in single-step prediction or multi-step prediction, which fully demonstrate the validity of RBF neural network prediction model based on PSO.

6   Conclusion

Because the prediction accuracy of RBF neural network chaotic time series prediction model is great influenced by the delay time τ in phase space reconstruction and the embedding dimension m, the paper put forward optimize the parameters of RBF neural network, the delay time τ in phase space reconstruction and the embedding dimension m by particle swarm optimization, then compared with RBF neural network prediction model which didn't be optimized. The experimental results show that this method can effectively improve the prediction precision of the chaotic time series.

References

1. Zhao, Y.P., Zhang, L.Y., Li, D.C., Wang, L.F., Jiang, H.Z.: Chaotic Time Series Prediction Using Filtering Window Based Least Squares Support Vector Regression. Acta. Phys. Sin. 62, 120511-1–120511-9 (2013)
2. Han, M., Xu, M.L.: A Hybrid Prediction Model of Multivariate Chaotic Time Series Based on Error Correction. Acta. Phys. Sin. 62, 120510-1–120510-7 (2013)
3. Yu, Y.H., Song, J.D.: Redundancy-Test-Based Hyper-Parameters Selection Approach for Support Vector Machines to Predict Time Series. Acta. Phys. Sin. 61, 170516-1–170516-13 (2012)
4. Zhang, C.T., Liu, X.F., Xiang, R.Y., Liu, J.K., Guo, J.: Multi-Step-Prediction of Chaotic Time Series Based on Maximized Mutual Information. Control and Decision 27, 941–944 (2012)
5. Arash, M., Majid, A.: Developing a Local Least-Squares Support Vector Machines-Based Neuro-Fuzzy Model for Nonlinear and Chaotic Time Series Prediction. IEEE Transactions on Neural Networks and Learning Systems 24, 207–218 (2013)
6. Takens, F.: Dynamical Systems and Turbulence. Springer, Berlin (1981)
7. Fraser, A.M.: Information and Entropy in Strange Attractors. IEEE Transactions on Information Theory 35, 245–262 (1989)
8. Kugiumtzis, D.: State Space Reconstruction Parameters in the Analysis of Chaotic Time Series-The Role of the Ttime Window Length. Physica D 95, 13–28 (1996)
9. Kim, H.S., Eykholt, R., Salas, J.D.: Nonlinear Dynamics Delay Times and Embedding Windows. Physica D 127, 48–60 (1999)
10. Packard, N.H.: Geom Etry From a Time Series. Phys. Rev. Lett. 45, 712–718 (1980)
11. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: IEEE Int. Conf. on Neural Networks (1995)
12. Shi, Y.H., Eberhart, R.: A Modified Particle Swarm Optimizer. In: Proc. of IEEE Int. Conf. on Evolutionary Computation (1998)
13. Parsopoulos, K.E., Vrahatis, M.N.: On the Computation of All Global Minimizers through Particle swarm Optimization. IEEE Trans. on Evolutionary Computation 8, 211–224 (2004)
14. Trelea, I.C.: The Particle Swarm Optimization Algorithm: Convergence Analysis and Parameter Selection. Information Processing Letters 85, 317–325 (2003)

Segmentation and Description of Human Facial Features Region

Yingli Wang*, Yao Wang, and Song Li

Electronic Engineering College, Heilongjiang University,
Harbin, Heilongjiang, China
midaspeking@sina.com

Abstract. Although there are many search engines based on the text content, that do not meet our extensive search for multimedia content, interests and effectiveness requirements.MPEG-7 is an important criterion for the multimedia content descriptions, so it has a great application value for us to have an in-depth study. Combined with the face recognition algorithm of face regions, feature relationships describing face images content in this paper, according to the study of MPEG-7 standard construction of face related structure models.

Keywords: MPEG-7 description, face structure, facial features division.

1   Edge Detection of Face Image

Extraction of face image is established on the basis of image segmentation. The extraction of face image is based on image segmentation, image segmentation is to separate the containing unique information of regions in the image from the image, each area contains characteristic parameters itself, it is convenient for the dividing area to express, localization and restore.The basic algorithm has many edge detections,such as:gradient operator, direction operator,laplace operator,canny operator and so on [1].

For the features extracted face image can use edge detection methods. The essence of edge detection is to extract the demarcation line between the object and the background in the image, it is particularly important for noise reduction, so we use Canny operator for the edge detection technique.

OpenCV uses the Canny operator to detect the image edge, Canny proposed canny operator in MIT master's thesis in 1986 [2].the function of cvCanny is using canny algorithm to find the edge of the input image and symbol of the edge in the output image, edge detection result is shown in Figure 1. There are two parameters in the cvCanny function, Parameter 1 is small threshold: Control edge connection, Parameter 2 is big threshold: Used to control the initial strong edge segmentation (Usually is 3 times higher than an smaller threshold) [3].

(a) Threshold=1   (b) Threshold=6   (c) Threshold=17   (d) Threshold=34

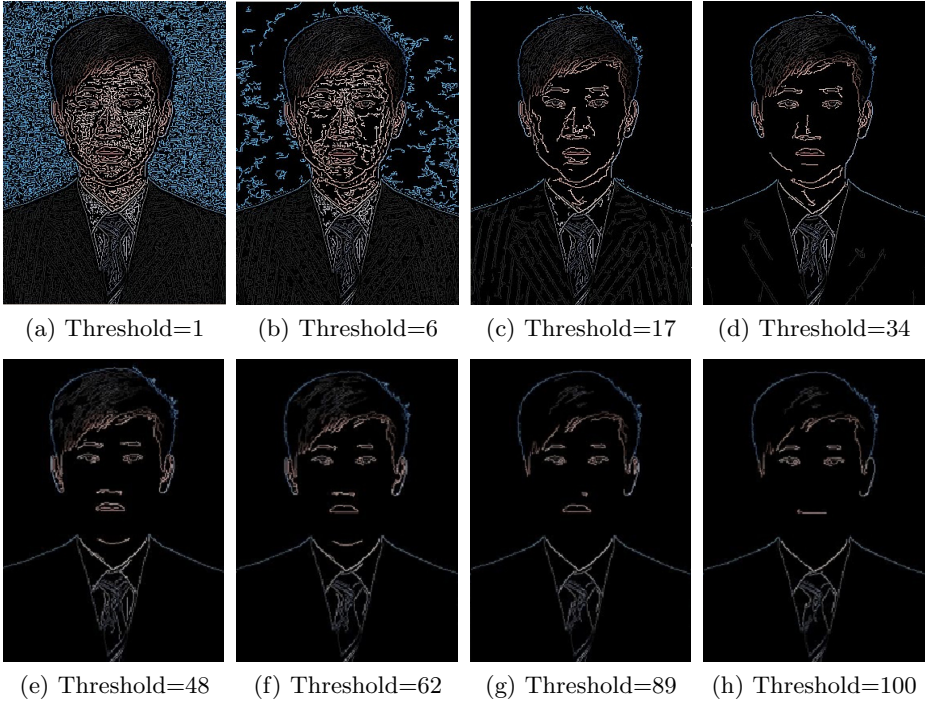(e) Threshold=48   (f) Threshold=62   (g) Threshold=89   (h) Threshold=100

Fig. 1. Canny edge detection with different threshold

As it's shown in Figure 1, it's the result of edge detection of image in different thresholds (from 1 to 100), it's used to determine the edge position through the threshold, the lower the threshold, the more edges can be detected, the greater noise it is influenced. Figure (a) shows the entire image boundaries are drawn, the boundary lines are more likely to find irrelevant features between images, extraction of image contour is important starting point for the start of facial features, but it is difficult to determine the specific contours which contained in the facial characteristics. By contrast, Figure (h) shows a high threshold will be lost thin or short segments [4]. But it can clearly mark the eyes, mouth contour information.

2   Eye Contour Partition

Eye extraction unit locks the concrete existence range of the human eye in the image, using the Canny algorithm to find a specific type of threshold contour of the human eye. Figure 2 shows a flow diagram of a specific, the specific process is shown in figure 2:

(1) Put a piece of original face image into the face location and regional separation module, detecting face through the face/eye training documents in
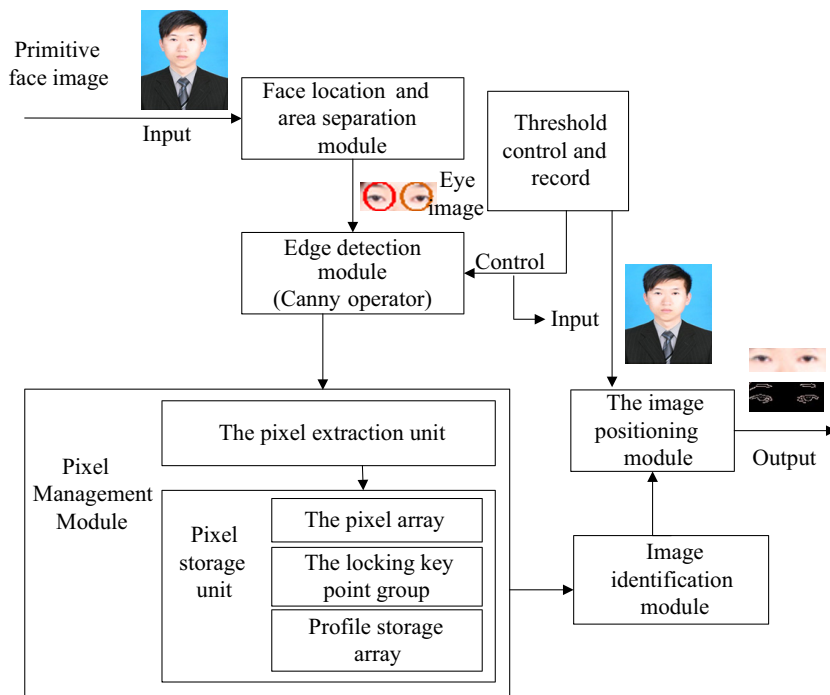
Fig. 2. Extraction processing block diagram of the human eye contour

OpenCV, determining the existence of effective human, and calibrating in interval in face image in the eyes, separating this area, through the general area which has existing human eyes, reducing the complexity of the more difficult to split eye contour image because of lacking the precise positioning [5].

(2) Put the eye image separation into the edge detection module, using the Canny operator to operator the image edge connected degree by different threshold. Each threshold corresponds to a segmented image, putting each pixel management module corresponding to each piece of image edge detection, selecting 100 thresholds value as the selected index in this paper, putting a human face image segmentation of 100 gray images with different edge into the pixel management module, further analysis in its edge contour.

(3) Pixel management module: Mainly consist of the pixel extraction units and the pixel storage units. Pixel storage unit consists of the pixel array, the locking key point group, the profile storage array three parts.

The pixel extraction unit: Mainly extract the coordinates of each segmentation image contains all the pixels and the pixels corresponding to the whole image, and to be marked with the corresponding number.

The pixel array: Contain all the pixels of 100 segmentation image storage.

The locking key point group: Extract the fixed pixel to have a locked storage in the different threshold corresponding to segmentation images.
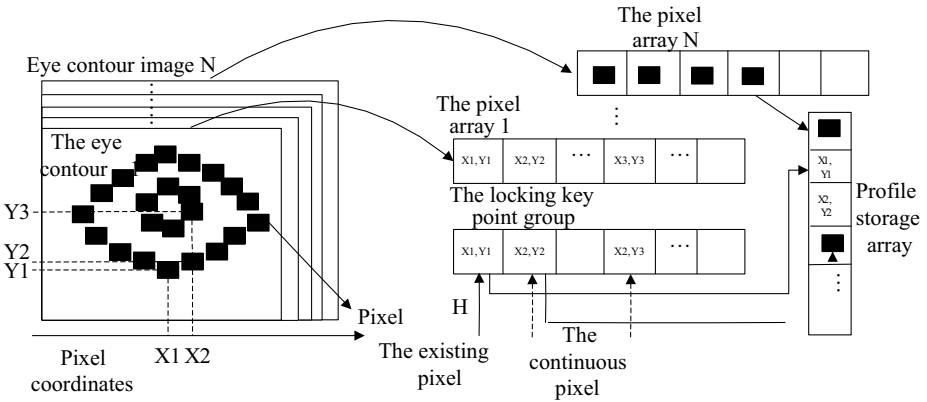
Fig. 3. Pixels management module block diagram

Contour storage array: Store a set of contour points which invariant of surrounding pixels to extract the coordinates for points in the array, In order to maximize the threshold for the benchmark, the principle of main functions are described, as it shown in figure 3.

An image picks up the picture of eye area after the face/eye location, through edge detection, corresponding to different thresholds to extract eye contour images of 1 to N ($N = 100$), putting the N image contour into pixel processing module, the pixel array corresponds to all the pixels in the N profile, as some contours with threshold adjustment become more clear, using different thresholds, comparing the pixel multiple pixel storage array, finding the pixel coordinates (X1, Y1) which always exist, To lock the key points in the storage group, which concluded that the point of eye specific contour points. Similarly through different threshold value to (X1, Y1) for continuous search near the origin (X2, Y2) in the contour corresponding array, the collection contains the outline of an array of points of the specific area of the eye contour [6].

(4) Eye contour point is not present in accordance with the changing of the threshold value in the image (such pixels can be found in the "the group of lock key points"), starting with these points, tracking edge by on the image, to find the pixel point continuous place. As the contour line of human, and has been recording edge position and stop record until the value is smaller than the lower, putting the processed image data into image identification module, according to the pixel points near the color regions of coherent, judging profile in an array of point is the eye contour, reducing the differences brought about by the image contour incomplete judgment.

(5) Original image positioning module: Through the above method to extract clear eye contour, it is accurate and clear in the original image positioning of the human eye, separating images. Through this process, it can reduce the processing of edge detection complexity, quickly and accurately extract the eye contour. Human eyes detection in the process of extracting the

inaccuracy will bring the eyes and eyebrows locked by the same, to find the human eye contour,eye center positioning differences, in order to solve this problem can help by the aid of combining color judgment, the eye contour area near the border color and color has a different color of the eyebrows by means of this method, it can easily rule out eyebrows contour was wrong judges for the possibility of the eye.

3   Partitioning for Eye

In this paper, it is based on adaboost algorithm to eyes as the specific research object, invoking CascadeClassifier kinds in the OpenCV, using its member functions to complete detection, it's own eye training files (haarcascade_eye.xml) that can detect the eye area well, and then separate it. Under the combination of VC and OpenCV to implement the following method.

```
String nestedCascadeName = "./haarcascade_eye.xml";
//The human eye training files
CascadeClassifier cascade, nestedCascade;
//Create a cascade classifier
IplImage* image = cvLoadImage(filename, 1);//load picture
CvSeq* faces = cvHaarDetectObjects();//Haier feature detection
circle(img, center, radius, color, 3, 8, 0);
// In the eyes with circular mark
```

In the OpenCV, if there is a face obscured the main parts or in the case of not straight to the human eyes, it is the basic detection not to come out, because of their own eye database training is not completely good enough, instead of training intact classifier is a very complicated and very large project, this is also the important defects of the algorithm.

4   Other Parts of Facial Contour Extraction

Geometrical characteristics of the human face is fixed, below the eye is the nose, below the nose is the mouth, depending on the physical location of these organs, the core of the above methods, to the basis of to provide complete eye contour, as the core of the midpoint of the eyes of the whole face, under the premise of the eye contour determined, the eyes beneath the center position as the mouth, nose candidate region. Threshold corresponding to multiple different segmented image data to detect the lower eye contour unchanged few pixels, as a person's mouth and nose fixed point, at the same time the highest point of the detected image contours designated as head profile, and find the edge of a continuous contour, positioned in the original judgment, the completion of the divided regions. Because of the effects of light, camera angles and other aspects, the key point of the nose and face are not always present, point to outline the difficulties caused, this is also the limitations of methods exist to solve major problems in the future.

5   Face Structure MPEG-7 Descriptors

In the index system, it may be the position of the corresponding color according to the specific features as auxiliary information in order to achieve a combination of characteristics, it called complementary frequency domain and airspace, the combination of features is called that based on the extracted features, determining a coordinate position to form a description of the associated face of diverse characteristics. Finally, it meets the index of the interface features generated by a combination of MPEG-7 description file.

MEPG is an international standard developed by the Committee referred to MPEG, it is under International Organization for Standardization ISO / IEC [7]. MPEG-7 standard content,you can be described by more unpredictable information, descriptor associated with the query image and its content also plays a vital role. MPEG-7 descriptors into the color, texture, contour, movement and distribution positions, it can be used to identify people's facial technology. Users can check their interesting graphic images in the database according to their needs. The client sends a query request, the main program will be based on a pre-selected index from the XML database query to the corresponding conditions in XML files, which files contain keywords and corresponding with the corresponding images, in order to find the correct image is returned to the client.

This paper-based MPEG-7 standard face image description and retrieval program that can achieve a standardized description of the face and fast retrieval [8]. System Development Platform: WindowsXP, system development tools: Visual C++, MPEG-7 library, xmlspy, SQL2005 database.

6   Indexing System Query Mechanism

The main program will image processing procedures: Firstly, according to the user's input text annotations as the advanced query semantics, as a high level of query conditions. Secondly, procedures of image reprocessing, face detection, facial orientation, coordinates extraction and so on, graphic feature extracting, the corresponding eigenvalue associated with all images and local information, And sent to the feature information associated repository. Thirdly, combines the corresponding characteristics and input conditions according to MPEG - 7 standard that can describe the image content description text files, and sent to an XML database, these XML files can be image retrieval system calls, the whole system application block diagram as shown in figure4, as shown in the central part of the system is important for the image feature extraction algorithm for the automatic and dynamic libraries for MPEG-7 description file parsing and generation.

7   Retrieval Program

System for all image preprocessing, filter out noise. As there is in the process of load the picture, the Non-face/face figure, but this design is mainly on face
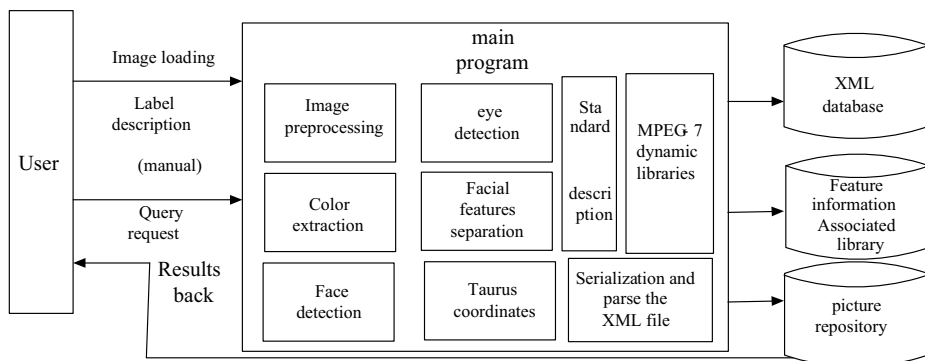
Fig. 4. System Application Block Diagram

image content description and queries, so it should pass face image detection programs (Adaboost algorithm), the distinction between Facial features and all the images stored in the image library, after the determination of face image, color image are extracted respectively (histogram), eye detection (training files), facial features extraction (contour detection), coordinates positioning (such as pixel location) algorithm procedure, Color, contour extraction, location information such as establishing the corresponding relationship each other. Application of the underlying use JOANNEUM RESEARCH developed MPEG-7 Dynamic Link Library, described by the key on the library to provide visual descriptors describe the program established to form a unified standard XML file, so as to achieve the purpose of parsing and serialization. In the process the query and load information to the database as the core, the establishment of information centers to provide similar hub station, the picture information, description file storage and correspond to each other.

In order to verify the applicability of the proposed structural model of the face, the paper development index query face image based on the MFC VC++6.0 on. Configure the environment as follows:

(1) Mp7Jrs2.2 dynamic link library on MPEG-7: means for generating and iden-
    tifying description, description scheme.
(2) xerces-c-src2_1_0 for parsing and serializing XML description file.

Figure 5 shows a query to people face after running the program for the MFC resulting image, the main function of the program is that the user will want to query image to the specified file into the next (This selection of c: \\ Face), when you click on the search button, Program will deal newly added image processing algorithms and XML documents generated, and the file and the file folder in accordance with the conditions of the original XML queries one by one comparison, which returned to the user the correct query results. Program can achieve facial features, size, color, semantic query in four ways.
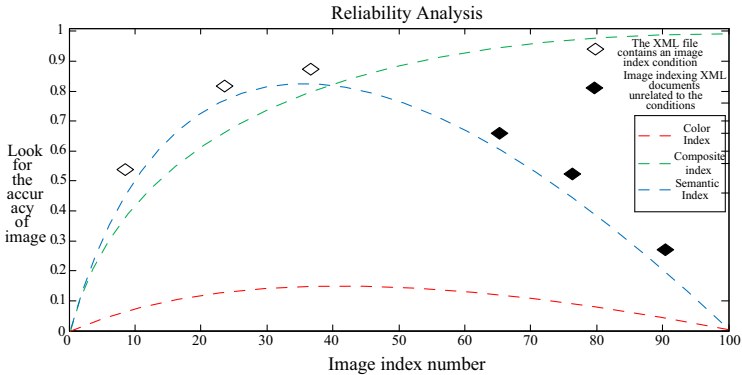
Fig. 5. MFC application

Fig. 6. The index system performance analysis

8   Performance Analysis of Experimental Results

For an image retrieval system, the accuracy of the query is evaluated quality of the system performance. Index model used in this paper are based on the color corresponding to the features, advanced semantics, the combination of conditions

in several ways, and the results of the experiment 100 times, recording each experiment data. And return the system to measure the accuracy of the results you want to find a face image through the actual data.

As can be seen from Figure 6, when using a combination of search criteria, the highest accuracy of the results, but it's computational complexity is the highest. When using the semantic query, the accuracy rate will be based on different conditions and consequent changes in the index, the color accuracy of the three most queries, because it is a condition limiting the least.

References

1. Ma, Y., Zhang, Z.H.: Comparison of Several Edge Detection Operators. J. Computer and Engineering Applications 46, 202–204 (2010)
2. Canny, J.: A computational approach to edge detection. J. Industry and Mine Automation 32, 54–56 (2004)
3. Wang, Z.C., Liu, X.D., Xue, L.X.: An Improved Method of Canny Edge Detection Operator. J. IEEE Transactions on Pattern Analysis and Machine Intelligence 8, 679–698 (1986)
4. Gourier, N., Hall, D., Crowlery, J.L.: Estimating Face Orientation from Robust Detection of Salient Facial Features. Visual Observation of Deicti Gestures, Cambridge (2004)
5. Sun, J.D., Zhao, S.: Image Low-level Feature Extraction and Retrieval Technology. Electronic Industry Press, Beijing (2009)
6. Tong, X.J., Cui, M.G.: Image Encryption Scheme Based on 3D Baker with Dynamical Compound Chaotic Sequence Cipher Generator. J. Signal Processing 89, 480–491 (2009)
7. Tang, Y., Li, G.H., Ni, N.: Multimedia Content Description Based on MPEG-7C. J. Computer Engineering Science 25, 46–49 (2003)
8. Wei, Y., He, Y.W., Ni, H.F., Zhang, W.: Shape Description Method in the Application of Image Retrieval. J. Systems Engineering and Electronics 31, 1755–1762 (2009)

Design and Implementation in Image Compression Encryption of Digital Chaos Based on MATLAB[*]

Zhiqiang Li, Xiaoxin Sun, and Qun Ding[**]

Electronic Engineering, Heilongjiang University, Harbin, Heilongjiang, China
`xiaoqiang-123000@163.com, sxxszcb@126.com, qunding@aliyun.com`

Abstract. A chaotic encryption algorithm used digital image compression and encoding technology based on discrete cosine transform and discrete wavelet transform is proposed in this paper. By taking the redundancy of images and the shortcomings of human visual into consideration, this passage conducts the original image compression firstly, and then uses the discrete Logistic chaotic sequence to achieve image encryption and transmission. From the theoretical analysis and experimental results: this compression encryption algorithm is feasible and it decreases the redundant information of the image, furthermore, it reduces the storage space and improves the efficiency of data transmission, computer simulations proved the security of the image processing method.

Keywords: DCT transform, discrete wavelet transform, compression, Logistic mapping, encryption, MATLAB.

1 Introduction

With the fast development of computer science and multimedia technology, digital images have been widely applied into human lives. However, the number of Internet crime has increased rapidly, thus the secure transmission of digital images is becoming an important research field in communication area. At the same time, there are lots of redundancy information exists in images that limit data storage and transmission severely [1,2]. In order to avoid these two problems, paper proposed an image compression encryption method used discrete cosine transform and discrete wavelet transform, focusing the Logistic algorithm to achieve image encryption. Despite the wide variety of wavelets, people select different wavelets to get different image compression effects. Compared with the wavelet compression method, the DCT compression method is a common used standard, which is more convenient and efficient. Therefore, we select DCT

transform to compress the original images and then conduct Logistic encryption processes. Experimental results show that the compressed image is more clearly and the security of image transmission is stronger. The compression encryption method has high compression efficiency and easy to operated.

2 MATLAB Implementation of Image Compression Encoding Technology

2.1 DCT Transform

Fourier transform (FT transform) and discrete cosine transform (DCT transform) are widely used in digital image processing. The FT conversion has a significant drawback that it uses complex to represent a parameter, thus storage space will be two times than the real number. However, the DCT transform not only has the same function with FT transform but also needs fewer data, which is suitable in the image processing field [3].

DCT transform uses different amplitude and sine function with different frequency to describe image approximately, and a dedicated function (dct2 function) in MATLAB 7.1 Image Processing Toolbox is used to calculate the two-dimensional image of the DCT. In a general digital image, most of its visual information is concentrated on a few DCT coefficients. The special nature of DCT transform makes it applied in JPEG encoding area (the current international standard). The detail steps of doing a DCT transform is to make large quantization intervals on the high-frequency coefficients and small quantization intervals on the low-frequency coefficients of DCT [4,5]. The two-dimensional discrete cosine transform of a $M \times N$ matrix A is defined as following.

Transform:

$$F(u,v) = c(u) c(v) \frac{2}{N} \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} f(x,y) \cos\left(\frac{2x+1}{2N} u\pi\right) \cos\left(\frac{2y+1}{2N} v\pi\right) \quad (1)$$

Inverse transform:

$$f(x,y) = \frac{2}{N} \sum_{u=0}^{N-1} \sum_{v=0}^{N-1} c(u) c(v) F(u,v) \cos\left(\frac{2x+1}{2N} u\pi\right) \cos\left(\frac{2y+1}{2N} v\pi\right) \quad (2)$$

Where

$$x,y,u,v = 0,1,\cdots,N-1, \qquad c(u) = c(v) = \begin{cases} \frac{1}{\sqrt{2}} & u=0, v=0 \\ 1 & u \neq 0, v \neq 0 \end{cases} \quad (3)$$

$F(u,v)$ is the DCT coefficient (also known as the weight applied to each function), $c(u) c(v) F(u,v) \cos \frac{\pi(2x+1)u}{2M} \cos \frac{\pi(2y+1)v}{2N}$ is the basic function of discrete function.

JPEG encoding algorithm uses the 8×8 matrix to do the two-dimensional discrete cosine transform, as follow:

$$F\left(u,v\right) = \frac{1}{4}c\left(u\right)c\left(v\right)\sum_{x=0}^{7}\sum_{y=0}^{7}f\left(x,y\right)\cos\left(\frac{2x+1}{16}u\pi\right)\cos\left(\frac{2y+1}{16}v\pi\right) \quad (4)$$

Where

$$x,y,u,v = 0,1,\cdots,7, \qquad c(u) = c(v) = \begin{cases} \frac{1}{\sqrt{2}} & u=0, v=0 \\ 1 & u \neq 0, v \neq 0 \end{cases} \quad (5)$$

In MATLAB 7.1 Image Processing Toolbox, function dct2 and idct2 are used for the two-dimensional DCT transform and inverse transform respectively. The article uses huiwenlou.jpg image to illustrate the two-dimensional DCT transform and inverse transform in MATLAB platform. The results are shown in Fig. 1.
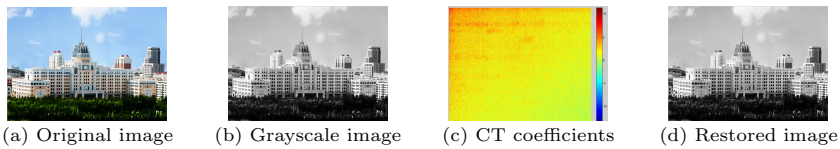

(a) Original image   (b) Grayscale image   (c) CT coefficients   (d) Restored image

Fig. 1. Cosine transform and inverse transform

2.2 Simulation and Analysis of Image Compression Based on DCT Transform

DCT transform method is the key element in JPEG compression algorithm. The basic principle of DCT compression is that: Firstly, decompose the original image into 8×8 or 16×16 digital image blocks. Secondly, do the two-dimensional DCT transform for each block. Finally, encode and transmit the quantized DCT coefficients and build up the compressed image formats. In the receiving side, the system will decode the quantized DCT coefficients, and do the two-dimensional inverse DCT transform for each 8×8 or 16×16 blocks. The last operation will be completed after these divided blocks composing of a single re-image. Fig. 2 shows the principle of image compression based on DCT transform.
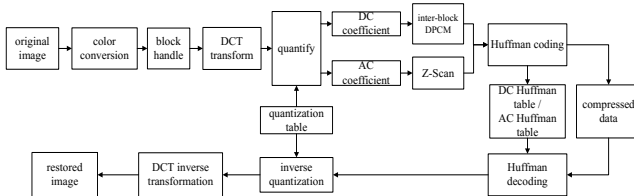

Fig. 2. Principle of image compression based on DCT transform

No matter whatever the image is, the majority of its DCT coefficient values are very closing to zero. If discarding these zero values , there is no significant deterioration to the quality of reconstructed image. Therefore, the use of DCT image compression method can save a lot of storage space. Soon after, this article will divide the size of 512×512 huiwenlou.jpg image into 8×8 blocks, calculating its DCT coefficients, and only retain 10 coefficients from 64 DCT coefficients. Then we use these 10 coefficients to do the inverse transform in order to reconstruct the compressed image. In this experiment, we build up MATLAB simulation platform and get the compressed image as well as other relevant results. The results are shown in Fig. 3.
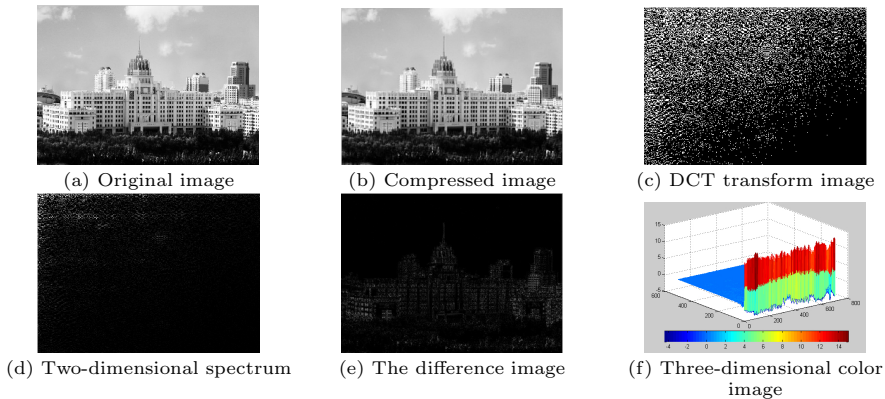

(a) Original image   (b) Compressed image   (c) DCT transform image

(d) Two-dimensional spectrum   (e) The difference image   (f) Three-dimensional color image

Fig. 3. Compression results based on DCT transform

The experiment results are summarized as follows:

1. Program running time is 4.726774 seconds.

2. When choose 10 coefficients from the 64 coefficients, the compression ratio is 6.4:1.

3. Mean square error is 0.0031.

4. Peak Signal to Noise Ratio is 73.1632 dB.

2.3   Wavelet Transform

From above, we have gain a little understand about the JPEG compression algorithm, wherever, we would like to talk about the JPEG 2000 algorithm. The core idea of JPEG 2000 compression algorithm is the wavelet transform. Wavelet transform is a time-frequency localization analysis method, which has the fixed window size and variable shapes. Wavelet transform extracts the useful signal information through two ways: one is getting high quality time resolution from the high frequency portion of the signal, the other is achieving better frequency resolution from low frequency portion of the signal [6,7].

Suppose the function is $f(t)$, and it is a square integral function which meets the condition $f(t) \in L^2(R)$, the definition of continuous wavelet transform is defined as follows:

$$WT_f(a, b) = \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} f(t)\psi^* \left(\frac{t-b}{a}\right) dt, a \neq 0 \tag{6}$$

Where, $\frac{1}{\sqrt{|a|}}\psi^* \left(\frac{t-b}{a}\right) = \psi_{a,b}(t)$ is the displacement and scale expansion volume generated by mother wavelet, a is the scale parameter, b is the translation parameter.

If we perform scale and displacement on the continuous wavelet transform in accordance with powers of two, we will get the discrete wavelet transform (also called binary wavelet transform). The definition of discrete wavelet transform is as follow:

$$W_k[f(x)] = \frac{1}{2^k} \int_{-\infty}^{+\infty} f(t)\psi^* \left(\frac{x-t}{2^k}\right) dt \tag{7}$$

Where, $\psi(t)$ is the mother wavelet.
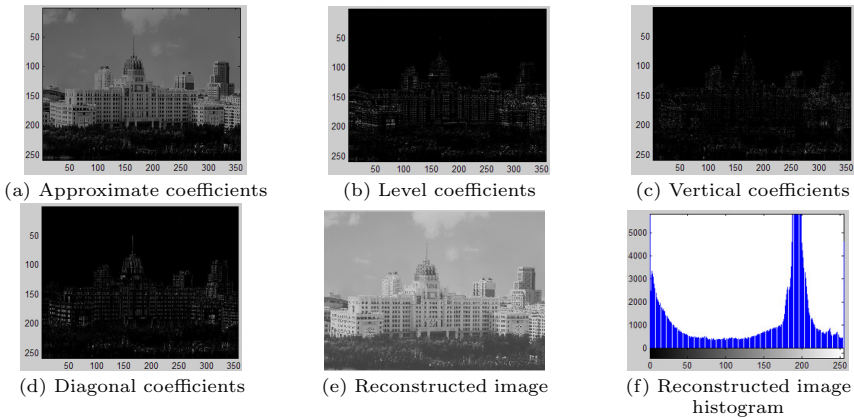

(a) Approximate coefficients   (b) Level coefficients   (c) Vertical coefficients

(d) Diagonal coefficients   (e) Reconstructed image   (f) Reconstructed image histogram

Fig. 4. Wavelet decomposition

Wavelet transform has overcome the shortcoming of Fourier transform, and provides a time–frequency humane window to users. Wavelet transform extracts the low-frequency information from the approximate weight and abandons the high-frequency information until users get the desired image compression effects. The results are shown in Fig. 4.

2.4 Simulation and Analysis of Image Compression Based on Wavelet Transform

The important information of digital images is concentrating on the upper left. By means of the decomposition technique layer by layer, the wavelet compression method can extract useful information from the image. By using the MATLAB

platform, the image compression process mainly includes obtaining the compressed threshold and completing the image compression. This article uses the two-dimensional wavelet transform to finish huiwenlou.bmp image compression. The simulation results are shown in Fig. 5.
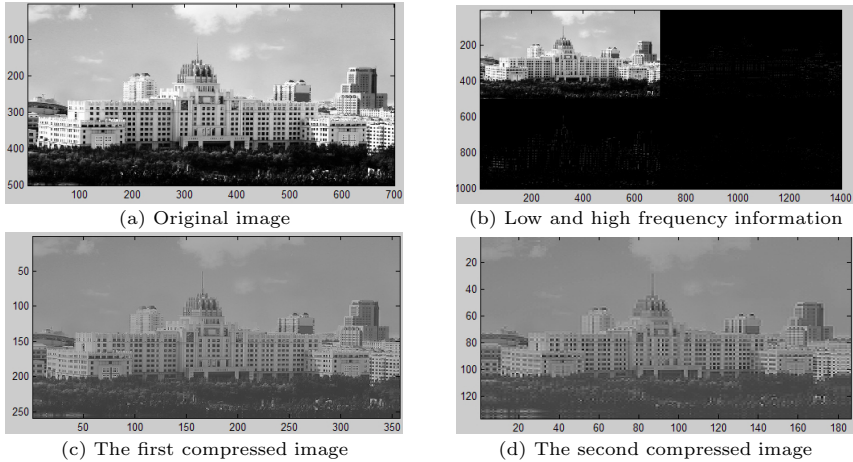

(a) Original image

(b) Low and high frequency information

(c) The first compressed image

(d) The second compressed image

Fig. 5. Compression result based on wavelet transform

The experiment results of the wavelet transform is summarized as follows:

1. The zero percentage of wavelet coefficients is about 47.3740%.

2. The percentage of the remaining energy of the compressed is nearly 99.7997%.

3. The first compression ratio is 15.23:1. And the second compression ratio is 55.46:1.

4. The MSE value before and after image compression is 0.2296.

5. The peak signal to noise ratio (PSNR) value before and after image compression is 54.5202 dB.

6. The program running time is 2.509440 s.

2.5   The Contrast between DCT Compression and Wavelet Compression

This article compared the DCT compression method with wavelet compression. We find DCT-based image compression technology has more advantages: First and foremost, DCT is an orthogonal transformation and it can convert the image from the spatial domain to the frequency domain. Only a small number of data points also can contain the main message of the image. Meanwhile, the whole compression process of DCT is simple and easy. Secondly, by taking a subjective view, we find the DCT compression ratio quite well and the difference between recovery image and the original image are not obvious. Last but not least, the mean square error of DCT compression is much smaller than the wavelet compression and the peak signal to noise ratio is much larger than wavelet compres-

sion. From above, it is easy for us to understand the great advantages of DCT compression method.

3 MATLAB Implementation of Chaotic Image Compression Encryption Technology

3.1 Chaotic Image Compression Encryption Scheme Design

Chaotic image encryption method takes advantage of the extreme sensitivity of chaotic dynamical system to the initial conditions. At the same time, the chaotic signals have excellent hidden features, unpredictable and low complexity characteristics [8,9]. These characteristics are consistent with secret communication system.

The principle of image compression encryption algorithm is summarized as follow: First, we input the DCT compressed image to the encryption system and use the key X to generate the real values chaotic sequence Y and Z; Second, using Y and Z to generate the grayscale matrix G and the symbol matrix S that is matching with the compressed image; Finally, the XOR is needed to operate on the gray matrix and the image compression, and their results WG will dot with the symbols matrix S again[10,11,12]. Only by following these steps can we get the compression encryption images. The schematic of the design is shown in Fig. 6.

After JPEG compression process, the huiwenlou.bmp image will be input into the Logistic encryption system. In chaotic image encryption experiments, a random key$m(1) = 0.11$is selected and we can get the compressed encrypted image information.
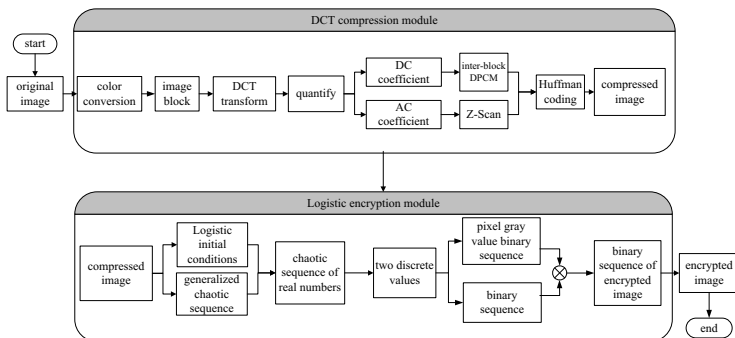

Fig. 6. The schematic of image compression encryption

In order to get a more intuitive view about the images, this article uses the histogram analysis method. Since a good encryption system should be sensitive to the plaintext as well as the keys [13], the experiment chooses the correct key$m(1) = 0.11$and error key$m(1) = 0.13$ to do the tests. Due the image encryption key are matched with the decryption key that changing the decryption key will lead to a failure. The results are shown in Fig. 7.
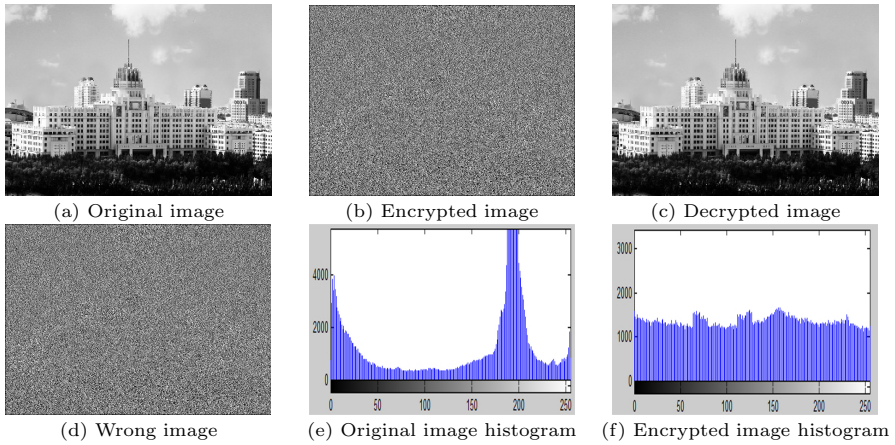
(a) Original image   (b) Encrypted image   (c) Decrypted image

(d) Wrong image   (e) Original image histogram   (f) Encrypted image histogram

Fig. 7. Chaotic image encryption and histogram analysis

From the simulation results we know the DCT compressed image would not bring any significant distortion to the image, and the restored image can better reflected its original appearance. Then after Logistic system encryption, the outline of compressed image has been covered completely, and no one is able to distinguish the details of compressed image. The scheme design is accurate and its performance meets the needs of users absolutely.

From the histogram (e) and (f), we can see the distribution of original image histogram is sharp and steep, but the distribution of encrypted image histogram is more evenly soothing. This result shows that the safety of the encryption system is high enough.

3.2   Anti-cropped Test on the Chaotic Image

The article use the imcrop function to extract a rectangular portion from the encrypted image, then using the key $m(1) = 0.11$ to decrypt the cropped image. The simulation result is shown in Fig. 8.
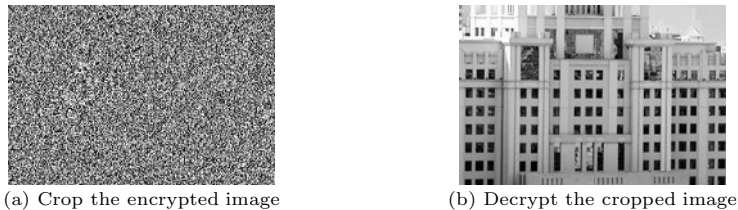

(a) Crop the encrypted image   (b) Decrypt the cropped image

Fig. 8. Anti-cropped test on the chaotic image

3.3 Anti-rotation Test on the Chaotic Image

This article operates the rotation processing on encrypted image (respectively rotated 35° and 145°), then using the key $m(1) = 0.11$ to decrypt the rotated image. The simulation result is shown in Fig. 9.

It can be seen from Fig. 9 that we restore the original image completely after entering the correct decryption key, and the changing of pixel relative position on the encrypted image has no harmful effects on our studies. Though performing different degrees of rotation on the image, we can get the desired results. Therefore, this algorithm has a good anti-rotation performance.
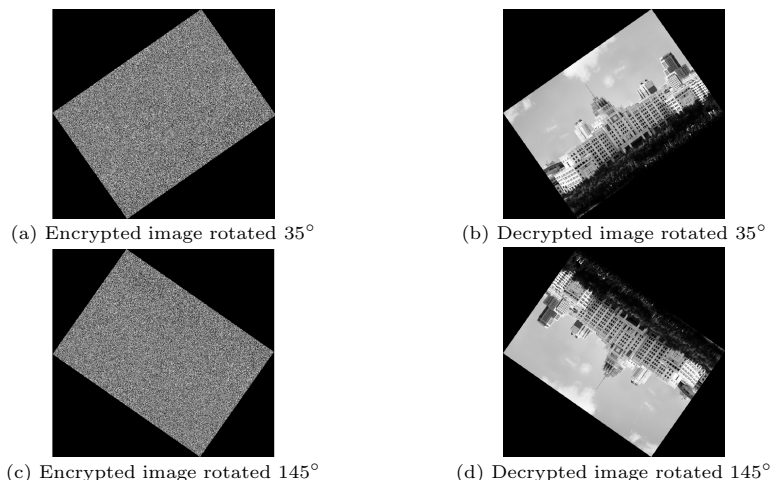

(a) Encrypted image rotated 35°

(b) Decrypted image rotated 35°

(c) Encrypted image rotated 145°

(d) Decrypted image rotated 145°

Fig. 9. Anti-rotation test on the chaotic image

4 Conclusion

This paper proposed a new chaotic encryption compression algorithm based on DCT transform and discrete wavelet transform techniques. This design not only removes redundant information from the image effectively but also saves storage space. Once the design operated can we get a higher efficiency of data transmission and a more secure network communication environment.

References

1. Lv, J.H.: Research on Image Compression Technology Based on MATLAB. J. Xian Coking Technol. 12, 35–38 (2008)
2. Cao, Y.R., Zheng, J.M.: Implementation of Image Compression Based on MAT-LAB. J. Comput. Eng. Des. 12, 2998–3000 (2009)
3. Milanovic, V., Zaghoul, M.E.: Improved Masking Algorithm for Chaotic Communications Systems. J. Electron. Lett. 32, 11–12 (1996)
4. Chen, Y.H.: Research of Image Compression Based on Discrete Cosine Transform. J. Mod Electron. Tech. 34, 86–88 (2011)

5. Chen, L.: The Realization of Matlab by Image Coding Compression Algorithm on DCT. J. Electron. Des. Eng. 19, 168–170 (2011)
6. Zhang, D.F., Ma, L., Fan, L., Liang, Z.H.: Algorithm Research on Image Compression Technologies with Wavelet Transform. J. Acta Scientiarum Naturalium Universitatis Sunyatseni. 47, 42–45 (2008)
7. Lu, X.Q., Zhang, S.C.: Wavelet Transform Based Image Compression Technique. J. Journal of Baotou University of Iron and Steel Technology 21, 59–63 (2002)
8. Ding, Q., Pang, J., Fang, J.Q., Peng, X.Y.: Designing of Chaotic System Output Sequence Circuit Based on FPGA and Its Possible Applications in Network Encryption Card. J. International Journal of Innovative Computing, Information and Control 3, 449–456 (2007)
9. Li, Z.Q., Sun, X.X., Du, C.B., Ding, Q.: JPEG Algorithm Analysis and Application in Image Compression Encryption of Digital Chaos. In: International Conference on Instrumentation & Measurement, Computer, Communication and Control, pp. 185–189. IEEE Press, Shenyang (2013)
10. Ding, Q., Pan, J., Wang, L., Chen, G.R.: The Cipher Code Parameter Selection and Impact on Output Cycles. In: International Workshop Chaos-Fractals Theories and Applications, pp. 143–147. IEEE Press, Shenyang (2009)
11. Zhang, J.S., Xiao, X.C.: A Kind of Less Parameters Second Order Volterra Used in Chaotic Time Sequence Adaptive Prediction. J. Acta Phys. Sinica. 50, 1248–1254 (2001)
12. Li, Z.Q., Sun, X.X., Du, C.B., Ding, Q.: Hardware Design and Implementation of Wi-Fi Technology Based Encryption System. In: International Conference on Sensor Network Security Technology and Privacy Communication System, pp. 144–147. IEEE Press, Harbin (2013)
13. Geng, X.L., Du, C.B., Ding, Q.: Image Encryption System of New Chaotic Algorithm Based on DSP. In: the Second IEEE International Conference on Instrumentation & Measurement, Computer, Communication and Control, pp. 614–618. IEEE Press, Hangzhou (2012)

Depth Map Coding Based on Arbitrarily Shaped Region

Ruizhen Liu and Anhong Wang

Institute of Digital Media & Communication, Taiyuan University of Science and Technology
{lanshuye_ty,wah_ty}@163.com

Abstract. In 3D video processing systems, efficient depth map coding is vital since the quality of the synthesized virtual views highly depends on the accuracy of the depth map. In this paper, we propose an efficient depth map coding method based on arbitrarily shaped region. A depth map is first divided into some regions along the detected edges, and then the edges are coded without loss using a directional 8-connected chain code and an arithmetic codec while the pixels inside the arbitrarily shaped regions are coded by a Differential Pulse Code Modulation (DPCM) with uniform quantization. Experimental results show that our method can obtain better quality of rendered views than JPEG at the high bitrate.

Keywords: 3-D video, depth map coding, region division, DPCM, uniform quantization.

1   Introduction

Recently, with the development of three dimension (3D) display and interactive multimedia system, people are pursuing the immersive perceive experience of the reality scenes, which leading to the high development of 3D video [1].

For 3D video, multiview video coding (MVC) [2] is the conventional format, which is compressed by H.264/AVC video coding standard. The central idea is to exploit the statistical dependencies between different video sequences and the spatial/temporal correlation in each single sequence. Since this approach brings about large amounts of data to be stored and transmitted, another solution using the data format of multiview video plus depth (MVD) [3] is represented. MVD uses the transmitted video views as well as their corresponding depth maps to generate the intermediate virtual views, where depth map has much lower data than the color texture data. At the decoder side, a number of desired intermediate views can be synthesized from the neighboring viewpoints, via some depth-image-based rendering (DIBR) techniques [4].

Depth maps are characterized by shaped edges and large regions of nearly constant or slowly varying sample values. Most depth maps have the same format as color maps (YUV 8-bit) and thus can be compressed using color map codecs such as H.264/AVC. However, depth map has a number of different characteristics, which makes color map codec-based techniques less efficient for depth map coding. A significant difference is the depth-level distribution. Depth map only represents the

distance between the capturing camera and an object and it has no texture. Therefore, most regions within an object have very similar depth levels. Another difference of depth map is its lower temporal consistency than color map. The last, depth map is only used to generate the virtual view synthesis and not presented to the user [5]. Therefore, efficient depth map coding should not use the traditional compression methods as color maps in the 3DV systems. Nowadays, various approaches have been proposed in attempting to compress depth map. Some approaches focus on moving data redundancy by using the correlation between color and depth sequences. For example, paper [6] argued that depth sequences should share motion vectors with the corresponding color sequences, and [7] more recently proposed a view synthesis prediction method for depth video coding. Others focus on the rendering quality of synthesized views since the main function of depth data in 3DV systems is to assist the synthesis of high quality virtual viewpoint views [8][9]. The depth quality is greatly restricting the quality of synthetic viewpoint in 3DV system [10]. However, in the forementioned approaches, the depth map is compressed without considering the characteristics of edges, which will consequently affect the synthesized quality.

In this paper, we propose a depth map coding based on arbitrarily shaped region. A depth map is first divided into some regions along the detected edges, and then the edges are coded without loss using a directional 8-connected chain code and an arithmetic codec while the pixels inside the arbitrarily shaped regions are coded by a DPCM with uniform quantization. Experimental results confirm that our method can obtain better quality of rendered views especially at high bitrate.

The outline of the paper is as follows. The proposed method is described in Section 2 in detail, and Section 3 presents the coding performances and experimental results of the proposed method. Finally, we conclude this paper in Section 4.

2   The Proposed Method

Fig.1 shows the flowchart of our proposed method, which consists of the encoder and decoder.
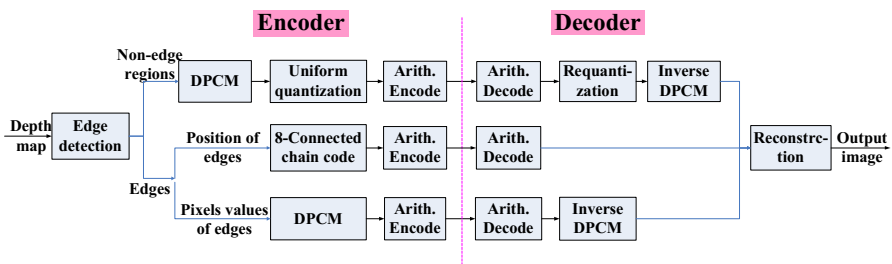

Fig. 1. Flowchart of our proposed method

2.1   Edge Detection

In this paper, Canny edge detector [11] is to detect a given depth map's edges, including the following steps:

- Gaussian smoothing,
- image gradients computing,
- non-maximum suppression to every gradient vector,
- two-threshold hysteresis edge testing.

Here, the low hysteresis threshold is 0.4 times of the high threshold. As shown in Fig.2, first, for the depth map (Fig.2.(a)), when the gradient value is less than the low hysteresis threshold, we can get the corresponding gray values of the depth map and set them to 0, and then we obtain the image 1 (Fig.2.(b)). In the same way, when the gradient value is less than the high threshold, we can get the image 2. Because the threshold value of image 2 (Fig.2.(c)) is higher, and it not only removes most of the noise, but may also lose the useful edge information, based on this observation, we add the edges of the depth map according to the image 1.
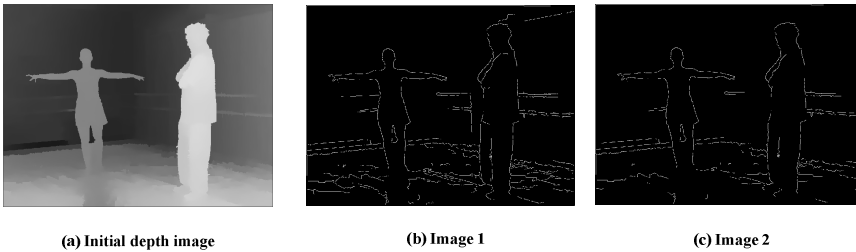

(a) Initial depth image   (b) Image 1   (c) Image 2

Fig. 2. (a) Initial depth image of "Ballet", (b) Image 1 from the low hysteresis threshold, (c) Image 2 from the high hysteresis threshold

2.2   8-Connected Chain Coding the Position of Edges

The depth map contains no texture but sharp object edges and smooth areas inside each edge. The depth plays an important role on the synthesized virtual views since the distortion of depth data especially around the object edges will lead to geometry changes and occlusion variations of the texture in the view synthesized. Hence, in order to preserve a lossless edge, a lossless coding is performed for the position of edges.

Fig. 3 shows an illustration of detected edges in a 8×8 depth block. Here, the positions 1 represent the edges and the positions 0 represent the non-edge regions. After a directional 8-connected chain coding for describing every continuous edge-pixels, an arithmetic coding is implemented to the set of edges for the entropy coding.
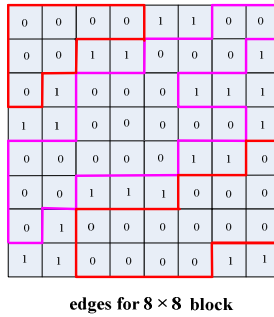
0	0	0	0	1	1	0	0
0	0	1	1	0	0	0	1
0	1	0	0	0	1	1	1
1	1	0	0	0	0	0	1
0	0	0	0	0	1	1	0
0	0	1	1	1	0	0	0
0	1	0	0	0	0	0	0
1	1	0	0	0	0	1	1

edges for 8 × 8 block

Fig. 3. An example for detected edges in a 8×8 depth pixel block

2.3 Coding the Pixels Values of Edges

With the purpose of lossless coding the edges, all pixels values on edges are arranged as a row vector, and then this vector is encoded by DPCM, aiming at removing data correlations. Finally, the results will be inputted to an arithmetic codec. At the decoder side, we obtain all pixels values of the edges with the help of the edge position.

2.4 Scanning and Coding the Non-edge Regions

Since depth maps only express the distance information between the capturing camera and object, pixels in each region inside the same edge have very similar distance and they are smooth. Therefore, for a certain region, there are strong correlations between the adjacent pixels, and we can put all the pixels into a one-dimensional vector for efficient coding.

Firstly, we scan each non-edge region from the first row, when it comes to the border, the scanning just jumps to the next line until the first regional scanning is completed. Next, we regard the coordinates of positions 0 (e.g. Fig.3) that after the edge pixels that we met in the first scan as the starting points of the next scanning, like this, each scanned regions can be of arbitrarily shaped. Do it again and again until all the non-edge pixels have scanned, we'll be able to get more than one non-edge region of arbitrarily shaped.

Secondly, we put all pixels of each region in a row. After doing this, we encode each of one-dimensional vectors by a simple DPCM with the purpose of removing correlations. In the process of DPCM, in order to avoid the spreading of prediction error in DPCM, we conduct a down-sampling.

For example, a one-dimensional error vector has 19 elements, we first take out of the first and tenth components and transmit them accurately to the decode side, while for the rest components, we deal them with a simple DPCM, which is defined as:

$$e(n+1)=e(n+1)-e(n) \tag{1}$$

where e denotes the element in a one-dimensional error vector, n is the position of the corresponding elements in the vector, and where n is a positive integer.

Thirdly, a uniform quantization is performed to the error data after DPCM. Typically, the quantization used in our work can be described as follows:

$$y_q = \text{round}(E \times \text{level}/(y_{max} - y_{min})) \tag{2}$$

where E denotes the one-dimensional error vector with length N after DPCM, and N represents the number of pixels in a certain region (the different value N means that the non-edge region is of arbitrarily shaped), and level is the resolution of the uniform quantization (it can be 8, 16 or 32 and etc.), and y_{max} and y_{min} denote the maximum and minimum values of the vector E, respectively.

At last, after quantization, we put the data into an arithmetic codec for entropy coding.

2.5   The Tradeoff between Edges and Non-edge Regions

The different allocation of bitrate for edges and non-edge regions will influence the final performance of depth map and the synthesized virtual views. Generally, there is a tradeoff between edges and non-edge regions.

When a low threshold is used to detect a given depth map's edges, the number of edges will increases, at the same time, the bitrate of edges also increases. Meanwhile, the resolution of the uniform quantization can also influence the total bitrate, and with the reduction of quantization precision, the bitrate for the non-edge regions is decreasing. Based on this, we have done a lot of experiments to find a tradeoff between edges and non-edge regions.

For example, table 1 shows the distribution of bitrate for edges and non-edge regions for a certain PSNR of the depth map Ballet. Hence, we could find the optimal combination (in case of total bitrate of 0.148BPP) to achieve the best result.

Table 1. The depth map ballet for a certain PSNR

Bitrate of edges(BPP)	Bitrate of non-edge regions(BPP)	The total bitrate of image(BPP)	PSNR of depth map(dB)
0.049	0.09	0.148	30.24
0.056	0.088	0.1562	30.24
0.092	0.066	0.17	30.24
0.082	0.096	0.197	30.24
0.088	0.098	0.2	30.24

3   Experimental Results

The performance of the proposed framework was evaluated using the multiview depth video sequences Kendo, BookArrival, and Ballet (1024×768 @15Hz) provided by Microsoft at the camera position 1, position 8, and position 4.

To confirm the performance of our proposed method, three groups of performance comparisons are demonstrated here: the R-D performance of depth map, the R-D performance of the synthesized virtual views, and comparison of the synthesized virtual views. . In all experiments, we compare our method with JPEG2000 and JPEG

standard under the same conditions. The depth images (Kendo，BookArrival, and Ballet) are tested in the experiment and the Peak Signal-to-Noise Ratio (PSNR) is used to evaluate the distortion.

The R-D performance of depth map is presented in Fig. 4 and PSNR of the coded depth map is plotted against the different accuracy of the quantization. For the given depth images, it is obvious that the depth map compressed by our method is better than JPEG under high BPP. Unfortunately, there is a gap between JPEG2000 and JPEG under the low bitrate conditions. The reason is that the edges information is lossless coded, and it costs a lot of bits, therefore, the total bitrate will goes up. Especially, for image Kendo, when the bitrate is used to compress depth map is 0.79BPP, and the depth map can be lossless recovered.

Fig. 5 compares the R-D performance of the synthesized view among the three schemes, where the synthesized view of the image is generated using the original view and the compressed depth map which are both at the certain position. Fig. 5 shows that our algorithm can obtain better quality of rendered views at high bitrate. This is attributed to the lossless compression of the edges, which is important to the synthesis of virtual viewpoint. Another advantage of our method is that the shapes of non-edge regions are adaptive, unlike JPEG whose block-size is fixed. These adaptive shaped regions satisfy the characteristics of the depth map and can be compressed more efficiently than the fixed block. Consequently, the R-D performance of the synthesized view is better than JPEG at high bitrate.

The synthesized views are presented in Fig. 6, where bitrate used to compress depth maps are respectively 0.51BPP for Kendo, 0.7BPP for BookArrival, and 1.1BPP for Ballet. In the process of DIBR, we use warping and medium filtering, and the hole-filling is not implemented. The perceived quality of the synthesized views using our method performs better than JPEG especially at high bitrate.
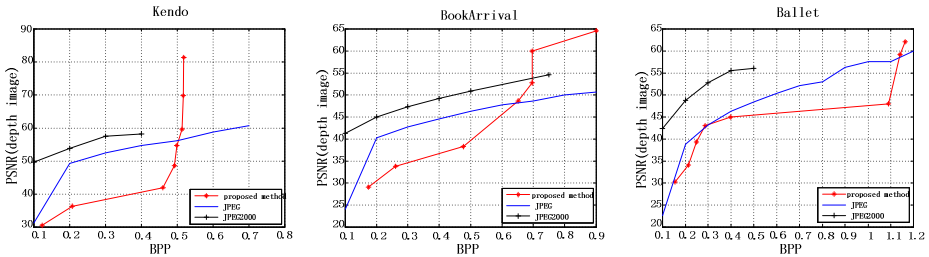

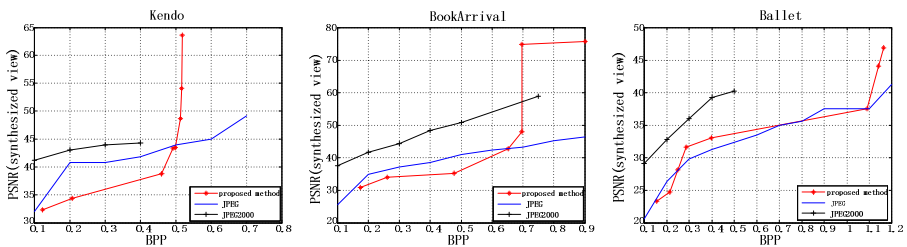**Fig. 4.** The rate-distortion comparison of the depth images

Fig. 5. The rate-distortion comparison of the synthesized view

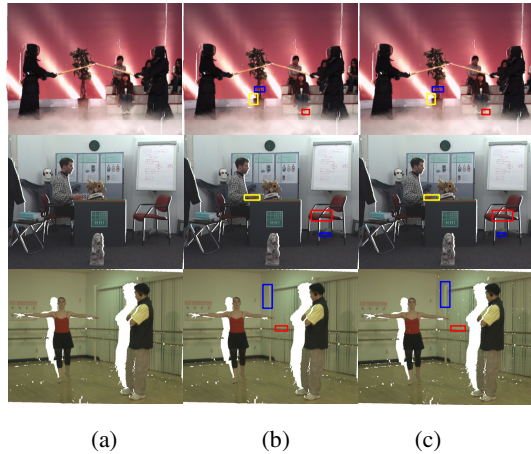(a)   (b)   (c)

Fig. 6. Synthesized views using compressed depth maps with 0.51BPP for *Kendo*, 0.7BPP for *BookArrival*, and 1.1BPP for *Ballet*. From up to down: *Kendo, BookArrival, and Ballet*. (a) ground truth, (b) compressed with our proposed method, (c) compressed with JPEG.

4   Conclusion

In this paper, we propose an efficient depth map coding method based on arbitrarily shaped region. The analyzed characteristics of depth map coding enable us to divide a depth map into arbitrarily-shaped regions along the detected edges, but not the rectangular blocks like that in traditional coding algorithm. The edges are transmitted using a directional 8-connected chain code and an arithmetic codec. The pixels of arbitrarily shaped regions are first encoded by a down-sampling with simple DPCM and uniform quantization. Experimental results show that our proposed method can obtain better quality of rendered views in the case of high bitrate.

However, there are still some issues to be considered in the future, e.g., the rate-distortion optimization for the depth map should be taken into account. The relation between depth distortion and view synthesis distortion needs more investigation, and can achieve the best results.

References

1. Isgrò, F., Trucco, E., Kauff, P., Schreer, O.: Three-dimensional image processing in the future of immersive media. In: 14th IEEE Transactions on Circuits and Systems for Video Technology, pp. 288–303. IEEE Press, New York (2004)

2. Merkle, P., Smolic, A., Muller, K., Wiegand, T.: Efficient prediction structures for multiview video coding. In: 17th IEEE Transactions on Circuits and Systems for Video Technology, pp. 1461–1473. IEEE Press, New York (2007)
3. Merkle, P., Smolic, A., Muller, K., Wiegand, T.: Multi-view video plus depth representation and coding. In: 17th IEEE International Conference on Image Processing, pp. I–201. IEEE Press, New York (2007)
4. Fehn, C.: Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV. In: International Society for Optics and Photonics, pp. 93–104. SPIE Press, USA (2009)
5. Kang, M.K., Ho, Y.S.: Depth video coding using adaptive geometry based intra prediction for 3-d video systems. In: IEEE Transactions on Multimedia, pp. 121–128. IEEE Press, New York (2012)
6. Oh, H., Ho, Y.-S.: H.264-based depth map sequence coding using motion information of corresponding texture video. In: Chang, L.-W., Lie, W.-N. (eds.) PSIVT 2006. LNCS, vol. 4319, pp. 898–907. Springer, Heidelberg (2006)
7. Na, S.T., Oh, K.J., Lee, C., Ho, Y.S.: Multi-view depth video coding using depth view synthesis. In: IEEE International Symposium on Circuits and Systems, pp. 1400–1403. IEEE Press, New York (2008)
8. Zhao, Y., Zhu, C., Chen, Z., Tian, D., Yu, L.: Boundary artifact reduction in view synthesis of 3D video: from perspective of texture-depth alignment. In: IEEE Transactions on Broadcasting, pp. 510–522. IEEE Press, New York (2011)
9. Lai, P., Ortega, A., Dorea, C.C., Yin, P., Gomila, C.: Improving view rendering quality and coding efficiency by suppressing compression artifacts in depth-image coding. In: International Society for Optics and Photonics, pp. 72570O–72570O. SPIE Press, USA (2009)
10. Zhang, J., Hannuksela, M.M., Li, H.: Joint multiview video plus depth coding. In: 17th IEEE International Conference on Image Processing, pp. 2865–2868. IEEE Press, New York (2010)
11. Canny, J.: A computational approach to edge detection. In: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 679–698. IEEE Press, New York (1986)

Emotional Impact on Neurological Characteristics and Human Speech

Pavol Partila[1], Jaromir Tovarek[1], Jaroslav Frnda[1], Miroslav Voznak[1],
Marek Penhaker[2], and Tomáš Peterek[2]

[1] Department of Telecommunications, VSB – Technical University of Ostrava,
17. listopadu 15, 70833 Ostrava, Czech Republic
[2] Department of Cybernetics and Biomedical Engineering,
VSB – Technical University of Ostrava, 17.listopadu 15, 70833 Ostrava, Czech Republic
{pavol.partila,jaromir.tovarek.st,miroslav.voznak,
marek.penhaker,tomas.peterek}@vsb.cz

Abstract. This article discusses impact of human emotions on physiological characteristics and their changes. Many fields require applications that provide information about the emotional state of a human. Today's research is mainly concerned with increasing the accuracy of the methodology for obtaining this information. Studied subjects were psychologically stimulated to change their neutral calm state to stress. Subjects were measured physiological characteristics and the change of speech also. Blood samples, ECG and EEG form part of the neurophysiological data that were collected during the neutral state and during stress. Voice activity was recorded from reading text that read, patients in both emotional state. Features extraction was focused on the Mel-frequency Cepstral coefficients and their dynamic and accelerated derivations. Change in emotional state from neutral to stress was recognized by using a GMM classifier that has been trained and tested by mentioned speech features. Psychological stimulus was induced using professional psychological methods. The measurement was performed in a special EMC interference protected chamber to prevent undesirable electrical influences from the external environment especially on sensitive EEG measurement.

Keywords: Emotional state, cortisol, ECG, EEG, MFCC, GMM.

1 Experiment

The objective of the experiment was to determine the effect of stress on neurophysiological changes both in human brain activity and in responses of the cardiovascular and hormonal systems. The cardiovascular activity was monitored with standard electrocardiography (ECG). To determine hormonal responses, changes in the cortisol level were monitored in blood samples taken both before and after the stress. In order to avoid any undesirable interference, the experiment was performed in an EMC shielded chamber at the Department of Telecommunications of the Czech Technical University in Prague. The experiment was carried out between 7 and 11 am to avoid

circadian rhythms in measured values. 18 healthy subjects (14 men and 4 women) aged between 20 and 22 years took part in the testing. All test subjects agreed with their participation in the experiment by signing an informed consent. EEG was monitored with 19 EEG leads placed according to the internationally recognised 10/20 system. The limit value of contact resistance between the electrode and skin was below 5 KΩ. ECG was monitored via a single-channel bipolar lead. A g. USBamp (g.Tech, Graz, Austria) biological signal amplifier was used to measure biological signals. Blood pressure values were recorded using the oscillometric method with an Omron M6 blood pressure meter. Cortisol levels were determined from blood samples analysed in a certified biochemical laboratory of the University Hospital in Ostrava.

1.1   Experiment Description

After a subject arrived to the workplace, both EEG and ECG leads were attached and a blood pressure cuff was put on. After that a certified nurse introduced a cannula to take blood samples. The subject then moved to the test chamber. Prior to starting the measuring, blood pressure and pulse were taken both before and after the relaxation phase. A relaxation phase followed during which a series of relaxing photographs were shown and an accompanying narration recorded by a psychologist was played. The presentation included pictures of beaches, animals and children. After the last photograph had been shown, the test subjects were asked to close their eyes and focus only on their thoughts. This stage lasted 8 minutes and its purpose was to evoke sleep. At the end, the test subjects were asked to open their eyes to make sure no sleep symptoms would be registered in the EEG signal. After 30 seconds, the test subjects were asked to close their eyes again in order to eliminate the artefact of blinking. At that stage, the actual EEG and ECG measuring began and lasted for 30 seconds. After that, blood pressure was measured and a blood sample was taken. To determine any emotion in voice, the subject was asked to read a short text. During the second phase of measuring, a series of logical tasks were displayed and the subject was supposed to perform a calculation or answer a question within a time limit. To enhance the stress, each question was assigned a score depending on the difficulty. A subject with the highest score received a reward of 40 euros which is about a 1/20 of the average salary in the Czech Republic. A typical example was to count the number of triangles in a picture. This testing phase lasted for 5 minutes and at the end, the subjects were asked to close their eyes again for 30 seconds during which brain and heart activity was monitored. After that, blood pressure was measured and a blood sample was taken and the subject was asked to read the same short text as in the relaxation phase.

1.2   Experiment Results

The results of the experiment can be seen in Figure 1. Cortisol levels (upper left square) after stress were higher only in 7 subjects. A certain disadvantage of using this stress hormone as a stress indicator is that the time necessary for the hormone to be released from suprarenal glans to circulation is not known. The same figure (upper right corner) shows the values of systolic BP before (green) and after stress exposure

(red); the difference is marked in blue. Out of the total number of 18 test subjects, blood pressure increased in 15 of them and the average increase was 4.5 mmHg with a standard deviation of ± 6.5 mmHg. Diastolic blood pressure increased on average in 11 test subjects by an average value of 1.67 mmHg with a standard deviation of ±7.30 mmHg.
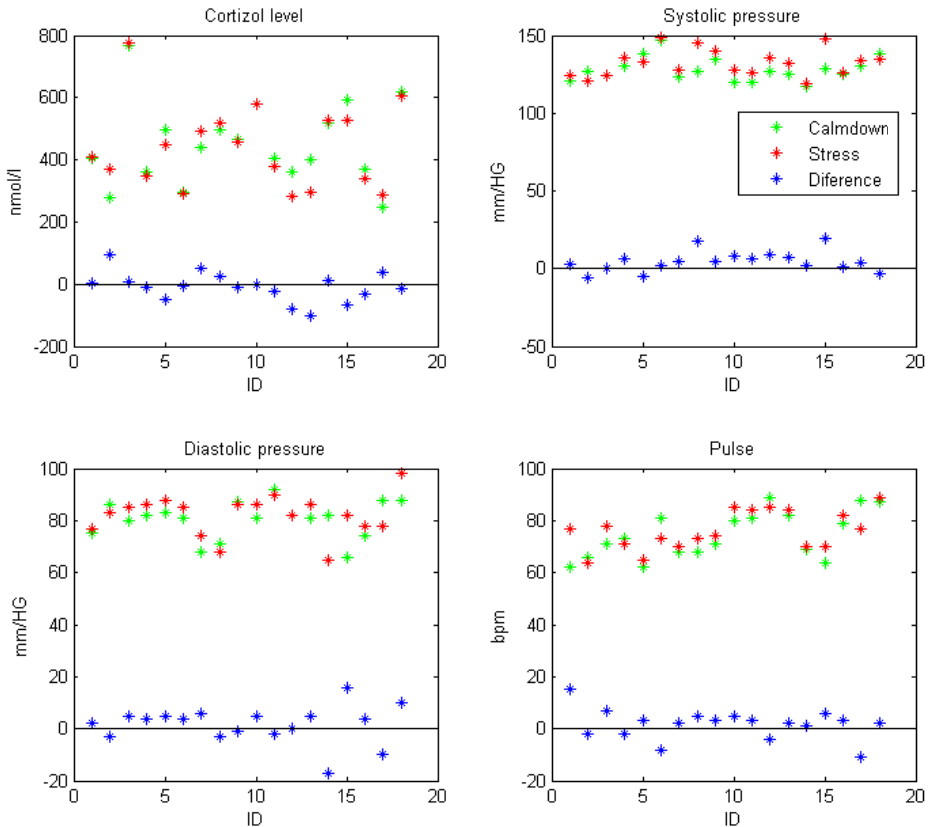

Fig. 1. Experiment results

The fourth monitored parameter was the pulse. This stress indicator increased in 13 subjects by an average of 1.8 bpm and the standard deviation was 5.78 bpm.

2   Stress Recognition from the Human Speech

This chapter deals with the recognition of the stress condition of the patients. Psychologically stimulated patients read it before the selected text. The database was created recordings of their speech in normal and stress conditions. The research compares the impact of emotions stimulated patients given the well-known Berlin emotional database. Gaussian Mixture Model was used to classify the state of stress. [6]

2.1   Features Extraction

For speech recognition, the most commonly used features are cepstral coefficients. Cepstral coefficients are derived from an inverse discrete Fourier transform (IDFT) of logarithm of short-term power spectrum of a speech segment as:

$$c(m) = \sum_{k=0}^{N-1} \ln\left[\left|X(k)\right|\right] e^{\frac{j2\pi mk}{N}} \tag{1}$$

Where X(k) is the FFT-spectrum of speech x(m). As the spectrum of real-valued speech is symmetric, the DFT can be replaced by discrete cosine transform (DCT). To obtain MFCC features, the spectral magnitude of FFT frequency bins are averaged within frequency bands spaced according to the Mel scale which is based on a model of human auditory perception. The scale is approximately linear up to about 1000 Hz and approximates the sensitivity of the human ear as:

$$f_{mel} = 1125\log(0,0016 f + 1) \tag{2}$$

Number 13 is extracted coefficients, c_0 included. For each vector of coefficients were derived dynamic coefficients of the delta Δc_m and delta-delta $\Delta^2 c_m$ (acceleration coefficients), which reflect the temporal changes of coefficients vectors c_m. The final number of coefficients is therefore 39. [1][2][3][4][5]

$$c_m^{all} = \left[c_m, \ \Delta c_m, \ \Delta^2 c_m \right] \tag{3}$$

2.2   GMM Classifier

A Gaussian Mixture Model is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continues measurements or features in biometric system, such as vocal tract, in speaker recognition systems as well. Probability distribution of the parameter vectors derived from human speech can be described using GMM.

$$p\left(o \mid \lambda^s\right) = \sum_{i=1}^{M^s} w_i^s p_i^s(o) \tag{4}$$

Where M is number of components for s class, w_i, i=1,…,M, are weights of components complying condition that sum of all weights is 1, p means the probability density of the components represented by the mean value μ and covariance matrix C_i.

Speaker model can be described mentioned mixture characterized by the equation below.

$$\lambda^s = \left\{ w_i^s, \mu_i^s, C_i^s \right\}, \quad i = 1,…,M^s. \tag{5}$$

The criterion of maximum likelihood is found lambda parameters with maximum p probability density based on sequence parameters $O = \{o_1, o_2, \ldots o_n\}$ obtained from speech, seen below. [7][8]

$$\lambda^s = \arg \max \ p\left(o \mid \lambda^s\right)$$

(6)

2.3 Results

The extracted speech features were training and test data in relation cyrca 50 percent. GMM classifier was trained data from the Berlin database of recordings and stimulated patients. A different number of components of the GMM has been reached various classification accuracy, as shown in Table and Figure bellow.

Table 1. Impact of GMM components number

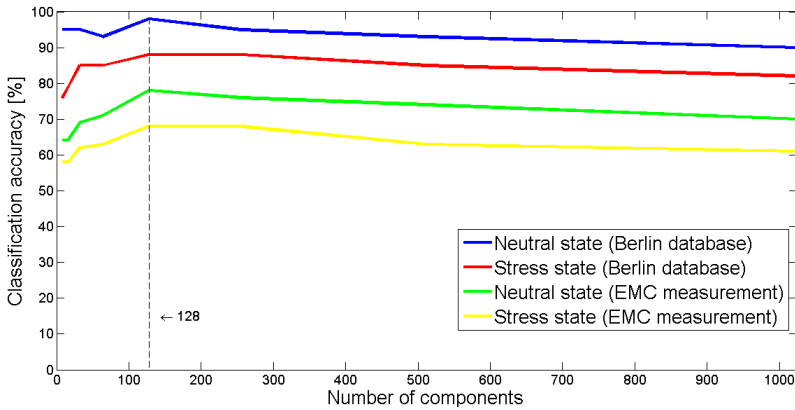| Number of components	8	16	32	64	128	256	512	1024
Stress state (Berlin database) [%]	76	79	85	85	88	88	85	82
Neutral state (Berlin database) [%]	95	95	95	93	98	95	93	90
Stress state (EMC measurement) [%]	58	58	62	63	68	68	63	61
Neutral state (EMC measurement) [%]	64	64	69	71	78	76	74	70

Fig. 2. Dependence of the classification accuracy on the number of GMM components and databases

The best recognition ability was GMM classifier with 128 components. The best success rate reached classifier in recognizing neutral state in the Berlin database. Percentage recognition neutral state, given the stress was 78 percent in the processing of speech patients from EMC chamber. Figure below shows the comparison of the two data sets and classification accuracy for training and testing of this data.
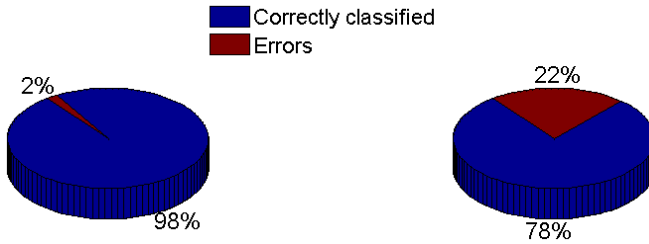
Fig. 3. Percentage of classification accuracy for neutral emotional state recognition. Left pic shows percentage for Berlin EmoDB and right shows accuracy percentage of classification voices from EMC measurements.

2.4   Conclusion

If we compare all physiological and neurophysiological measured values, we will find out that there were only two instances where all of them increased. If we take into account the fact that cortisol secretion speed is not known and cannot be 100% relied on to indicate immediate stress, only the three remaining parameters can be monitored: systolic pressure, diastolic pressure and pulse. All three of them were increased in 6 test subjects.

One of the benefit of this research is that the sound quality has a considerable influence on the extraction features from speech and hence the classification accuracy of GMM classifier. Less speech recognition accuracy with patient's recordings from the EMC chamber is caused by the lower quality of the speech signal. This is due to the fact that in the EMC chamber was not possible to provide studio quality sound, as it was in the creation of the Berlin database. Another reason is that some patients were not affected by the psychological stimulus with equal intensity.

References

1.   Iliou, T., Anagnostopoulos, C.-N., Narayanan, S.: Comparison of Different Classifiers for Emotion Recognition. In: 2009 13th Panhellenic Conference on Informatics, pp. 102–106. IEEE (2009), doi:10.1109/PCI.2009.7
2.   Zulfiqar, A., Muhammad, A., Martinez Enriquez, A.M.: A Speaker Identification System Using MFCC Features with VQ Technique. In: 2009 Third International Symposium on Intelligent Information Technology Application, pp. 115–118. IEEE (2009), doi:10.1109/IITA.2009.420

3. Jayanna, H.S., Mahadeva Prasanna, S.R., Martinez Enriquez, A.M.: Analysis, Feature Extraction, Modeling and Testing Techniques for Speaker Recognition. IETE Technical Review 26(3), 181 (2009), doi:10.4103/0256-4602.50702

4. Hossan, A., Sheeraz, M., Gregory, M.A.: A novel approach for MFCC feature extraction. In: 2010 4th International Conference on Signal Processing and Communication Systems, vol. 26(3), p. 181. IEEE (2010), doi:10.1109/ICSPCS.2010.5709752

5. Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A.: Acoustic emotion recognition: A benchmark comparison of performances. In: 2009 IEEE Workshop on Automatic Speech Recognition, vol. 26(3), pp. 552–557. IEEE (2009), doi:10.1109/ASRU.2009.5372886

6. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A Database of German Emotional Speech. In: Proc. Interspeech, Lisbon, pp. 1517–1520 (2005)

7. Metallinou, B., Lee, S., Narayanan, S., Rigoll, G., Wendemuth, A.: Audio-Visual Emotion Recognition Using Gaussian Mixture Models for Face and Voice: A benchmark comparison of performances. In: 2008 Tenth IEEE International Symposium on Multimedia, vol. 26(3), pp. 250–257. IEEE (2008), doi:10.1109/ISM.2008.40

8. Metallinou, A., Lee, S., Narayanan, S., Rigoll, G., Wendemuth, A.: Audio-Visual Emotion Recognition Using Gaussian Mixture Models for Face and Voice: A benchmark comparison of performances. In: 2008 Tenth IEEE International Symposium on Multimedia, vol. 26(3), pp. 250–257. IEEE (2009), doi:10.1109/ICAPR.2009.89

Part IX

Intelligent Data Analysis and System Reliability Modeling

An Effective Approach Based on Partial Duplication for Reducing Soft Error Rate in SRAM-Based FPGA

Baolong Guo[1], Guochang Zhou[2], Jinfu Wu[1], Xiang Gao[2], and Yunyi Yan[1]

[1] School of Aerospace Science and Technology
Xidian University
Xi'an 710071, China
`blguo@xidian.edu.cn`
[2] China Academy of Space Technology(Xi'an)
Xi'an 710000, China

Abstract. In this paper, we present an effective approach for reducing soft error rate (SER) in SRAM-based FPGA. First, the entire system is divided into several modules according to its function. Then the soft error rate of each module is calculated by an analytical estimation method. Finally, rather than performing mitigation for all the modules in the system to achieve high reliability, the modules with highest soft error rate have a priority to be mitigated, i.e. we perform mitigation for the entire system based on partial duplication. Experimental results verify our proposed method.

1 Introduction

Field Programmable Gate Arrays (FPGAs) are utilized in many applications such as networking and embedded applications due to their high performance, low Non Recurring Engineering (NRE) cost and Fast-Time-To-Market [1]. However, SRAM-based FPGAs are vulnerable to *Single Event Upsets* (SEUs) that cause soft rate [2] and their applications in aerospace field are narrowed. SEUs are generated by cosmic particles, energetic neutrons, and alpha particles hitting the surface of silicon devices. One possible solution to this problem is to use radiation-hardened FPGA devices. These devices, however, are much more expensive than Commercial-Off-The-Shelf (COTS) FPGAs. Thus the COTS devices are affordable when cost, i.e. area and/or power, is a major issue [3]. Moreover, radiation-hardened devices are few generations behind state-of-the-art COTS devices.

To design one high reliable system mapped into an SRAM-based FPGA, it's necessary to study the effect of soft errors at the system level and identify the most vulnerable modules in the system. Conventional methods [4–6] are based on em Fault Injection (FI) strategies. Using this methodology, a limited number of error sites are targeted for fault injection. Several workloads are then run to measure the number of detected failures by comparing the results of each run to

the clean run. These steps make FI approaches both very time-consuming and inaccurate. Moreover, these approaches cannot be used during design phases since they need physical implementation. To overcome the shortcomings aforementioned, Asadi *et al.* [7,8] presented an analytical approach to estimate the soft error rate of systems mapped into SRAM-based FPGAs, which can achieve orders of magnitude faster than fault injection method while is very accurate.

After the soft error rate estimation, the most vulnerable modules in the system are identified. The next step is performing SEU mitigation for the entire system. The previous SEU mitigation techniques impose 100%-200% overhead in terms of area and power [2,9–11]. This extra overhead, in turn, affects the performance of the system mapped into SRAM-based FPGA. The extra power overhead even limits the widespread use of these devices in reliable embedded applications.

In this paper, we present an effective approach for reducing soft error rate (SER) in SRAM-based FPGA. First, the entire system is divided into several modules according to their functions. Then the soft error rate of each module is calculated by an analytical estimation method. Finally, the entire system is mitigated based on partial duplication within the limits of cost, such as area and/or power, achieving a high reliable system with low cost. The proposed approach does not require physical implementation, i.e. only a synthesis tool and a software program are used.

The rest of the paper is organized as follows. Section 2 describe the whole approach proposed to reduce the soft error rate of systems based on FPGAs. Experimental results are presented in Section 3. Finally, Section 4 concludes the paper.

2   Proposed Algorithm

Reliability is a major issue in aerospace applications. Since SRAM-based FPGAs have soft error susceptibility due to SEU, it is necessary to perform mitigation for system mapped into SRAM-based FPGAs so as to achieve a high reliability. However, there are cost constraint in many applications, especially in aerospace applications where area and power are major issues. Redundant mitigation, in turn, would affect the performance of the system. As a result, it's impractical to perform mitigation for the entire SRAM-based FPGA system.

To balance high reliability and low cost for SRAM-based FPGAs system, we propose an approach to reduce the soft error rate of the system based on partial duplication under cost restriction, i.e. area. Considering the high time and space complexity introduced during the estimation of soft error rate of an entire system, we divide the entire system into several modules according their functions in the system first and compute their soft error rates separately. The calculation of soft error rate for each module is conducted using an analytical SER estimation method. Then, according to the observation that the primary outputs have a very high soft error rate since they are always sensitized, we traverse the system from the modules near the primary outputs to the modules near the primary inputs in a greedy manner to find the modules with highest

soft error rate and make them duplicated within cost constraint, that is partial duplication.

The flowchart of the proposed approach is illustrated in Figure 1. The details of the analytical SER estimation method and mitigation based on partial duplication are presented in the following subsections, respectively.
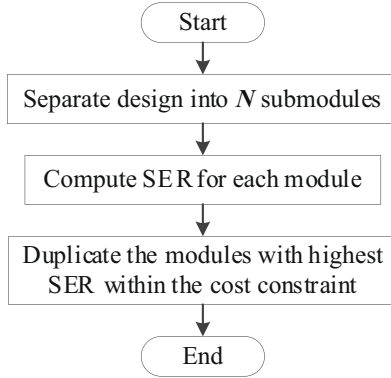

Fig. 1. Flowchart of the proposed algorithm

2.1   Analytical SER Estimation Method

The analytical SER estimation method is exploited to calculate the soft error rate for each module of the system mapped into SRAM-based FPGAs. In the analytical SER estimation method, the node error rate (PR_i) based on FPGA used resources (placement and routing information) for the given mapped design is calculated, followed by the calculation of the error propagation probabilities (N_i) using the gate-level netlist. The error propagation calculation rules for elementary gates are shown in Table 1. Based on the node error rate PR_i and the error propagation probabilities N_i, the soft error rate due to node i in the module, S_i, can be compute as follows:

$$S_i = PR_i \times N_i \tag{1}$$

Having the soft error rates computed for all nodes, we compute the soft error rate for the entire module in the clock after the particle hit according to Equation 2.

$$S = 1 - \prod_{i=1}^{n}(1 - S_i) = 1 - \prod_{i=1}^{n}(1 - PR_i \times N_i) \tag{2}$$

In addition, the soft error rate of the entire module in c clock cycles after the particle hit is calculated according to Equation 3.

$$S = 1 - \prod_{i=1}^{n}(1 - PR_i \times (1 - (1 - N_i)^c))) \tag{3}$$

Table 1. computing probability at the output of a gate in terms of its inputs

AND	$P_1(out) = \prod_{i=1}^{n} P_1(X_i)$
	$P_a(out) = \prod_{i=1}^{n}[P_1(X_i) + P_a(X_i)] - P_1(out)$
	$P_{\bar{a}}(out) = \prod_{i=1}^{n}[P_1(X_i) + P_{\bar{a}}(X_i)] - P_1(out)$
	$P_0(out) = 1 - [P_1(out) + P_a(out) + P_{\bar{a}}(out)]$
OR	$P_0(out) = \prod_{i=1}^{n} P_0(X_i)$
	$P_a(out) = \prod_{i=1}^{n}[P_0(X_i) + P_a(X_i)] - P_0(out)$
	$P_{\bar{a}}(out) = \prod_{i=1}^{n}[P_0(X_i) + P_{\bar{a}}(X_i)] - P_0(out)$
	$P_1(out) = 1 - [P_0(out) + P_a(out) + P_{\bar{a}}(out)]$
NOT	$P_0(out) = P_1(in)$
	$P_a(out) = P_{\bar{a}}(in)$
	$P_{\bar{a}}(out) = P_a(in)$
	$P_1(out) = P_0(in)$

The following are the steps of the analytical SER estimation method for one individual module.

1. Extract detailed placement and routing information of the specified module mapped into an SRAM-based FPGA using the Xilinx design language (XDL).
2. Obtain all signals from n_i to any reachable primary output PO_j and/or flip-flop ff_j. This is achieved using the forward and backward Depth-First Search (DFS) algorithm [12].
3. Levelize signals on these paths using the topological sorting algorithm [12]. Topological sort of a directed acyclic graph is an ordered list of the vertices such that if there is an edge (u, v) in the graph, the u appears before v in the list.
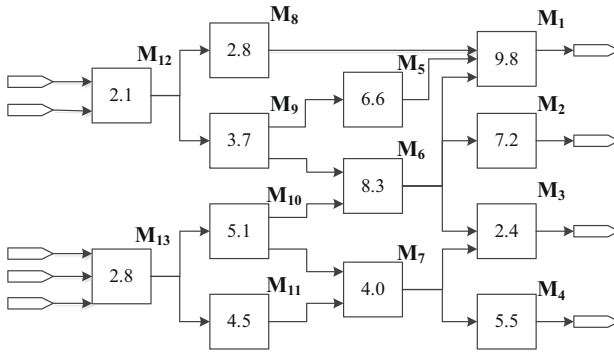
Fig. 2. Example of mitigation based on partial duplication

4. Traverse the paths in order. Compute signal probability (SP_i) and apply propagation probability rules, shown in Table 1, to compute propagation probability (PP_i) for node i. Error propagation probability N_i is the product of signal probability SP_i and propagation probability PP_i of node i.
5. Calculate the node error rate PR_i based on the raw error rate of the device, the error model, and the number of SRAM configuration bits used for implementing node i in the FPGA.
6. Calculate soft error rate due to node i according to Equation 1 and further calculate the soft error rate of the specified module according to Equation 2 and 3.

2.2 Soft Error Mitigation Based on Partial Duplication

Once the soft error rate of all the modules in the system are calculated, a mitigation based on partial duplication within specified cost overhead is conducted for the system. The basic idea of the mitigation based on partial duplication is to traverse the entire system from the modules near the primary outputs to the modules near the primary inputs in a greedy manner to find the set of the modules with highest soft error rate and make them mitigated. The mitigation processes modules from the modules near the primary outputs for two reasons. The first is based upon the observation that the modules near the primary outputs have a high soft error rate since they are always sensitized. The second is that the soft error rate of the entire system is measured on the primary outputs. The details of the mitigation based on partial duplication is presented below.

First, a priority queue (indexed by soft error rate) is initialized with the modules near the primary outputs most. At each iteration of the mitigation method, the module m with the highest soft error rate is removed from the priority queue and added to a set s. The current cost overhead is updated to reflect the latest addition to the set s and all modules that are inputs to the module m are inserted to the priority queue. This process terminates when the size of set s equals (or just exceeds) the specified cost overhead. Mitigation can

be varied immediately with the change of the specified cost overhead, due to the continual update of the cost overhead.

To better understand the procedure of the mitigation based on partial duplication, an example, shown in Figure 2, is presented. In Figure 2, the soft error rate of each module in the system, as well as the connection relationship between each other, is provided. The only constraint to mitigation is the area overhead that is allowed for partial duplication. First, modules M_1 to M_4 which near the primary outputs are inserted into a priority queue ModuleQ. Then, the module with highest soft error rate, in this case, module M_1 is removed from the priority queue ModuleQ and added to a set s. The current area overhead is updated to reflect the latest addition to the set s and all inputs of module M_1, i.e. modules M_5, M_6 and M_8, are added to the priority queue ModuleQ. Choose the module with highest soft error rate, i.e. module M_6, and perform the same operations to it as to module M_1. In this manner, the set s grows until the area overhead of all the modules due to mitigation added to the set s exceeds the specified area overhead constraint. Finally, in this example, modules M_1, M_6, M_2, M_5, M_4, M_{10} and M_{11} are duplicated to meet the high reliability (low soft error rate) under the limits of area overhead.

Table 2. Experimental results of our proposed algorithm

Soft Error Rate Reduction (%)					
Circuit			Area Overhead		
Name	No. PIs	No. POs	20%	33%	50%
s298	3	6	56.1	63.6	79.3
s344	9	11	67.8	71.9	85.2
s832	18	19	49.6	61.1	77.9
s1196	14	14	63.9	69.3	78.8
s1423	17	5	53.1	71.7	83.9
s3330	40	73	65.6	75.8	89.1
s6669	83	55	66.7	73.9	83.4
s9234	36	39	48.8	62.2	72.7
s15850	77	150	53.4	63.1	82.1
s38584	38	304	65.5	74.6	79.1
Average Reduction			59.1	68.7	81.2

3   Experimental Results

The proposed approach is implemented and applied to ISCAS89 sequential benchmark circuits. A framework for the soft error rate estimation method and mitigation based on partial duplication described before is implemented in C++.

Table 2 presents the reductions in the soft error rate that achieved by our proposed approach. The soft error rate reduction percentage is computed according to:

$$\left(\frac{\text{Original SER} - \text{Reduced SER}}{\text{Original SER}} \right) \times 100\% \tag{4}$$

The last row presents the average reduction in the soft error rate that is observed using the proposed approach. Note that an order of magnitude reduction in the soft error rate can generally be achieved with a 50∼60% area overhead.

4   Conclusion and Future Work

Soft errors due to single event upsets are the main reliability threat of digital systems. In particular, systems mapped into SRAM-based FPGAs are vulnerable to single event upsets and their applications in aerospace field are narrowed. In this paper, a fast but accurate soft error rate estimation method is applied to compute soft error rate for all the modules divided from the system, followed by a soft error mitigation method based on partial duplication to significantly reduce the total soft error rate of the system within the specified area overhead. Experiments on the benchmark circuits show the effectiveness of our proposed approach.

Potential future work in the area includes, e.g. exploring a better way to divide the system according to its function accurately, which can achieve more accurate mitigation for the system. In addition, advanced mitigation methods should be investigated to better balance the reliability and cost constraint.

References

1. Asadi, H., Tahoori, M.B., Mullins, B., Kaeli, D., Granlund, K.: Soft Error Susceptibility Analysis of SRAM-Based FPGAs in High-Performance Information Systems. IEEE Transactions on Nuclear Science 54(6) (2007)
2. Carmichael, C., Fuller, E., Fabula, J., Lima, F.: Proton Testing of SEU Mitigation Methods for the Virtex FPGA. In: Proceedings of the Military and Aerospace Applications of Programmable Logic Devices (MAPLD), Washington D.C. (2001)
3. Rebaudengo, M., Sonza Reorda, M., Violante, M.: Simulation-Based Analysis of SEU Effects on SRAM-Based FPGAs. In: Proceeding of the 12th International Conference on Field-Programmable Logic and Application (FPL2002), Montpellier, France (2002)
4. Asadi, G., Miremadi, G., Zarandi, H.R., Ejlali, A.: Fault Injection into SRAM-Based FPGAsfor the Anlysis of SEU Effects. In: Proceeings of the IEEE International Conference on Field-Programmable Technology (FPT), Tokyo, Japan, pp. 428–430 (2003)
5. Gokhale, M., Graham, P., Johnson, E., Rollins, N., Wirthlin, M.: Dynamic Reconfiguration for Management of Radiation-Induced Faults in FPGAs. In: Proceedings of the 18th International Parallel and Distributed Processing Symposium (IPDPS 2004), Santa Fe, New Mexico, pp. 145–150 (2004)

6. Graham, P., Caffrey, M., Zimmerman, J., Johnson, D.E., Sundararajan, P., Patterson, C.: Consequences and Categories of SRAM FPGA Configuration SEUs. In: Proceedings of the Military and Aerospace Applications of Programmable Logic Devices (MAPLD), Washington D.C. (2003)
7. Asadi, G., Tahoori, M.B.: An Analytical Approach for Soft Error Rate Estimation of SRAM-Based FPGAs. In: Proceedings of the Military and Aerospace Applications of Programmable Logic Devices (MAPLD), Washington D.C. (2004)
8. Asadi, G., Tahoori, M.B.: Soft Error Rate Estimation and Mitigation for SRAM-Based FPGAs. In: Proceedings of the ACM/SIGDA 13th International Symposium on Field Programmable Gate Arrays, FPGA 2005, Monterey, California, USA (2005)
9. Lima, F., Carmichael, C., Fabula, J., Padovani, R., Reis, R.: A Fault Injection Analysis of Virtex FPGA TMR Design Methodology. In: Proceedings of the Radiation Effects on Components and Systems Conference (RADECS 2001), Grenoble, France (2001)
10. Lima, F., Carro, L., Reis, R.: Designing Fault Tolerant Systems into SRAMM-Based FPGAs. In: Proceedings of the IEEE/ACM Design Automation Conference (DAC), pp. 650–655 (2003)
11. Carmichael, C., Fuller, E., Blain, P., Caffrey, M.: SEU Mitigation Techniques for Virtex FPGAs in Space Applications. In: Proceedings of the Military and Aerospace Applications of Programmable Logic Devices (MAPLD), Washington D.C. (1999)
12. Cormen, T.H., Leiserson, C.L., Rivest, R.L., Stein, C.: Introduction to algorithms, 2nd edn., pp. 549–551. MIT Press & McGraw-Hill (2001)

Research on the System Reliability Modeling Based on Markov Process and Reliability Block Diagram

Guochang Zhou[1], Baolong Guo[2], Xiang Gao[1], Dan Zhao[2], and Yunyi Yan[2]

[1] China Academy of Space Technology(Xi'an)
Xi'an 710000, China
[2] School of Aerospace Science and Technology
Xidian University
Xi'an 710071, China
zhaodan115@gmail.com

Abstract. The on board computer systems have a high requirement on the reliability because of single event upset. So this paper analyzes the system reliability model, and presents a method to model the reliability of on board computer system. This paper defines the basic module of the system reliability using Markov model and proposes a system reliability prediction algorithm using reliability block diagram. By simplifying the reliability block diagram using the graph theory, to predict the reliability of a system.

1 Introduction

In the space radiation environment, the performance of electronic devices is often limited by its susceptibility to single event effects (SEE)[1], particularly radiation-induced single event upset (SEU). Thus they are protected through a very expensive process called radiation-hardening. But because of the cost, in many cases, soft mitigations are used to protected the computer systems from SEU.

Various methodologies have been developed over years to mitigate the system-level impact of device errors. For example, triple modular redundancy (TMR), configuration memory scrubbing, error detection and correction (EDAC) and so on. In fact, the effect of every methodology is different, so the problem is how to estimate the effect of a mitigation method.

Edmonds[2] analysed the single event upset rates in devices with the TMR mitigation. The probability of a system error during a single cycle was derived, and used this result to derive an equation that calculated the system error rate. And it is common to find microprocessor with the memory protected by Hamming code, one example is the Atmel SPARC microprocessor that is protected by EDAC[3]. However, the current forms of EDAC have various limitions, and bit errors can become observable under certain conditions. Different types of EDAC differ in terms of what those conditions are. The purpose of using these methods is to improve the system reliability.

Several methods exist to model system reliability. Markov modeling, named for the Russian mathematician Andrei Markov, is one such method[4]. The underlying assumption of Markov models is that the probability of a state transition depends only on the current state. The reliability of a system can be modeled as a very simple Markov process with two states: functional and failed. At the system level, reliability block diagram (RBD)[5] is a very popular and useful tool for evaluating the system reliability. RBD has been an active area of research for decades, even more so now with the advent of the embedded systems.

In this paper, a estimation method is presented for on board computer system reliability. The presented method combines Markov models with RBD. A system may contain one or more modules which are basic units of the system. Different modules may have their own reliability models with respective parameters. By using markov chain, we build a model for every component. After obtaining the reliability of every component, the system reliability can be evaluated according to the architecture and relationship of the components, and the architecture and relationship were illustrated by the RBD.

2   System Model

2.1   Reliability Definition

Reliability engineering is the study of product failure occur, the law of development, during product life-cycle process, and the purpose is to prevent failure, to eliminate failure and to improve efficiency. A reliability model for modeling a product presents a bathtub curve and can be divided into three regions, infant mortality, steady-state operation, and wearout[6]. Each region can be modeled with a different reliability function. In this paper, we focus on steady-state operation during use.

Reliability is an important quality index. Reliability is the probability that a system will perform its intended function under specified working condition for a specified period of time. Mathematically, the reliability function $R(t)$ is the probability that a system will be successfully operating without failure in the interval from time 0 to time t:

$$R(t) = P(T > t), t \geq 0 \tag{1}$$

where T is a random variable representing the failure time or time-to-failure.

2.2   Markov Model

The on board computer systems are often bombarded with the high energy particles in the space radiation environment, such as proton and heavy ion. So in the electronic devices may occur SEUs, and the output of this devices may be unreliable. A on board computer system may contain multiple modules. The reliability of a module can be modeled as a very simple Markov process. We
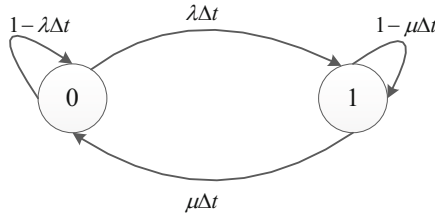
Fig. 1. A two-state Markov model

suppose that there are only two states: functional and failed, in the state space, and they can transform each other.

A graphical representation of this Markov model is shown in Figure 1. State 0 represents the functional state and state 1 represents the failed state. The state of a module transitions from functional to failed at a rate of λ, as labeled on the arc from state 0 to state 1. Also, it is assumed that a module can be repaired at a rate of μ, as labeled on the arc from state 1 to state 0.

The reliability function of the Markov model depicted in Figure 1 can be derived by first converting the process to a continuous-time model and then solving a set of differential equations for the probability the system is in state 0, the functional state, at time t. To convert to a continuous-time Markov model, a transition matrix $\mathbf{T}$ is needed. The transition matrix for Figure 1 is

$$\mathbf{T} = \begin{bmatrix} 1 - \lambda\Delta t & \lambda\Delta t \\ \mu\Delta t & 1 - \mu\Delta t \end{bmatrix} \tag{2}$$

where the entry in row m, column n represents the probability of transitioning from state m to state n.

From the transition matrix a set of equations can be defined that represent the probability of being in state 0 or state 1 at time $t + \Delta t$. The distribution of the probabilities at time $t + \Delta t$ is equal to the product of the distribution of the probabilities at time t multiplied by the transition matrix. State mathematically,

$$[p_0(t + \Delta t), p_1(t + \Delta t)] = [p_0(t), p_1(t)]\, \mathbf{T} \tag{3}$$

Multiplying out yields the system of equations

$$p_0(t + \Delta t) = (1 - \lambda\Delta t)p_0(t) + \mu\Delta t p_1(t) \tag{4}$$

$$p_1(t + \Delta t) = \lambda\Delta t p_0(t) + (1 - \mu\Delta t)p_1(t) \tag{5}$$

Subtracting $p_0(t)$ from both sides of Equation 4, $p_1(t)$ from both sides of Equation 5. Then dividing both sides of both equations by Δt gives

$$\frac{p_0(t + \Delta t) - p_0(t)}{\Delta t} = -\lambda p_0(t) + \mu p_1(t) \tag{6}$$

$$\frac{p_1(t + \Delta t) - p_1(t)}{\Delta t} = \lambda p_0(t) - \mu p_1(t) \tag{7}$$

And then taking the limit of Equations 6 and 7 generates

$$p_0^{'}(t) = -\lambda p_0(t) + \mu p_1(t) \tag{8}$$

$$p_1^{'}(t) = \lambda p_0(t) - \mu p_1(t) \tag{9}$$

In order to solve this set of differential equations more easily, we convert them to the LaPlace domain. Taking the LaPlace transform of Equations 8 and 9 yields

$$s P_0(s) - P_0(0) = -\lambda P_0(s) + \mu P_1(s) \tag{10}$$

$$s P_1(s) - P_1(0) = \lambda P_0(s) - \mu P_1(s) \tag{11}$$

Solving this set of Equations 10 and 11 generates

$$P_0(s) = \frac{s + \mu}{s^2 + (\lambda + \mu)s} P_0(0) + \frac{\mu}{s^2 + (\lambda + \mu)s} P_1(0) \tag{12}$$

$$P_1(s) = \frac{\lambda}{s^2 + (\lambda + \mu)s} P_0(0) + \frac{s + \lambda}{s^2 + (\lambda + \mu)s} P_1(0) \tag{13}$$

Supposing that at the beginning the module is in the functional state, that is $P_0(0) = 1, P_1(0) = 0$. Substituting the two values into Equations 12 and 13 gives

$$P_0(s) = \frac{s + \mu}{s^2 + (\lambda + \mu)s} \tag{14}$$

$$P_1(s) = \frac{\lambda}{s^2 + (\lambda + \mu)s} \tag{15}$$

Taking the inverse LaPlace transform of Equations 14 and 15 results in

$$p_0(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t} \tag{16}$$

$$p_1(t) = \frac{\lambda}{\lambda + \mu} - \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t} \tag{17}$$

$p_0(t)$ is the probability of the module at the functional state and $p_1(t)$ is the probability of the module at the failed state. So the reliability of the module is $p_0(t)$, that is

$$R(t) = \frac{\mu}{\lambda + \mu} + \frac{\lambda}{\lambda + \mu} e^{-(\lambda + \mu)t} \tag{18}$$

2.3   Reliability Block Diagram

Reliability block diagram (RBD) is one of the most common and conventional tools of system reliability analysis. A major advantage of using the reliability block diagram approach is the ease of reliability expression and evaluation. A reliability block diagram shows the system reliability structure. It is made up of individual blocks and each blocks and each block corresponds to a system module or function. A RBD graphically represents a series-parallel system in which its components are combined into blocks in series, in parallel or in k-out-of-n configuration[7] as shown in Figure 2. They are all standard modules. Standard series module has only two units. One unit input node is directly linked to the module input node, and the output node is linked to the other unit input node. The standard series is shown in Figure 2(a).

(a) Series system   (b) Parallel system   (c) k-out-of-n system
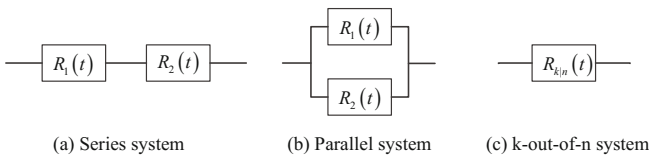
Fig. 2. Reliability block diagram

The mathematical formula to calculate the reliability of standard series module:

$$R_s(t) = R_1(t) \times R_2(t) \tag{19}$$

All the input nodes if the units in the standard parallel modules are directly linked to the module input node, and the output nodes are linked to the module output node. This is the standard parallel module, is shown in Figure 2(b). The mathematical formula to calculate the reliability of standard parallel module:

$$R_s(t) = 1 - \prod_{i=1}^{n}(1 - R_i(t)) \tag{20}$$

A k-out-of-n system requires that at least k modules out of a total of n must be operational in order for the system to be working. Usually a voter is needed. The reliability of a k-out-of-n system can be denoted by a symbol $R_{k|n}(t)$. If the voter is perfect and all the modules have reliability R, the formula to evaluate the reliability of these blocks, which can be obtained via following equation:

$$R_{k|n}(t) = \sum_{i=k}^{n} C_n^i R^i (1 - R)^{n-i} \tag{21}$$

All said above were standard modules. But in a real system, the topology relation of the modules may be very complex. So analyzing the topology relation is necessary.

3   System Reliability Prediction

An algorithm was proposed[8] to predict the system reliability. Basic algorithm described in Figure 3. The key parts of the prediction algorithm were searching for standard modules and simplifying the reliability block diagram.
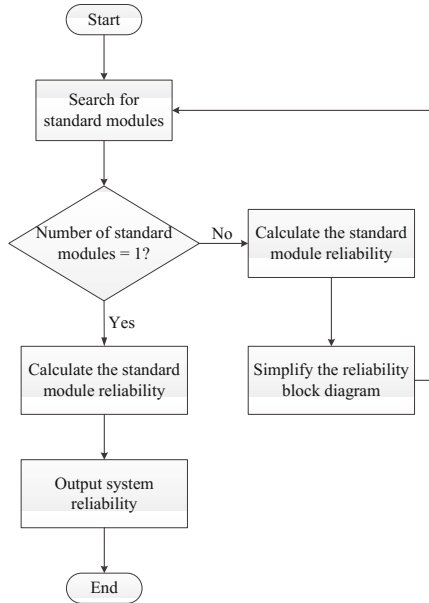

Fig. 3. The prediction algorithm

In a real on board computer system, the topology relation of the modules is usually very complex. Assuming that it is a mesh structure. In order to predict the system reliability, decomposing and simplifying the reliability block diagram is necessary.

Step 1: Search for the standard module. According to the definition of standard modules and its design features, from input to start the search, traverse the reliability block diagram to obtain the modules and their connected relation. And then storing the relation in a graph, one of data structures.

Step 2: Apply the mathematical formula of standard module to calculate the standard module reliability.

Step 3: Determine the number of standard modules, if the number of standard modules is equal to one, go to the sixth step, otherwise go to fourth step.

Step 4: Simplify the reliability block diagram. Apply independent unit instead of the standard modules.

Step 5: Go to the first step, and continue to search and simplify system reliability block diagram.

Step 6: Output system reliability. Apply the mathematical formula of standard module to calculate the standard module reliability, the reliability of the standard module is the system reliability, the output the reliability value.

In this algorithm, the most important steps are searching for the standard module and simplifying the reliability block diagram. For a complex system, the structure of its reliability block diagram may be a mesh network. It can be expressed by a directed graph $G = (V, E)$. So the graph theory can be used searching for the standard module. Every module is a vertex in the directed graph. It is difficult to analyze the reliability of a mesh network because a regular structure is not provided within such networks. The core of the problem is how to compute network reliability using a series-parallel graph.

The series-parallel graph was defined more formally in[9]. By defining a finite linear directed graph to be an ordered quadruple $G = (N, A, S, T)$ where

1. N is a finite set of elements called nodes.
2. A is a subset of $N \times N$, called the set of edges.
3. S is the subset of N containing those nodes that have no incoming edges. These are the entrance nodes.
4. T is the subset of N containing those that have no outgoing edges. These are the exit nodes.
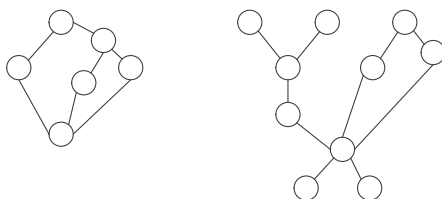

Fig. 4. Examples of series-parallel graphs

Note that a series-parallel graph is by definition acyclic. The definition of series-parallel also implies that a series-parallel graph cannot have redundant edges. That is, if there is an edge from vertex x to vertex y, there is no other path from x to y.

Figure 4 shows several examples of series-parallel graphs, the direction of the edges is not shown explicitly.

First we decompose the graph, obtaining a tree representation of the graph that shows the series and parallel combinations that formed the graph. The leaves of the tree correspond to nodes of the graph, and each internal tree node represents either a series or parallel combination of its subtree. When the decomposition is parallel, we label the internal node with the particular interpretation(maximum, minimum, probabilistic, or k-out-of-n) assigned to the set of parallel subgraphs. If a node is internal, it represents the series or parallel combination of two or more subgraphs.

Second step is to calculate the graph's reliability. We do that by visiting all of the nodes of the tree in postorder.

4   Example

If a system is composed of components, each having an independent repair facility, the proposed method can be used to compute the instantaneous reliability of the system. Consider the series-parallel system of components in Figure 5. This is the example in[10]. Use our method, we can compute the reliability. For the subsystems in parallel, we must take the product of the component unreliability(the system is unreliable only when all parallel subsystems are unreliable). This is the "maximum" combination. For a series of components, the reliability is the product of the component reliability(the system is reliable only when all subsystems are reliable). This is the "minimum" combination of the components.
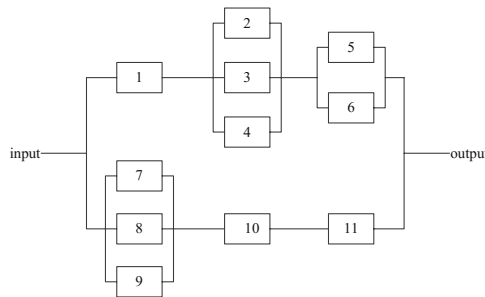
Fig. 5. Reliability block diagram

We use the same parameters as in[10]. Those parameters are $\lambda_5 = \lambda_6 = 0.005, \lambda_1 = 0.001, \lambda_i = 0.01$ for all other i, $\mu_5 = \mu_6 = 1/6, \mu_1 = \mu_{10} = \mu_{11} = 1/5$ and $\mu_i = 1/7.5$ for all other i. In Equation 18, as t approaches infinity, $R(t)$ approaches the steady-state reliability. According to the data above and the system structure, calculating the reliability of the system is 0.9994.

5   Conclusion

From the above, we proposed a method to estimate the reliability of the on board computer system. First of all, assume that each module is subject to failure, and has own independent repair facility, then modeled the module using markov chain. In the system level, reliability block diagram was used to estimate the system reliability.

In the algorithm, the most important steps were searching for the standard module and simplifying the reliability block diagram, therefore, a series-parallel graph was used to deal with the complex system structure. Decomposing the series-parallel graph, obtaining a tree representation of the graph that shows

the series and parallel combinations that formed the graph. By visiting all of the nodes of the tree in postorder to calculate the graph's reliability, that is the system reliability.

References

1. Petersen, E.: Single event effects in aerospace. Wiley-IEEE Press (2011)
2. Edmonds, L.D.: Analysis of Single-Event Upset Rates in Triple-Modular Redundancy Devices. NASA Jet Propulsion Laboratory (2009)
3. Velazco, R., Bessot, D., Duzellier, S., Ecoffet, R., Koga, R.: Two CMOS memory cells suitable for the design of SEU-tolerant VLSI circuits. IEEE Transactions on Nuclear Science 41, 2229–2234 (1994)
4. McMurtrey, D., Morgan, K., Pratt, B., Wirthlin, M.: Estimating TMR Reliability on FPGAs Using Markov Models. Brigham Young University Department of Electrical and Computer Engineering, Tech. Rep. (2007)
5. Sahinoglu, M., Chittoor: RBD Tools Using Compression, Decompression Hybrid Techniques to Code, Decode, and Computer Reliability in Simple and Complex Embedded systems. IEEE Transactions on Instrumentation and Measurement 54(5), 1789–1799
6. Crowe, D., Feinberg, A.: Design for Reliability. CRC Press LCC (2001)
7. Lin, C.M., Teng, H.K., Yang, C.C., et al.: A mesh network reliability analysis using reliability block diagram. In: 2010 8th IEEE International Conference on Industrial Informatics (INDIN), pp. 975–979. IEEE (2010)
8. Yanjun, L., Wei, W.: Research on the system of reliability block diagram design and reliability prediction. In: 2011 International Conference on Engineering Design and Manufacturing Informatization (ICSEM), vol. 2, pp. 35–38. IEEE (2011)
9. Sahner, R.A., Trivedi, K.S.: Performance and reliability analysis using directed acyclic graphs. IEEE Transactions on Software Engineering, 1105–1114 (1987)
10. Modarres, M.: A method of predicting availability characteristics of series-parallel systems. IEEE Trans., Rel. R-33(4), 308–312 (1984)

Precision Mosaicking of Multi-images Based on Conic-Epipolar Constraint

Meng Yi[1], Yan Chu[1], and Yunyi Yan[2]

[1] School of Electronic and Control Engineering, Chang'an University
Xi'an 710046, China
[2] School of Aerospace Science and Technology, Xidian University
Xi'an 710071, China
`yimeng0120@gmail.com`

Abstract. In this paper a robust mosaic method based on conic-epipolar constraint is proposed. The main characteristics of the proposed method include: (1) Several new methods are presented to realize fast and accurate interest points extraction under various different scenes, including SURF based feature points detection, interest points selection based on uniform distribution. (2) the transformation parameters are estimated using the invariant of conic-epipolar constraint and the most "useful" matching points are used to register the images. Experiment results illustrate that the proposed algorithm carries out real-time image registration and is robust to large image translation, scaling and rotation.

Keywords: Image mosaic, Image registration, Multi-images, Conic-epipolar constraint.

1 Introduction

Image mosaicking is the process of smoothly piecing together overlapping images of a scene into a larger image. An image mosaic is created from a set of overlapping images by registering and resampling all images to the coordinate space of one of the images. Nowadays, image mosaic has been widely applied in many computer vision domains, including robot localization, remote sensing, medical image analysis, pattern recognition, enhanced visualization and ground moving target detection and tracking [1-3]. Image mosaicking can be mainly divided into feature-based methods and area-based methods. As showing good invariant under a certain mathematical transformations and grayscale distortion, Image mosaic algorithm based on the feature-based methods has been draw widely attention in recent years. A feature is a local point or area which is invariant to illumination changes, noise disturbance and geometric deformations. As an important local feature of the image, the corners are found at various types of highly textured surfaces of the image, while only use a small amount of local information. Therefore, massive literature use corner as the invariant feature. Successful mosaicking of overlapping images using feature points has been reported, Szeliski [4] reviewed and classified existing mosaicking methods.

There are essentially three parts to any feature-based image mosaic algorithm: invariant feature detection and invariant Feature matching. Invariant feature matching is the most important step in feature-based image mosaic. The most popular methods are landmark-based descriptors and distribution-based descriptors (SIFT [5], GLOH [6] and PCA-SIFT [7]), which extract distinctive feature that are invariant to noise and viewpoint. However, due to the noise and occlusion, some matching points will be more accurate than others [8]. In this study, conic-epipolar is adopted [9]. Compared with other methods, this algorithm can derive an explicit relation between the local affine approximations and the epipolar geometry, then distinguish more accurate matching points from the less accurate ones and then register the images as accurately as possible. The result is further employed for estimating the transformation matrix, based on the estimated location of the epipole. It is shown to be more accurate and to require less iterations in RANSAC based estimation, than the feature point based approaches that employ the affine relation to generate correspondence points and then calculate the transformation model from the points relations.

The flowchart of the proposed algorithm framework is shown in Fig.1. This paper is structured as follows. In section 2, the proposed registration algorithm is describes in detail. Section 3 presented and evaluated the experimental results of the performance of the algorithm. Section 4 concludes.
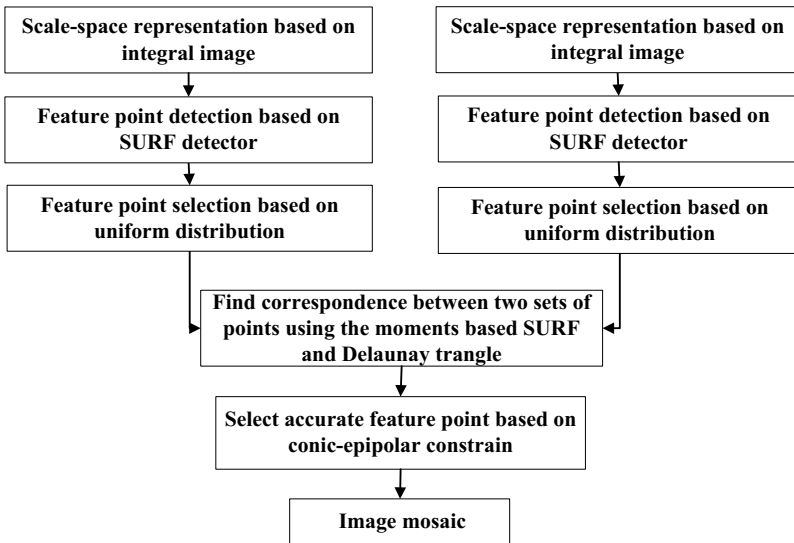

Fig. 1. Flowchart of the proposed image mosaic algorithm

2   Aproach

Assuming (x, y) represents a point in the base image and (X, Y) represents the same point in the image overlapping the base image,   the projective transformation between corresponding points in the images can be written as:

$$\begin{bmatrix} X \\ Y \end{bmatrix} = s \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} \Delta x \\ \Delta y \end{bmatrix} \tag{1}$$

If the number of correspondences n is assumed to be greater than the number of model parameters to be estimated i.e $n > 4$, and let it be a set of n point correspondences between the source and target image. We can solve this as a set of $2n$ equations, Therefore, the problem of the above model is transformed into function $B = Am$, as shown in Eq(2).

$$B = \begin{bmatrix} x_{t,1} - x_{t-1,1} \\ y_{t,1} - y_{t-1,1} \\ \vdots \\ x_{t,N} - x_{t-1,N} \\ y_{t,N} - y_{t-1,N} \end{bmatrix} \quad A = \begin{bmatrix} -y_{t-1,1} & 1 & 0 \\ x_{t-1,1} & 0 & 1 \\ \vdots & & \\ -y_{t-1,N} & 1 & 0 \\ x_{t-1,N} & 0 & 1 \end{bmatrix} \quad m = \begin{bmatrix} \theta, \Delta X, \Delta Y \end{bmatrix}^T \tag{2}$$

We use the over-determined linear equations to solve for m, the initial value of m is calculated by $m = (A^T A)^{-1} A^T B$.

2.1   Feature Points Extraction

SURF feature points are robust to illumination change, scale zooming, affine transformation and noise [10], Based on the excellent performance, SURF detector extract feature points between the frames and match the feature points, and then estimate the background global motion vector. In this paper, the intensity, quantity and distribution were considered in the process of feature point extraction section. the main direction of SURF descriptor extraction method was accelerated, and the use of grids eliminate the weak feature points, and improve differentiation and correct matching rate of the feature points.

2.1.1   SURF Detector
The detector is based on the determinant of the Hessian matrix.

$$H(\text{x},\sigma) = \begin{bmatrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma) \end{bmatrix} \tag{3}$$

Here $L_{XX}(x,\sigma)$ refers to the convolution of the second order Gaussian derivative $\partial^2 g(\sigma)/\partial x^2$ with the image at point x=(x, y) and similarly for L_{xx} and L_{xy}.

The second order scale normalised Gaussian is the chosen filter as it allows for analysis over scales as well as space. We can construct kernels for the Gaussian

derivatives in x, y and combined x and y such that we calculate the four entries of the Hessian matrix. Use of the Gaussian allows us to calculated at different scales. convolution with the kernel allows for rotation invariance.

Lowe found performance increase in approximating the Laplacian of Gaussians by a diference of Gaussians. In a similar manner, Bay[24]proposed an approximation to Laplacian of Gaussians by using box filter representations of the respective kernels, as shown in Fig.2.
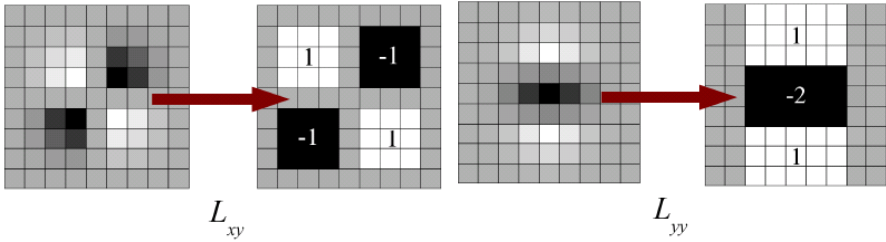

$$L_{xy} \qquad\qquad L_{yy}$$

Fig. 2. Approximation to the laplacian of Gaussians

The following formula as an accurate approximation for the Hessian determinant using the approximated   Gaussians:

$$\det\left(H_{approx}\right) = D_{xx}D_{yy} - \left(0.9D_{xy}\right)^2 \tag{4}$$

The value of the discriminant is used to classify the maxima and minima of the function by the second order derivative test. If the determinant is negative then the eigenvalues have different signs and hence the point is not a local extremum, if it is positive then either both eigenvalues are positive or both are negative and in either case the point is classified as an extremum. The determinant here is referred to as the response at location x=(x, y, σ). The search for local maxima of this function over both space and scale yields the interest points for an image.

2.1.2   Point Detection

In this paper, the main direction of feature points through calculating the moment, and moment are defined as follows:

$$M_{ij} = \sum_x \sum_y x^i y^j I(x, y) \tag{5}$$

Here, the direction of the blob regions determined by the following:

$$c_x = \frac{M_{10}}{M_{00}}, c_y = \frac{M_{01}}{M_{00}} \tag{6}$$

$$C_{ori} = \tan^{-1}(\frac{c_y}{c_x}) \tag{7}$$

In order to further improve the matching speed, in this paper, by means of the grid distribution, we further optimize the distribution of feature points. First of all, we set the uniform distribution points on the image grids, and each detected Hessian feature point is been marked as j from the nearest grid point sequence, here, $j \in (1, M \times N)$. With regard to the same marked feature point, we select the maximum of Hessian value as feature points. This feature point selection method will get $M \times N$ feature points. In this paper, M, N Satisfy the following principle: M is the image's height divided by 20, N is the image's width divided by 20.

2.2   Correspondence between Points

In order to obtain more accurate matching pairs, we choose Delaunay triangulation edges to find the matching pairs between feature points in overlapping images. For Aerial Video Image with general perspective changes (the view-points for the two images are not significantly different), the orientations of the corresponding edges in local areas should be relatively consistent. Consider, a set of J Delaunay triangle edge in the source image $F = \{F_1, F_2,...,F_j...,F_J\}$, we first hypothesize a line by randomly selecting a Delaunay triangle edge with two feature points, from which we label the line as either "standing" or "lying". We compute this label as

$$L_j = \left|x_{j1} - x_{j2}\right| / D_{j1,j2} \tag{8}$$

Where x_{j1} and x_{j2} is the x_j coordinates of start point and end point of a line, respectively. $D_{j1,j2}$ refer to the distance between two feature points. If the value of L_j is larger than 0.71, the line is labeled as "lying". Otherwise, the line is regarded as a standing line. After labeling lines, we build a model $v_j = (L_j, S_j, R_j)$ at each line F_j. If the line is lying, S_j refer to the angle of the line to x-axis.The definition for S_j are just inverse if the line is a standing line. The parameter R_j is to compare the strength contrast of corner response R (describe in formula (9)) in a local buffering region along both sides of the line. Assuming the equation expression of a line l is $Ax + By + C = 0$, for a local region centered at the line, the average strength on one side of the line is l_1 ,and the average strength on the other side of the line is l_2. A strength contrast R_j for each line is assigned as $l_2 - l_1$.

The model at the Delaunay triaunay edge is given by:

$$D = (\frac{\left\|v_j - \tilde{v}_j\right\|}{2\omega^2}) \tag{9}$$

We carry out matching of edge pairs from the two images. then the candidate edge is not considered as a possible matching edge and is excluded from further matching process

2.3   Conic-Epipolar Constraint

In computer vision, conics are widely accepted as one of the most fundamental image features together with points and lines [11]. Like points and lines, conics are invariant under planar projective transformations. This means that a perspective image of a plane conic is still a conic. The motivation for studying the geometry of conics arises from the fact that sometimes point or line correspondences are not available or could be noisy while higher order curves, such as conics, can still be identified robustly .

Before we start our analysis of conic correspondences, we will recapitulate some of the basic concepts in projective geometry, especially concerning conics and quadrics [12].

Where P denotes the standard 3×4 camera matrix and λ a scale factor. Here $X = [XYZ]^T$ and $X = [\mathrm{xy1}]^T$ denote homogeneous point coordinates in the 3D scene and in the image respectively. Given two cameras, defined by the camera matrices P_1 and P_2, and the images of a point X in space, i.e.

$$\lambda_1 x_1 = P_1 X \text{ and } \lambda_2 x_2 = P_2 X$$

the two corresponding points, x_1 and x_2 fulfil the epipolar constraint,

$$x_1^T F x_2 = 0 ,$$

where F denotes the fundamental matrix. The fundamental matrix has seven degrees of freedom. Since each point correspondence enforces one constraint, we need at least seven points to determine the entries of F, which can only be done up to a three-fold ambiguity.

The image, under a perspective projection(2), of a quadric is a conic. This relation is conveniently expressed as follow:

$$\lambda l = PLP^T , \quad \lambda \neq 0 ,$$

where λ is a scale factor.

Of special interest in projective geometry is the absolute conic. It is defined by the equations $X_1^2 + X_2^2 + X_3^2 = X_4 = 0$. It is a virtual conic, i.e. it contains no real points, positioned in the ideal plane, $X_4 = 0$. Euclidean transformations in the projective space are exactly those transformations that leave the position of the absolute conic invariant. Knowing the calibration of a camera is equivalent to knowing the image of the absolute conic.

A conic, and two points not lying on the conic define a single invariant given by[21]:

$$I = \frac{(x_1^T C x_2)^2}{(x_1^T C x_1)(x_2^T C x_2)}$$

where C is the matrix of the conic: a conic takes the form $ax^2 + bxy + cy^2 + dx + ey + f = 0$, which can be expressed as the quadratic form:

$$[x, y, 1] \begin{bmatrix} a & \dfrac{b}{2} & \dfrac{d}{2} \\ \dfrac{b}{2} & c & \dfrac{e}{2} \\ \dfrac{d}{2} & \dfrac{e}{2} & f \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = 0 \tag{10}$$

or as $x^T C x = 0$. The envelope of the conic and two lines is the dual system. The envelope of a conic C is the conic C^{-1}, and so two lines and a conic yield an invariant as well.

Then we can select the best corresponding feature points using the invariant of two points and 1 conic by comparing the distance of the feature points in one image and the coordinates points in another image. If the combination gives the small distance, then the best matching points will be selected to determine the parameters of the projective transformation.

At last, we use PROSAC algorithms for getting the optimum parameters of the model. PROSAC algorithms are the improved RANSAC algorithm, which is quicker than the RANSAC algorithm. The algorithm gets basic subset from the matching results of sampling point and the original data, then estimate the fundamental matrix using the obtained basic subset, that is, the affine matrix H from one image to another image.

3   Experimental Results

This section presents the example of the proposed image mosaic algorithm. The algorithm has been implemented in C++ and all experiments have been carried out on DELL Intel Xeon E5410 2.33-GHz desktop computer with 9GB of RAM, with Windows 7 Enterprise Professional Edition.

A mosaicking example is shown in Fig. 3. First, images (a), (c) and (d) are registered to image (b) to create one side of the mosaic image. Then, images (e), (g) and (h) are registered to image (f) to create another side of the mosaic image. Finally, two mosaic images create the overall mosaic (i). The largest rectangular area within the obtained mosaic as shown by (i) is used as the final result.
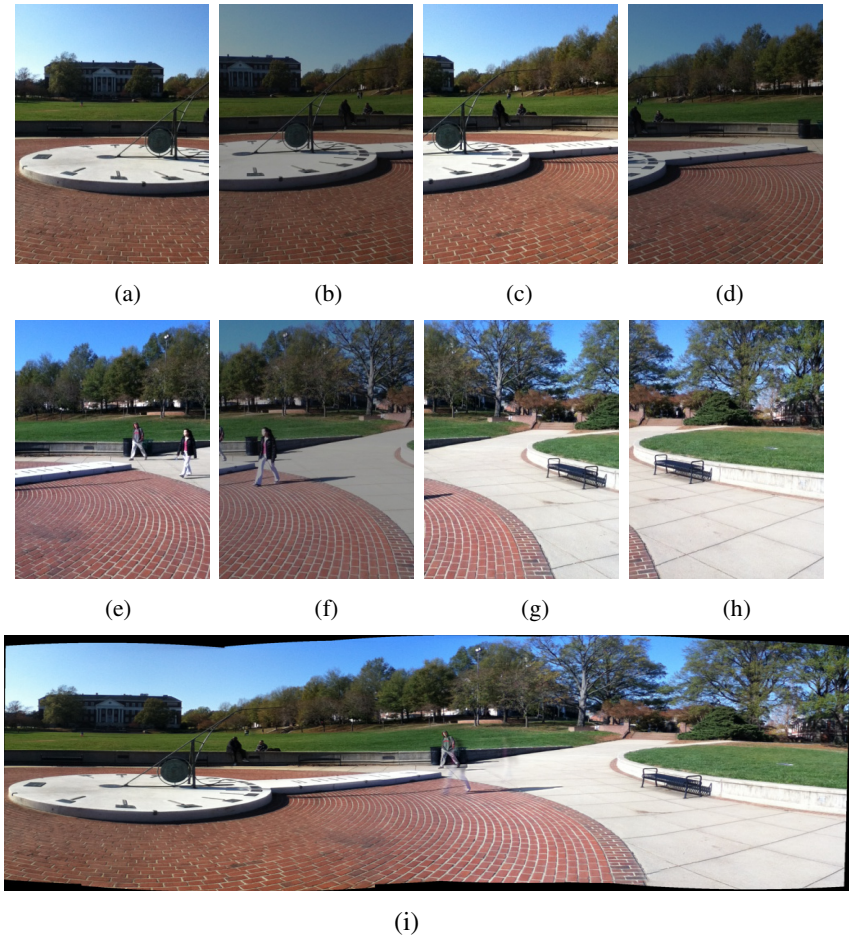
(a)   (b)   (c)   (d)

(e)   (f)   (g)   (h)

(i)

Fig. 3. (a)-(h) Image sequence. (i) Final mosaicking result.

4   Conclusions

In this paper, an automatic image mosaic technique based on conic-epipolar constraint was proposed. Keypoints are first extracted by SURF detector, then the correspondence between the feature points of the input image and reference image are found using SURF and delaunay triangulation, finally the input image with the correct mapping model is transformed for image fusion and complete image mosaicking. Experimental results demonstrate the proposed algorithm can achieve accuracy and create precision mosaic.

References

1. Zitova, B., Flusser, J.: Image registration methods: A survey. Image Vis. Comput. 21(11), 977–1000 (2003)
2. Ali, S., Reilly, V., Shah, M.: Sneddon: Motion and appearance contexts for tracking and reacquiring targets in aerial videos. In: IEEE CVPR, pp. 1–6 (2007)
3. Parag, T., Elgammal, A., Mittal, A.: A framework for feature selection for background subtraction. In: Proc. Computer Vision and Pattern Recognition, pp. 1916–1923 (2006)
4. Shum, H.-Y., Szeliski, R.: Construction of panoramic image mosaics with global and local alignment. Int. J. Comput. Vis. 36(2), 101–130 (2000)
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. 15(6), 415–434 (1997)
6. Mikolajczyk, K., Schmid, C.: A Performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. 27(10), 1615–1630 (2005)
7. Ke, Y., Sukthankar, R.: PCA-SIFT: A more distinctive representation for local image descriptors. In: IEEE Int. Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 506–513 (2004)
8. Krish, K., Heinrich, S., Snyder, W.E.: Global registration of overlapping images using accumulative image features. Pattern Recognition Letters 31(2), 112–118 (2010)
9. Rothwell, C.A., Zisserman, A.: Using projective invariants for constant time library indexing in model based vision. In: BMVC 1991 (1991)
10. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006, Part I. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
11. Mundy, J.L., Heller, A.: Geometric Invariance in Computer Vision. MIT Press, Cambridge (1992)
12. Kahl, F., Heyden, A.: using conic correspondences in two images to estimate the epipolar geometry. In: Proceedings of the International Conference on Computer Vision (1998)

A Fast Image Enhancement Algorithm Using Bright Channel

Long Chen[1], Wei Sun[2], and Jiaxing Feng[1]

[1] Xi'an Research Institute of China Coal Technology & Engineering Group Corp ,
82 jinye 1st Road Xi'an National Hi-tech Industrial Development Zone.
710077, Xi'an, Shaanxi, China
[2] Xidian University ,South Taibai Road 2 Xidian University 284#,
710071, Xi'an, Shaanxi, China
{chenlong,fengjiaxing}@cctegxian.com,
sunweitom@tom.com

Abstract. After summarizing the poor-illumination image enhancement methods and analyzing the shortcomings of the currently well-performed multi-scale Retinex algorithm, this paper proposed a new image speedy algorithm with detailed illumination component information. It combined illumination imaging model with target reflection features on RGB color channel, raised a new bright channel concept, and obtained computation method of illumination components by analysis. Then, illumination components were gained precisely through image bright channel gray-scale close computation and fast joint bilateral filtering. Consequently, target reflection components on RGB channel could be solved by illumination/reflection imaging model. The proposed algorithm can get excellent edge details through simple and quick computation. After being removed from the illuminative effects, the images gained are natural-colored, highly visible, and with no halo artifacts. This paper also resolved color casting problem. Compared with NASA method based on multi-scale Retinex, the proposed algorithm improved computation speed, received vivid colors and natural enhancement result.

Keywords: image enhancement, Retinex algorithm, bright channel, joint bilateral filtering.

1 Introduction

During the image shooting process, the objects in the scene will get non-uniformed illumination due to light changes, environmental illumination, object blocking, and so on, which degrades the image quality and affects the identification of the local information. And the consequent image detail feature extracting, target identification, tracking, and understanding will not be performed. Therefore, enhancing the images with non-uniformed illumination has been widely concerned. Braun and others [1] used Sigmoidal function for tone mapping, and gained function parameters by combining image mean and variance. This method is very simple and quick, but when

the image dynamic range is very large, local details will get lost. Partially overlapped sub-block histogram equalization proposed by Kim et al. [2] basically eliminated blocking artifacts in non-overlapped sub-block method, and reduced computation compared with overlapped sub-block method. Although local sub-block gray distribution was concerned by this method, image layering was weakened. Automatic color equalization mode proposed by Rizzi et al.[3] used local self-adaptive filtering when the spatial distribution of color and luminance was taken into account. It enhanced the color contrast of the image. However, for the images that do not meet the gray world assumption, this method will lead to significant color distortion. Fattal [4] transferred images into gradient domain, decreased larger gradients and increased smaller ones when keeping the gradient direction. This method enhanced the details of the dark area in the image but weakened the details of the bright area. Based on light reflection imaging model, spatial homomorphic filtering algorithm [5] took the logarithm of the image and got it through a low-pass filter, then estimated illumination components by using the filtered result. The speed and performance of this algorithm was superior to the traditional frequency homomorphic filter due to the reduction of dynamic region, but it's very hard to choose the weights of the filter model. Land [6] proposed Retinex theory by combining illumination reflection imaging model with perceptual characteristics of human visual system to luminance and color. Jobson and his fellows [7-9] proposed Single Scale, Multi-Scale and Multi-Scale Color Restoration Retinex algorithms. The result of Single Sale Retinex algorithm depended on the magnitude of the scale constant, Multi-Scale Retinex algorithm would cause halo and large computation, Multi-Scale Color Restoration Retinex algorithm[10] restored the colors of the images which violated gray-world assumption, but the color correction results were not significant.

To solve the above problems, this paper deduced a solution to illumination component in optical model after deep analysis on optical image enhancement principle. This paper raised image RGB bright channel principle from a new perspective, thoroughly studied the speedy algorithm of illumination computation, and proposed Max Intensity Channel Image Enhancement (abbr. MICIE) algorithm. The operation steps of the algorithm are as follows: 1) Make preliminary estimation of the illumination component by using RGB bright channel; 2) Make precise estimation of the illumination component by eliminating the influences of low-reflectivity to targets through fast joint bilateral filter; 3) Further compute the estimated scene illumination to get the reflection component of the target through illumination reflection imaging model, and truncate interval between [0, 1]. After the above 3 steps, the colors and illumination of the input image will be quickly restored.

2   Analysis of Multi-scale Retinex Algorithm

2.1   Illumination Reflection Imaging Model

In this model, images are composed of illumination components and reflection components[11], shown as figure 1, and the equation is:

$$I(p) = L(p)R(p) \tag{1}$$

Where $I(p)$ is image collected by the sensor, $L(p)$ is lamination component, i.e. the illumination intensity received by the objects in the scene, $R(p)$ is the target reflection coefficient, which is the inherent characteristic of the object, $p = (x, y)$ is the location of one pixel point in the image.
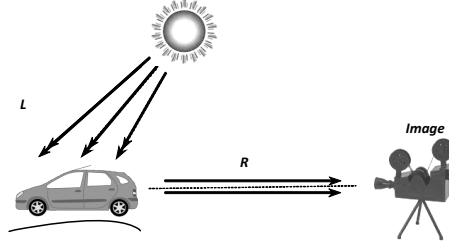

Fig. 1. Illumination imaging model

According to the model presentation in figure 1, it's known that as for image $I(p)$ which is influenced by lumination, if the scene lumination component $L(p)$ is known, reflection coefficient $R(p)$ of the target can be calculated. This means scene image which is not influenced by lumination is gained. However, only $I(p)$ in equation (1) is known, so the solution to the reflection component is an ill-conditioned problem. In general, a certain method is used to estimate lumination component, and then reflection component is solved. And Retinex algorithm is one typical representation among these methods.

2.2 Multi-scale Retinex Algorithm and Its Disadvantage

Retinex theory combined lumination reflection imaging model with perceptual characteristics of human visual system, and regarded that the objects perceived by human visual system depended on illumination and the reflection ability of the objects to illumination. SSR algorithm estimated illumination component through Gaussian filtering to the input image, and transferred data to logarithmic domain to solve reflection component. Equation is shown as follows:

$$R^c(p) = \log I^c(p) - \log[F(p) * I^c(p)] \tag{2}$$

Where $R^c(p)$ is the reflection component value of pixel point p on color channel c, $\log$ is natural logarithm, $*$ is convolution, $F(p)$ is Gaussian surround function, which is shown as:

$$F(p) = K \exp(-\frac{r^2}{c^2}) \tag{3}$$

Where K is normalization factor, c is scale. The smaller c is, the more outstanding image edge detail is, the larger dynamic region is reduced, and the more severely the colors are distorted; the bigger c is, the higher image color fidelity is, but the more indistinct the local detail is. Therefore, SSR algorithm cannot get both clear details and color restoration. While MSR algorithm can get better results in these aspects, it makes weighted average to the big, media, small scaled SSR algorithm results. It is shown as follows:

$$R_{MSR}^c(p) = \sum_{n=1}^{N} \omega_n R_n^c(p) \tag{4}$$

Where scale number N is 3, $R_n^c(p)$ is reflection component after SSR algorithm operation, ω_n is the corresponding weight of the n-th scale, and it is 0.33 when scale is 3.

Although MSR algorithm is superior to some other typical algorithm after combining the advantages of different scales, it still has two main disadvantages. Firstly, the premise of the Retinex algorithm is that illumination over the whole scene is changing slowly and evenly. In reality, illumination received by objects will have sudden changes due to illumination source changes, environmental illumination, object blocking, and so on. And these changes usually occur at the edge of the objects. As shown in figure 2(e), there's obvious halo artifacts at the edge between the white building and pine trees. Secondly, MSR algorithm has to process 3 scaled Gaussian filtering for each RGB channel respectively, 9 Gaussian filtering processes for all, which means larges computation, and is unsuitable for real-time application.

3   Image Enhancement Algorithm Based on Bright Channel and Fast Joint Bilateral Filtering

3.1   Bright Channel

In the natural illumination scene, when the reflectivity of the object in the scene approaches 1, i.e. $R(p) \to 1$, according to illumination reflection imaging model, image intensity gained will approach scene illumination intensity, that is $I(p) \to L(p)$. Therefore, as for one pixel in three RGB channel, if the intensity value of one channel is bigger, the reflectivity of this channel is also greater, and its gray value is closer to the illumination intensity of this pixel. If the largest value of each pixel in the RGB channel is chosen to form the bright channel chart, we could vividly call it Max Intensity Channel (MIC).

$$I_{MIC}(p) = \max_{c \in \{R,G,B\}} (I^c(p)) \tag{5}$$

Where, $I^c(p)$ is the gray value on a certain channel of the input image $I(p)$, $I_{MIC}(p)$ is the bright channel of the input image $I(p)$.

3.2   Estimation of Illumination Component of Reflection Component

According to equation (5), this paper obtained bright channel of the image by RGB three channels computation, and made illumination component rough estimation $L_{rough}(p)$, as shown in figure 2 (b). It is demonstrated as

$$L_{rough}(p) = I_{MIC}(p) \tag{6}$$

Obviously, the bigger the bright channel intensity value is, the closer to the scene illumination intensity. However, the objects in the scene are not all very bright-colored, and most areas are changing mildly, which means not every pixel point satisfies the conditions of $R(p) \rightarrow 1$ and equation (5). Therefore, directly using the estimation result of equation (6) will lead to distortion.

In order to eliminate the influence of low-reflectivity to targets, this paper introduced gray-scale close computation to filter the illumination component. This computation could reduce the sudden changed small dark area while keep the bright area unchanged. After this step, illumination was changing smoothly and mildly in most areas of the image. The result after processing is L_{close}, as shown in figure 2 (c), and demonstrated as

$$L_{close} = L_{rough} \bullet b = \left(L_{rough} \oplus b \right) \ominus b \tag{7}$$

Where b is the gray-scale morphological structuring element, and smoother results can be gained if disk structure is chosen.

As discussed before, there was a certain degree of illumination mutation at the edges of objects in the scene, and illumination component after gray-scale close computation lost local details and blurred edges. Therefore, in order to get smooth illumination component with edge details, joint bilateral filter was used to refine the bright channel images and results after close computation filtering. Bilateral filter [12] is an edge-preserving filter, its results are related to not only the spatial locations of the surrounding pixels, but also their gray-scale differences. As for gray-scale image I, bilateral filter is defined as

$$bf(I_p) = \frac{1}{W_p} \sum_{q \in \Omega} f(\|p - q\|) g(|I_p - I_q|) I_q \tag{8}$$

Where f is a Gaussian spatial filter whose center is located at point p, g is a Gaussian range filter of pixel gray-scale value with point p as its center, Ω is the spatial range of f, W_p is the weight sum of $f \cdot g$. Due to the introduction of

range filter, if point p is around the edge, weight of g then represents the information of the edge. When weight of g times spatial weight of f, not only edges will be preserved but also both two sides of the edge will be smoothened. Joint bilateral filter applies range filter g of bilateral filter into another image which is full of detailed information. The rough estimation of illumination component I_{rough} owns the detailed information of illumination, therefore, it is used to solve the range Gaussian kernel g. MICIE algorithm processes joint bilateral filtering for both L_{close} and L_{rough}, which can not only preserve the edge information of bright channel L_{rough}, but also further smoothen L_{close}. Namely,

$$L_{refined}(p) = \frac{1}{W_p} \sum_{q \in \Omega} f\left(\|p-q\|\right) g\left(\left|L_{rough}(p) - L_{rough}(q)\right|\right) L_{close}(q) \tag{9}$$

This paper used fast joint bilateral filter [13] to refine illumination component, which greatly improved computation speed. Refined illumination component can be seen in figure 2(d).

(a) Input Image   (b) Bright Channel image   (c) Close operation

(d) Joint bilateral filtering result   (e) Result of MSR   (f) Result of MICIE
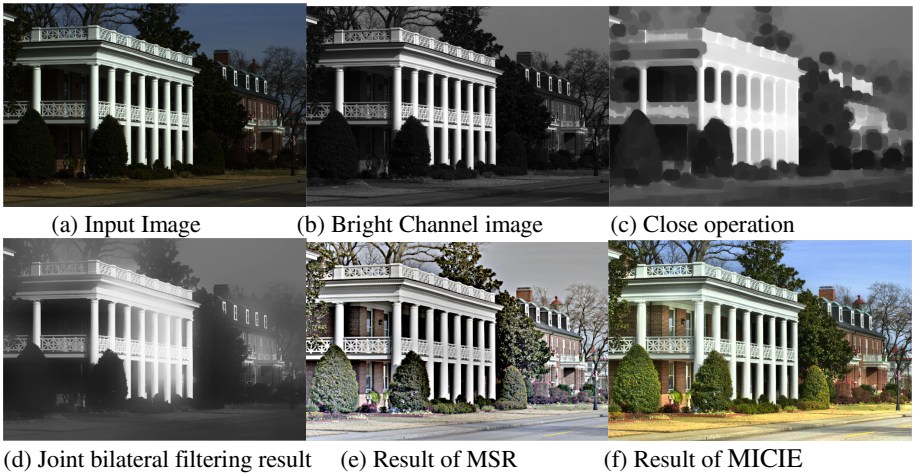
Fig. 2. Intermediate results of proposed algorithm

After estimation of illumination component, illumination reflection imaging model is transformed to solve the reflection component $R(p)$ of the scene, which is defined as equation (10), and the final enhanced result is shown as figure 2(f).

$$R(p) = \frac{I(p)}{L_{refined}(p)} \tag{10}$$

4   Comparison and Analysis of Experimental Results

4.1   Subjective Evaluation

In order to test the proposed algorithm, this paper enhanced some typical non-uniform illumination images on the NASA website [14], and the target images were composed of color-distorted and color-normal images. Figure 3 to figure 5 show the experimental results of the non-uniform illumination images without color distortions. It can be seen from figure 4 that the proposed algorithm better reappears the scene colors than NASA algorithm, and the sky is especially blue and clear. NASA algorithm uses Gaussian filter to give different weights according to target locations, which cannot preserve the illumination mutation at the scene edge; therefore, distortion of illumination estimation will occur at high-contrast edges. Comparisons between figure 3 (d) and (e) show the results of local details. There're obvious halo artifacts at the edge of white tower in figure 3 (d), while bilateral filter can adjust filter weights according to pixel gray-scale difference in the neighborhood, thus, halo is eliminated by this algorithm. Figure 4 demonstrates the input image and the details of the white sole base by these two algorithms. NASA algorithm preserves small part of soil at the sole edge, the white part is over-enhanced, and the colors are too saturated. While the proposed MICIE algorithm well preserves the sole details of the input image, and the colors are more natural and vivid. Detailed comparison can be seen in figure 4 (d) and (e). In figure 5, both two algorithms restore the details of the darker area (car window), but compared with result of NASA algorithm, the restored image by proposed MICIE algorithm are with more vivid colors and better-visible details. For example, the car compartment and house reflected on the window glass in figure 5 (e) are clearer than the ones by NASA algorithm in figure 5 (d), and the sky on the top right corner of the image is better restored.
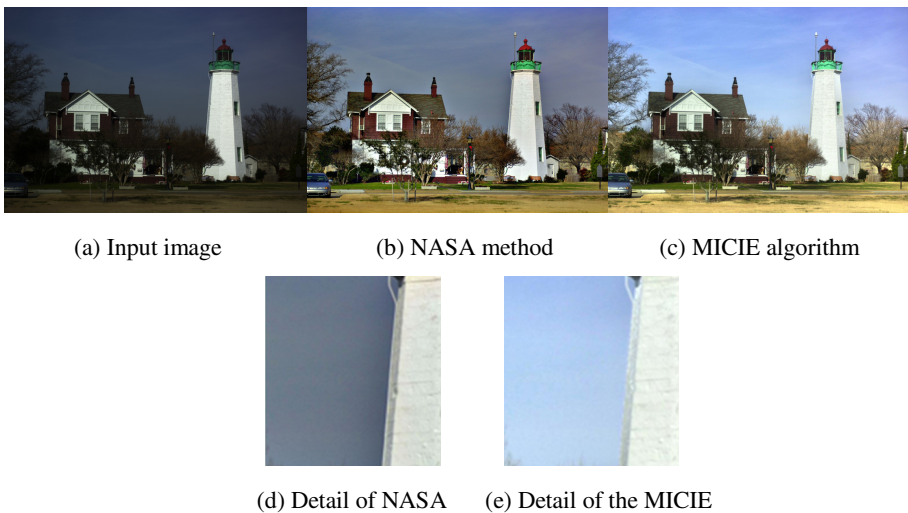

(a) Input image   (b) NASA method   (c) MICIE algorithm
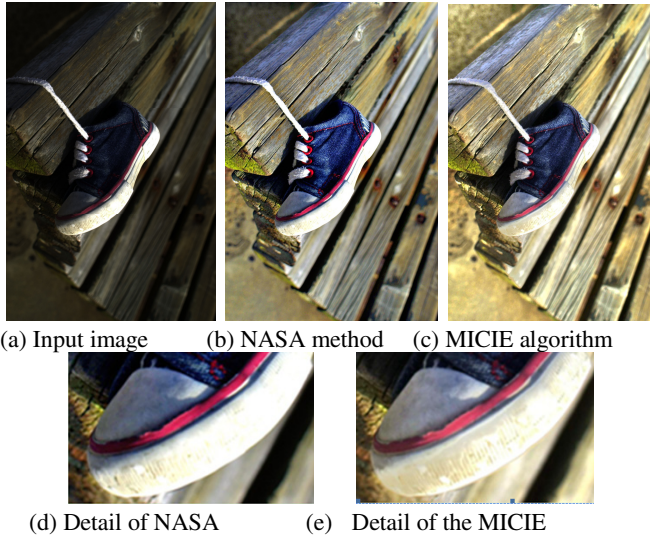

(d) Detail of NASA   (e) Detail of the MICIE

Fig. 3. Comparison of different algorithms

(a) Input image   (b) NASA method   (c) MICIE algorithm

(d) Detail of NASA   (e) Detail of the MICIE

Fig. 4. Comparison of different algorithms

(a) Input image   (b) NASA method   (c) MICIE algorithm

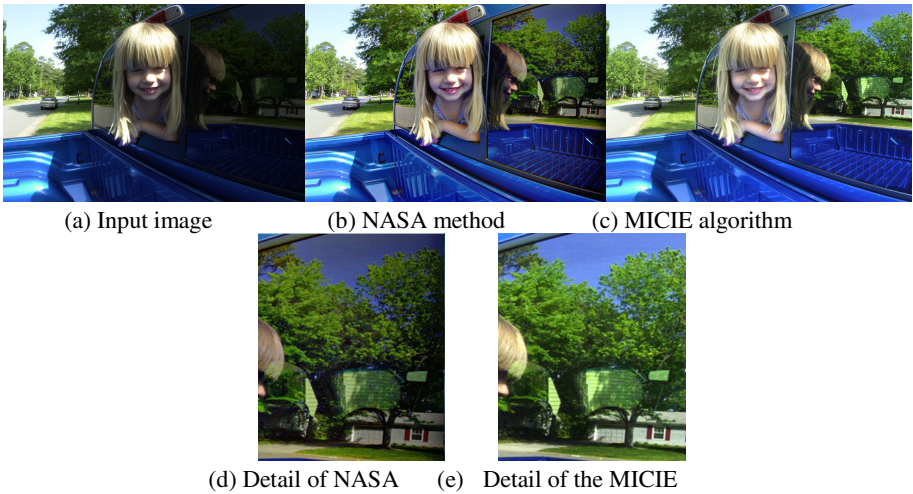

(d) Detail of NASA   (e)   Detail of the MICIE

Fig. 5. Comparison of different algorithms

4.2   Objective Evaluation

This paper made objective comparisons to the experimental results by using standard deviation and information entropy (See Table 1). Standard deviation was used to judge the image contrast, and information entropy reflected the size of the information amount included in the image. Since NASA algorithm is based on multi-scale Retinex, Table 2 showed the computing time of the proposed algorithm and multi-scale Retinex algorithm under spatial and frequency domains.

Table 1. Results comparison of different algorithms

Image	Algorithms	Std	Entropy
tower(Fig3(a))	NASA	48.4251	7.6255
2000x1312	MICIE	54.7450	7.7967
shoe(Fig4(a))	NASA	61.9346	7.6407
1312x2000	MICIE	69.1627	7.9105

Table 2. Computing time of different algorithms

Image	Algorithms	Time/s
tower	Spatial MSR	66.72
2000 x 1312	Frequency MSR	20.83
	MICIE	10.66
shoe	Spatial MSR	72.47
1312 x 2000	Frequency MSR	16.61
	MICIE	10.57

It can be seen from table 1 that the image contrast of this algorithm is close to the one of NASA algorithm, and more information of the enhanced image can be gained by MICIE algorithm than by NASA algorithm, so this algorithm can enhance the scene details well. Table 2 shows that MICIE algorithm has greatly improved computing time compared with multi-scale Retinex algorithm.

5   Conclusion

This paper proposed bright channel concept according to reflective characteristic of objects, and calculated illumination component fast and precisely from a new perspective. In this paper, image bright channel was used for rough estimation of the scene illumination, and gray-scale morphological filter and fast joint bilateral filter were used for the refinement of the rough estimation. By this method, the illumination estimated was kept smooth while detailed information of the illumination mutation was also preserved. The final reflection component images gained through illumination reflection imaging model were of vivid colors, outstanding details, and high visibility. This algorithm greatly improved computing time as well as well eliminated halo artifact which was the typical disadvantage of Retinex algorithm. As for color-distorted non-uniform illumination image, this paper also used gain offset correction and information statistics to further correct the colors and obtained good correcting results.

References

1. Braun, G.J., Fairchild, M.D.: Image lightness rescaling using sigmoidal contrast enhancement functions. Journal of Electronic Imaging 8, 380–393 (1999)
2. Kim, J.Y., Kim, L.S., Hwang, S.H.: An advanced contrast enhancement using partially overlapped sub-block histogram equalization. IEEE Transactions on Circuits and Systems for Video Technology 11, 475–484 (2011)
3. Rizzi, A., Gatta, C., Marini: A new algorithm for unsupervised global and local color correction. Pattern Recognition Letters 24, 1663–1677 (2003)
4. Fattal, R., Lischinski, D., Werman, M.: Gradient domain high dynamic range compression. In: Proc. of ACM, SIGGRAPH 2002, pp. 249–256. ACM, New York (2002)
5. Xiao, J., Song, S.H.P., Ding, L.J.: Research on the fast algorithm of spatial homomorphic filtering. Journal of Image and Graphics 3, 2302–2305 (2008)
6. Land, E.H.: An alternative technique for the computation of the designator in the retinex theory of color vision. Proceedings of the National Academy of Sciences (1986)
7. Jobson, D.J., Rahman, Z., Woodell, G.A.: Properties and performance of a center/surround retinex. IEEE Transactions on Image Processing 6, 451–462 (1997)
8. Jobson, D.J., Rahman, Z., Woodell, G.A.: A multi-scale retinex for bridging the gap between color images and the human observation of scenes. IEEE Transactions on Image Processing 6, 965–976 (1997)
9. Rahman, Z., Jobson, D.J., Woodell, G.A.: Retinex processing for automatic image enhancement. Journal of Electronic Imaging 13, 100–110 (2004)
10. Kimmel, R., Elad, M., Sobel, I.: A variational framework for Retinex. International Journal of Computer Vision 52, 7–23 (2003)
11. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Publishing House of Electronics Industry, Beijing (2007)
12. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: International Conference on Computer Vision, pp. 839–846. IEEE Press, Bombay (1998)
13. Paris, S., Durand, F.: A fast approximation of the bilateral filter using a signal processing approach. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 568–580. Springer, Heidelberg (2006)
14. NASA Research Center, http://dragon.larc.nasa.gov

Software Analysis for Transient Faults:
A Review of Recent Methods

Guochang Zhou[1], Baolong Guo[2], Xiang Gao[1], Weikang Ning[2], and Yunyi Yan[2]

[1] China Academy of Space Technology(Xi'an)
Xi'an 710000, China
[2] School of Aerospace Science and Technology
Xidian University
Xi'an 710071, China
`yyyan@xidian.edu.cn`

Abstract. Transient faults in a computer system could influence the behavior of the software and pose a big threat to the dependability of the system. So a variety of software analysis methods are invented to characterize the property of the program in the face of transient faults and yield a lot of instructive information that could be utilized to improve the dependability of the software. In this paper, we first summarize some typical software analysis methods that are related to transient faults and introduce each method briefly. Then we make a comment on them and recommend some prospective methods.

1 Introduction

Transient faults or soft faults refer to the faults that appear in the circuits and are not permanent, which are contrary to hard faults. The causes of transient faults include radiation, electromagnetic interference, power glitches, etc. For example, a single event upset (SEU) is a most common transient fault that the state of one bit in a space computer changing form 0 to 1 or from 1 to 0, which is usually caused by the strike of cosmic rays.

The faults may influence the behavior of the program and incur unpredictable results. In [1], the author presents a comprehensive picture of the impact of faults on LynxOS key features by means of a Software-Implemented Fault Injection tool. The results indicate that the impact of the faults is related to many factors, such as the location of the faults, the OS, the characteristic of the program, etc.

Many Software-Implemented Fault Tolerance techniques[2] are used to mitigate the influence of these faults. However, the techniques may introduce considerable overhead. So it becomes necessary to find out the most vulnerable or sensitive parts of the program and make a trade-off between performance and cost.

To analyze the impact of transient faults and get some useful information to mitigate their impact, a lot of work has been done from different point of views and have achieved some instructive conclusions. In [3], a structure's architectural vulnerability factor (AVF) is defined as the probability that a fault in

that particular structure will result in an error, which is used to quantitatively describe the vulnerability of an architecture. In [4], the author reviews some main approaches that are used to evaluated the effect of Single Event Transients (SETs) and Single Event Upsets (SEUs) in digital circuits described at different abstraction levels.

Most work focuses mainly on the analysis from the hardware perspective, just as the work mentioned above, the analysis from the software viewpoint is rather rare. So in this paper, we focus on the introduction of some purely software analysis methods that have been proposed recently.

The paper is organized as follows: In section 2, some basic concepts and the classification of software analysis are first introduced briefly. Then some program analysis techniques that are used in the face of transient faults are introduced. In section 3, we make a comment on the techniques introduced in section 2 and make a comparison between them. Finally, some prospective methods are listed out in section 4.

2   Software Analysis

We define software analysis as the process of the extraction of some useful information from the program. From the software reliability viewpoint, the information collected could be further utilized to improve the software dependability. There exist a variety of software analysis techniques [5]and it's impossible for us to cover all of them. In this paper, the scope of program analysis is limited. The software analysis techniques we are going to introduce are all related to transient faults. More specifically, what we are mainly concerned about includes: the sensitivity of the program to transient faults, the effect of transient faults on the behavior of the program, and the generation and propagation of transient faults in the program.

2.1   The Classification of Software Analysis

The software analysis methods could be classified according to different standards. The most commonly used standard is that if the program is executed during the process of analyzing. Based on this standard, the program analysis methods could be classified into two types: static analysis and dynamic analysis. The program analysis techniques we are going to introduce are cataloged based on this standard.

2.2   Static Analysis

Contrary to dynamic analysis, static analysis does not require that the program analyzed being executed. There are a variety of techniques and tools that are used to analyze the program statically. And different techniques could usually get different information of the program.

Some basic static analysis techniques, such as syntax analysis, control flow analysis, data flow analysis, etc, have been widely used in all kinds of compilers. Some basic static analysis techniques are also used to analyze the influence of transient faults on the program. However, related work is rather rare.

The static analysis methods known could be further classified based on a variety of standards. The program is usually analyzed from different point of views, such as the level of the language, the information one could collect, the type of transient faults, etc. In this paper, the available static analysis techniques we are going to introduce are organized based on the type of transient faults.

2.2.1   Pure Data Error Analysis

The most dangerous and deceptive situation is that the program terminates normally while the results are not correct due to the faults. This situation is mostly caused by the faults that appear in the memories storing data or in the registers. It's worthwhile to analyze the program in the face of this kind of error.

(a) In [6], the author proposes a C/C++ source-to-source compiler that could improve the dependability of C/C++ programs by means of code re-ordering and variable duplication. The approach adopted by the author is pure software and automated by means of a tool named RECCO. It could analyzes the program first and then transform the source code accordingly.

To find the most vulnerable variables, a reliability-weight is calculated for each variable in the program. The reliability-weight of a variable is calculated based on its life time and functional dependencies, which takes both error generation and error propagation properties into consideration. The higher the reliability-weight of a variable, the more sensitive it is. To reduce the reliability-weight of the variables, an local optimization approach by means of code-reordering is first implemented. Besides, based on the ranking of reliability-weight of the variables in the program, the user could choose the percentage of variables to be protected (duplicated), thus makes it possible to make a trade-off between dependability improvement and performance.

(b) In [7], a probabilistic model is built to describe the propagation of errors in the software.

The errors are categorized into two kinds: original errors and propagated errors. The basic idea is that errors are propagated through the program via computation. To analyze the error propagation properties, the computational data flow model is first built for a program, then the error flow model is set up based on this model. Besides, some rules about error propagation are given by the author. Based on the rules, the probability of an error at the program could be calculated.

(c) Considering the intrinsic inaccuracy of floating-point numbers, the final output of a program may not be affected by the faults under some conditions. In [8], the relation between bit-flip rate and the error of floating-point arithmetic operations is analyzed. The error in a floating-point number is limited in the fractional part.

2.2.2   Code and Data Error Analysis

Contrary to the situation of purely data error, the behavior of the program in the face of code error is usually more complex.

(a) In [9], a methodology named Fault Effect Analysis (FEA) is introduced by the author to calculate the probabilities of different possible behaviors of a given application affected by a transient fault. The fault model is assumed to be a single bit-flip or SEU in the code of the target application and the possible results of an application are summarized as: illegal instruction, system exception and normal termination (undefined results).

The results are achieved by means of analyzing the instruction sequence statically. A comparison is made between FEA and traditional fault injection methods.

(b) To analyze the sensitivity of a program to transient faults, two methods are proposed in [10] and [11] from different point of views.

In [10], some techniques are proposed to characterize the vulnerability of software resource to hardware faults in a micro-architectural independent way.

The techniques introduced derive from the techniques in [3] and the vulnerability of a program is quantified by Program Vulnerability Factor (PVF), which is related to ACE-bits of the architecture.

In [11], a new approach called PRASE is introduced to assess the reliability of a given program under the occurrence of soft errors. To quantify the vulnerability of the program, a Program Vulnerability Factor (PVF) is also defined.

The techniques used to calculate PVF includes mainly four parts: a probabilistic analysis of soft error generation, an analysis of soft error propagation, an optimization method using basic block analysis and a method to calculate the reliability of a given program.

2.3   Dynamic Analysis

Dynamic analysis means that the program to be analyzed must be executed to collect the information wanted. Compared with static analysis, dynamic analysis is usually more accurate but more limited because the input data is just a subset all the possible input data.

Fault injection could be seen as a kind of dynamic analysis method. It is the most widely used method in system analysis and a lot of fault injection techniques and tools[12] are invented. The faults are injected into the program while the program is running. And the property of the program is achieved by means of analyzing the results of the program after a finite number of experiments.

They are used not only in analyzing the program directly but also in verifying some other program analysis methods or models. In the following of this section, we will introduce some typical work that uses fault injection techniques to analyze the program.

(a)In [13], the author introduces a framework called EPIC that is used to profile the propagation and effect of data errors in software.

The whole system is first divided into many modules and each module is viewed as a black-box with multiple input and output signals. Then error

permeability is defined to quantitatively characterize the error propagation property between an input and output signal of a module. Based on the error permeability defined, a set of related measures are introduced to provide an insight on the error propagation and effect characteristics and vulnerabilities of a system. All the measures introduced could be classified into four groups including exposure, permeability, impact and criticality, just as the name EPIC indicates.

The hot-spots and vulnerabilities of a system could be identified using these measures. Based on the information collected, the designer could make a trade-off between system dependability and cost.

However, it may be hard to achieve the numerical values for the error permeability considering many factors. Thus fault injection is adopted by the author to achieve the estimates of error permeability values. The program is first run without faults, which is called the golden run, then different faults are injected into the system in the following experiments and the result is compared with the result of golden run.

(b)The static analysis techniques listed in 2.2 are usually proposed based on some experiments or experience, their effectiveness is usually tested by means of fault injection techniques, as is shown in [6][11].

3   Comparison and Comments

Just as we have introduced, there are a variety of techniques that are used to analyze the program. However, the methods are different in cost, accuracy, adaptability, etc. So it's important to choose the most proper method for a specific application.

Compared with fault injection techniques, most static analysis techniques are totally automated so the analyzing time could be shortened considerably. Although there are some fault injection tools that are used to test the program, the faults injected have to be prepared in advance.

The information collected by static analysis methods is usually more comprehensive. This may be explained from two aspects. First, the input data of fault injection techniques is limited. Second, the code being executed may not cover all the branches in the program. However, the information collected by fault injection techniques may be more accurate if the input data is well selected.

Besides, the techniques adopted by different methods usually varies a lot. This could be explained from several aspects.

First of all, the viewpoints adopted by different methods are different. A program is a structured text that contains abundant information. The program is usually quantitatively profiled based on different characteristics of the program.

Second, the fault models adopted are different. For example, the faults are further classified into data and code faults in section 2.2.

Third, the application is analyzed based on different language levels, from executable language to different high level languages.

Considering all these differences, the result obtained by different methods are not exactly the same. Some methods provide more macro descriptions of the program's vulnerabilities while the others provide more detailed ones.

4   Future Work

On the one hand, the more accurate the information we could collect from program analysis, the more effective our software protection will be.Thus it becomes necessary to build more accurate models for a specific purpose.

On the other hand, as the software becomes more and more complex, the cost of program analysis becomes high, thus it becomes necessary to invent some tools that could analyze the program automatically.

The compiler could collect a lot information about the program and could use the information collected to optimize the program. Besides, the process could to be highly automated and the process of program analysis or program enhancement is usually transparent for the programmers. To analyze the program and improve the dependability of the program by means of compiler seems an attractive way.

References

1. Madeira, H., Some, R.R., Moreira, F., Costa, D., Rennels, D.: Experimental evaluation of a cots system for space applications. In: Proceedings of the International Conference on Dependable Systems and Networks, DSN 2002, pp. 325–330. IEEE (2002)
2. Xu, J.J., Tan, Q.P., Xiong, Y.Q., Tan, L.F., Li, J.I.: Software fault-tolerance techniques for transient faults. Computer Engineering & Science 11, 030 (2011)
3. Mukherjee, S.S., Weaver, C., Emer, J., Reinhardt, S.K., Austin, T.: A systematic methodology to compute the architectural vulnerability factors for a high-performance microprocessor. In: Proceedings of the 36th Annual IEEE/ACM International Symposium on Microarchitecture, vol. 29. IEEE Computer Society (2003)
4. Anghel, L., Leveugle, R., Vanhauwaert, P.: Evaluation of set and seu effects at multiple abstraction levels. In: 11th IEEE International On-Line Testing Symposium, IOLTS 2005, pp. 309–312. IEEE (2005)
5. Hong, M., Qian-Xiang, W., Lu, Z., Ji, W., et al.: Software analysis: a road map. Chinese Journal of Computers 32, 1697–1710 (2009)
6. Benso, A., Chiusano, S., Prinetto, P., Tagliaferri, L.: A c/c++ source-to-source compiler for dependable applications. In: Proceedings of the International Conference on Dependable Systems and Networks, DSN 2000, pp. 71–78. IEEE (2000)
7. Xue-Jun, Y., Long, G.: Error flow model: modeling and analysis of software propagating hardware faults. Journal of Software 18, 808–820 (2007)
8. Li, S., Li, X.: Soft error propagation in floating-point programs. In: 2010 IEEE 29th International Performance Computing and Communications Conference (IPCCC), pp. 239–246. IEEE (2010)
9. Benso, A., Di Carlo, S., Di Natale, G., Prinetto, P.: Static analysis of seu effects on software applications. In: Proceedings of the International Test Conference, pp. 500–508. IEEE (2002)
10. Sridharan, V., Kaeli, D.R.: Quantifying software vulnerability. In: Proceedings of the 2008 Workshop on Radiation Effects and Fault Tolerance in Nanometer Technologies, pp. 323–328. ACM (2008)

11. Xu, J., Shen, R., Tan, Q.: Prase: an approach for program reliability analysis with soft errors. In: 14th IEEE Pacific Rim International Symposium on Dependable Computing, PRDC 2008, pp. 240–247. IEEE (2008)
12. Hsueh, M.C., Tsai, T.K., Iyer, R.K.: Fault injection techniques and tools. Computer 30, 75–82 (1997)
13. Hiller, M., Jhumka, A., Suri, N.: Epic: Profiling the propagation and effect of data errors in software. IEEE Transactions on Computers 53, 512–530 (2004)

A Novel LCD Driver Chip for Brightness Adjustment System

Hui Wang[1,2], Song-lin Wang[1,2],
Ling-xia Zhang[1], and Yunyi Yan[1]

[1] School of Aerospace Science and Technology
[2] Key Lab of High-Speed Circuit Design and EMC, Ministry of Education
Xidian University, Xi'an 710071, China
`whui94@126.com, slwang@mail.xidian.edu.cn,`
`zlxnpu@163.com, yyyan@xidian.edu.cn`

Abstract. Based on the liquid crystal display, in order to obtain the best image display effect and how to prolong the life of the liquid crystal the two problems, basing on BiCOMS technology of Samsung company, we design and implement a simple structure, economical and practical LCD brightness control circuit and a polarity switching circuit. The relationship between the gray voltage and gray scale brightness using in the circuit, so we can design economical and practical LCD brightness control circuit; at the same time in order to extend the life of liquid crystal, we design a polarity switching circuit, the value of liquid crystal voltage providing the data driving circuit become near zero in the average time and reduce the DC component, thus it can extend the life of liquid crystal.

Keywords: LCD, gray scale voltage, brightness mediation, polarity switch.

1 Introduction

With the advantage of small volume, thin thickness, light weight, low consumption, no radiation, no flicker and so on, Liquid crystal display (LCD) become one of the most rapid development, and gradually mature industry after the semiconductor industry. The best image display effect and the life of liquid crystal are the keys in display technology, therefore in the liquid crystal display, design of brightness adjustment module and polarity switch module are very important, it not only can control and adjust the LCD display effect, but also can effectively extend the life of liquid crystal. Based on a BiCOMS LCD driver chip, the author use the key factor that liquid crystal display quality is very dependent on the gray level, come up with a novel and economical and practical LCD brightness control circuit, it can realize adjustment of multistage brightness of various users; the author further design a mature principle and simple structure polarity switching circuit that makes the data driving circuit for liquid crystal voltage value on the time average approximation is 0, so as to reduce the DC component, to prolong the life of liquid crystal. Compared with the traditional design

idea, brightness adjusting circuit not only have principle mature and practical, but also effectively reduce the chip area; while the polarity switching circuit can extend the life of LCD, at the same time no additional power consumption. The test results show that the circuit working in good condition.

2   The Proposal and Analysis of the Design Thoughts

2.1   The Basic Principle of LCD Display

Liquid crystal display (referred to as LCD) display principle: putting the liquid crystal between two pieces of conductive glass, depending on electric field between two electrodes to drive, causing the liquid crystal molecules twisted to the column of electric field effect, in order to control the light transmission or shadowing function. Figure 1 shows the principle of TN-LCD display in power and without electric condition, wherein the polarizer allowed to pass light vibrating in a certain direction, when the glass substrate without added voltage, incident light along the arrangement of the LCD screen liquid crystal molecules are distorted by 90°, go through parallel along under the direction of the polarizer, the state of light is defined as white; when the glass substrate with voltage, the liquid crystal molecules under electric field, arranged along the direction of the electric field, light rays pass through the liquid crystal molecular space to maintain the original direction, it is covered by under polarizing plate, light is absorbed not revealing, this state of light is defined as black. The display of intermediate color between black and white depends on the applied voltage. According to the voltage, the liquid crystal display make the panel arrive a predetermined display effect.
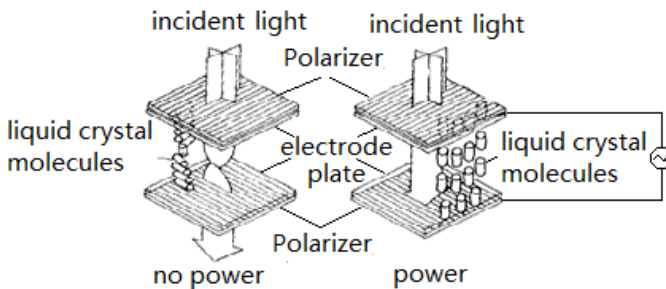
Fig. 1. Schematic structure of liquid crystal display device

2.2   The Proposal and Analysis of the Design Thoughts

In order to achieve better image display effect, LCD driver chip usually will have the brightness adjustment function module, the module provide R, G, B brightness

adjustment and signal polarity switch selection for all levels of users. For convenience to describe, as shown in Figure 2, is module diagram of brightness adjusting circuit of a digital camera LCD driver chip, we use it as an example.
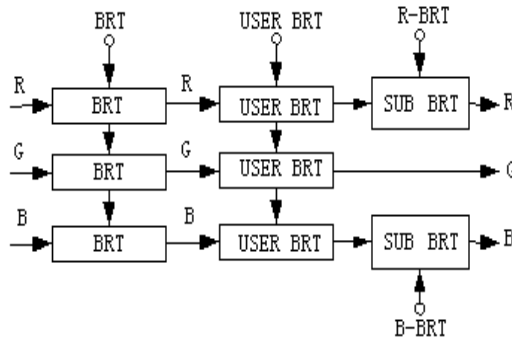
Fig. 2. Brightness control system block diagram

In order to meet the different needs of all levels of users to the image display effect, and the low power portable product requirements, the module must complete the following functions:

(1) provide brightness adjustment interface for direct users of the chip (such as: digital camera manufacturers); (2) completed the USER_BRT regulation (user brightness adjustment): provides the brightness adjustment interface indirect users of the chip(such as: digital camera user); (3) completed the SUB USER_BRT regulation (sub brightness adjustment): provide color temperature adjustment interface for indirect users of the chip; (4) provide positive, negative signals of different types of LCD required.

As shown in Figure 2, R, G, B signals is the result of a sample / hold circuit which was used as the input signal to brightness adjustment circuit, the user can through serial bus, after D/A conversion circuit to generate BRT, USER_BRT, R_BRT analog control signals, at the same time can adjust in R, G, B three signals brightness, can also be brightness adjustment on R, B signals. In order to realize aforementioned function, we need to design the following two circuits: brightness control circuit and polarity switching circuit.

3   The Realization of the Design

3.1   Brightness Control Circuit

The concrete circuit of brightness adjusting circuit as shown in Figure 3, it consist of three differential amplifiers, the circuit is mainly composed of voltage /current converter and proportional current mirror, to achieve the reference current of positive and negative polarity of polarity switch module by selecting the a and b values, for the extended linear region, increases the degeneration resistor of emitter R_1, R_2, R_3 and R_4.
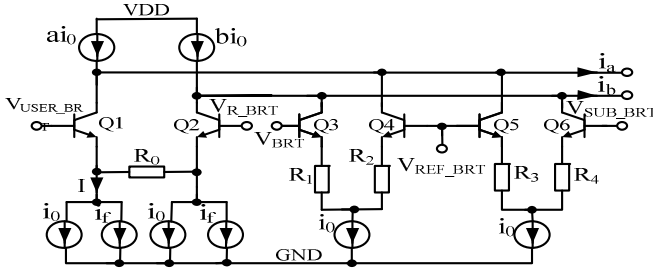
Fig. 3. R brightness adjustments

For the first differential pair (Q1, Q2), when a differential signal input, the I_{R0} current flowing through the resistor R_0, V_{R0} was defined as the voltage on the R_0, $V_1 = V_{USER\_BRT} - V_{R\_BRT}$. that is:

$$V_1 - V_{R_0} = V_{be\_Q1} - V_{be\_Q2} = V_T \ln \frac{i_{Q1} \times I_{SSQ2}}{i_{Q2} \times I_{SSQ1}} = V_T \ln \frac{I + I_{R_0}}{I - I_{R_0}} = V_T \ln \left(\frac{1 + \dfrac{V_{R_0}}{IR_0}}{1 - \dfrac{V_{R_0}}{IR_0}} \right) \tag{1}$$

When using the appropriate I and R_0 let the following formulas establish:

$V_T \ll V_{R0} \ll 0.9IR0$, $V_1 - V_{R0} \approx 0$, and ignore the effect of the base current, then:

$$i_{e\_Q2} = i_{c\_Q2} = I + \frac{V_{R\_BRT} - V_{USER\_BRT}}{R_0} \tag{2}$$

$$i_{e\_Q1} = i_{c\_Q1} = I - \frac{V_{R\_BRT} - V_{USER\_BRT}}{R_0} \tag{3}$$

Similarly, we can obtain i_{c_Q3}, i_{c_Q4}, i_{c_Q5} and i_{c_Q6}, according to KCL:

$$i_a = ai_0 - i_{c\_Q1} - i_{c\_Q4} - i_{c\_Q5} \tag{4}$$

$$i_b = bi_0 - i_{c\_Q2} - i_{c\_Q3} - i_{c\_Q6} \tag{5}$$

So:

$$i_a = (a-2) \times i_0 - i_f + \frac{V_{R\_BRT} - V_{USER\_BRT}}{R_0} + \frac{V_{BRT} - V_{REF\_BRT}}{R_1 + R_2} + \frac{V_{SUB\_BRT} - V_{REF\_BRT}}{R_3 + R_4} \tag{6}$$

$$i_b = (b-2) \times i_0 - i_f - \frac{V_{R\_BRT} - V_{USER\_BRT}}{R_0} - \frac{V_{BRT} - V_{REF\_BRT}}{R_1 + R_2} - \frac{V_{SUB\_BRT} - V_{REF\_BRT}}{R_3 + R_4} \tag{7}$$

By analysis of the AC small signal of equivalent circuit, the output stage current i_a (small signal AC) of V_{USER_BRT} and the output stage current i_b (small signal AC) of V_{R_BRT} equal in size, in the opposite direction (a phase difference of 180 °). At the same

time they have unequal DC level, due to active load of the output stage is the proportion of current source, it lead to bias current i_a, i_b different. The a and b values can be realized by proportional current mirror.

3.2 Polarity Switching Circuit

The input of R, G, B signals require periodic change its signal polarity. Let V_{COM} be a public electrode voltage of liquid crystal, every time R, G, B signal changes the polarity, signal with centered of VCOM flip up and down, the average voltage is the value of V_{COM}, POL_SW is the A/D module to the input of a control signal of the module, support the completion of the polarity reversal of R, G, B, and convert the current signal into a voltage signal to output. The concrete circuit is shown in figure 4.
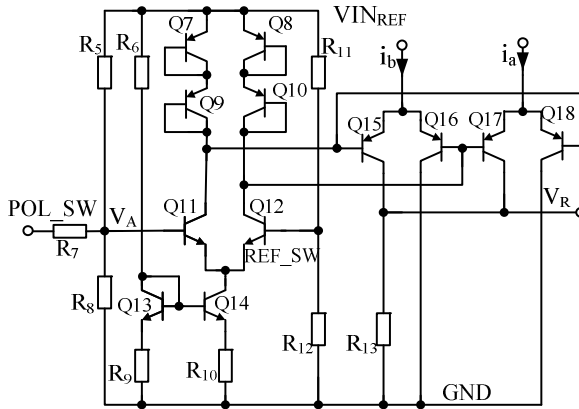
Fig. 4. Polarity switch architecture diagram

Consider Figure 4 TL ring 1:Q8, Q10, Q17, Q18, Q9, Q7, according to the law:

$$i_{c\_Q8} \times i_{c\_Q10} \times i_{c\_Q18} = i_{c\_Q17} \times i_{c\_Q9} \times i_{c\_Q7} \tag{8}$$

And:

$$i_{c\_Q17} + i_{c\_Q18} = i_a, i_{c\_Q8} = i_{c\_Q10}, i_{c\_Q9} = i_{c\_Q7} \tag{9}$$

So:

$$i_{c\_Q17} = \frac{i_{c\_Q7}^2}{i_{c\_Q8}^2 + i_{c\_Q7}^2} \times i_a \tag{10}$$

Similarly, consider the TL ring 2:Q8, Q10, Q16, Q15, Q9, Q7. That is

$$i_{c\_Q15} = \frac{4i_{c\_Q8}^2}{4i_{c\_Q8}^2 + i_{c\_Q7}^2} \times i_b \tag{11}$$

Thus:

$$V_R = R_{13} \times \left(i_{c\_Q15} + i_{c\_Q17}\right) = R_{13} \times \left(\frac{i_{c\_Q7}^2}{i_{c\_Q8}^2 + i_{c\_Q7}^2} \times i_a + \frac{4i_{c\_Q8}^2}{4i_{c\_Q8}^2 + i_{c\_Q7}^2} \times i_b\right) \tag{12}$$

If we ignore the base current, the value of i_{c_Q8} and i_{c_Q7} is equal to the collector current Q11, Q12, and depends on the base of potential V_A, V_{REF_SW} of Q11, Q12.

The circuit is shown in Figure 4, let VIN_{REF}=3V, R_{12}=40K, R_{11}=60K Ω, we can available:

$$V_{REF\_SW} \approx \frac{R_{12}}{R_{11} + R_{12}} \times VIN_{REF} \approx 1.2V \tag{13}$$

When SW_POL= "1", and let VIN_{REF}=3V, R_7=56K, R_{10}=36.7K, R_{S5}=47K, so we can get:

$$V_A \approx \frac{R_8}{R_8 + (R_5 \;//\; R_7)} \times 3V \approx 0.81V \tag{14}$$

At this time, Q12 conduction, Q11 approximate cut off, $i_{c\_Q8} \gg i_{c\_Q7}$, $V_{OUT} \approx R_{13} \times i_b$; When the SW_POL= "0",

$$V_A \approx \frac{R_7 \;//\; R_8}{R_5 + (R_7 \;//\; R_8)} \times VIN_{REF} \approx 1.6V \tag{15}$$

At this time, Q11 conduction, Q12 approximate cutoff, $i_{c\_Q8} \ll i_{c\_Q7}$, $V_{OUT} \approx R_{13} \times i_a$; let a=4, b=5, $R_1=R_2$, $R_3=R_4$, and $R_4=2R_2$, are:

$$\begin{cases} \text{positive:} V_R = R_{13}\left[\left(2i_0 - i_f + \frac{V_{R\_BRT(d)} - V_{USER\_BRT}}{R_0} + \frac{2V_{BRT} + V_{SUB\_BRT} - 3V_{REF\_BRT}}{4R_2}\right) + \frac{V_{R\_BRT(a)}}{R_0}\right] \\ \\ \text{negtive:} V_R = R_{13}\left[\left(3i_0 - i_f + \frac{V_{R\_BRT(d)} - V_{USER\_BRT}}{R_0} - \frac{2V_{BRT} + V_{SUB\_BRT} - 3V_{REF\_BRT}}{4R_2}\right) - \frac{V_{R\_BRT(a)}}{R_0}\right] \end{cases}$$

$V_{R\_BRT(d)}$, $V_{R\_BRT(a)}$ are respectively for DC and AC component of the input voltage, i_f is the feedback current. From the above formula, through adjust signal relationship between the signal V_{BRT}, V_{USER_BRT}, V_{SUB_BRT}, V_R and polarity. We can see that the function of adjustment of the input signal of the overall brightness and R signal brightness of the circuit work is very good.

4   Experimental Results and Analysis

Figure 5 is the simulation waveform of the circuit, we can see from the simulation waveform: when the brightness adjustment and input Vpp of polarity switch is 357mV, the output Vpp is 938mV (including the polarity reversal voltage interval, the actual exchange gain is 1), mach the design requirements.
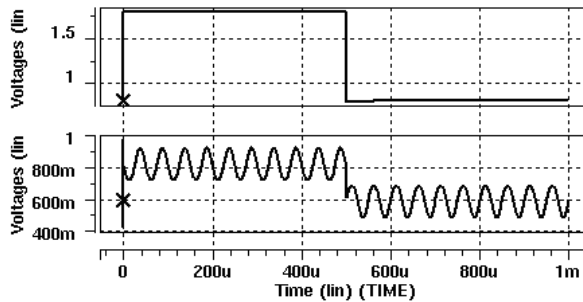
Fig. 5. R polarity switch simulation curve

5   Conclusion

With the development of the information industry and the microelectronics technology, the technology of liquid crystal display is constantly breakthroughs, with the advantage of small size, thin thickness, light weight, low consumption, no radiation, no flicker and so on, the liquid crystal display attract more and more attention, and has made a great progress in technology. Brightness mediation circuit introduced by this paper uses the BiCMOS technology of Samsung company, the design has been applied in a LCD drive chip, and use Cadence, Hspice and other EDA tools to complete the circuit design and simulation verification.

References

[1] Li, W., Guo, Q.: Applied Technology of Liquid Crystal Display. Publishing House of Electronics Industry, Beijing (1999)
[2] Li, Y., Lai, X.: Integrated and Design Technology of Electronic System. Publishing House of Electronics Industry, Beijing (2002)
[3] Zhu, Z.: The Semiconductor Integrated Circuit. Tsinghua University Press, Beijing (2001)
[4] Tong, S., et al.: Basic simulated electronic technology, 3rd edn. Higher Education Press, Beijing (2002)
[5] Lee, et al.: A Low-Power Poly-Si TFT-LCD with Integrated 8-bit Digital Data Drivers. In: Proceedings of the 18th International Display Research Conference, CONF. 18, September 28, pp. 285–288 (1998)
[6] Liang, Z., Du, Y., Liu, R., et al.: Design and realization of LCD driver module based on LPC2478. Electric Power Automation Equipment 30(7), 137–140 (2010)

Arrhythmias Classification Using Singular Value Decomposition and Support Vector Machine

Tomáš Peterek[1], Lukáš Zaorálek[1,2], Pavel Dohnálek[1,2], and Petr Gajdoš[1,2]

[1] IT4innovations, VSB - Technical Univesity of Ostrava,
Ostrava 708 33, Czech republic
tomas.peterek@vsb.cz
http://www.it4i.cz/
[2] Department of Computer Science, FEECS, VSB - Technical Univesity of Ostrava,
Ostrava 708 33, Czech republic

Abstract. The main aim of this work is to recognize arrhythmias in ECG records. Many algorithms for this task have been proposed in the past, but in our solution we try to reduce redundancy of information in the signals by Singular Value Decomposition. The reduced dataset is classified by Support Vector Machine. Our approach gives very satisfactory results which can be used in medical practice. This expert system should offer automated recognition between physiological beat and one of the three basic pathological beats: Premature ventricular contractions, Right bundle branch block and Left bundle branch block.

Keywords: SVD, SVM, Arrhythmias, PVC, LDA.

1 Introduction

The Electrocardiography (ECG) is a non-invasive diagnostic method which measures differences in electrical potentials in the heart. ECG is widely used in all branches of modern medicine such as angiology, sports medicine, critical care medicine, and so on. ECG consists of several waves (P,Q,R,S and T) and intervals between them. Each change of shape of this wave or length of these intervals can mean a certain kind of disease. Nowadays, it is not necessary to be monitored in hospitals. Instead, a patient can be monitored during regular daily activities. One of the biggest disadvantages of monitoring is the length of captured records. Record length is obviously days, weeks, or in extreme cases, months. It is not realistically possible to evaluate the whole record manually. Therefore, an expert system for automatic recognition of a pathological beat has to be used. One of the very dangerous types of arrhythmia is Premature Ventricular Contractions (PVC). A physiological beat arises in the sinoatrial node and spreads through the atrioventricular node, and from there it goes to His branch and Tawara's arms and arrives to the Purkinje fibers. Due to electrical stimulation, mechanical contractions occur in the heart muscle and blood is ejected to the body. In cases where the action potential does not arise in the sinoatrial node but in the artioventricular node, it is classified as PVC. When having PVC beats, patients

can feel palpitations and shortness of breath. An early occurrence of PVC beats can lead to a heart attack or to death in extreme cases. From time to time, the ECG of a healthy person can contain an occurrence of the PVC beat. Therefore, it is necessary to know the right number of occurrences of PVC beats and to start therapy if necessary. The difference between a physiological and a PVC beat is possible to see in Picture 1. In some cases, however, the difference between a physiological and a pathological beat can be very minimal. Right bundle branch block (RBBB) - During this pathological beats the right ventricle is not activated by regular impulses. During the RBBB activation of the right ventricle is delayed. The left ventricle is contracted later than the left one. Left bundle branch block (LBBB) is less common than RBBB. During the LBBB activation of the left ventricle is delayed. The left ventricle is contracted later than the right one. [1] [2] [3]
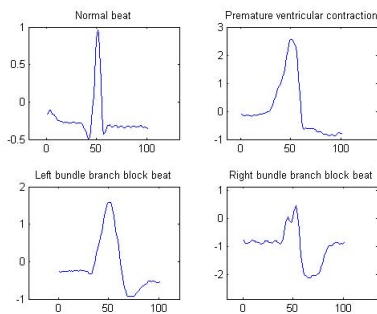
Fig. 1. A difference between physiological and pathological beats

2   Dataset

ECG records were obtained from the MIT-BIH arrhythmia database. The database consists of 48 ECG records. Each record contains a measurement from two ECG leads: ML2 and V5. The sampling frequency was set to 360 Hz and with an 11- bit resolution over a 10 mV range. The each beat was annotated by two cardiologists and is used like a reference database for comparing different algorithms for arrhythmia recognition or QRS detection algorithms. In addition to normal and PVC beats, the database also contains other types of arrhythmias such as Left bundle branch block, Right bundle branch block, supra ventricular premature beat, and so on. [5][4]

3   Experiment and Methods

Due to the known annotation, each beat was extracted from the original record and was preprocessed for classification. We used 35 neighboring points on the
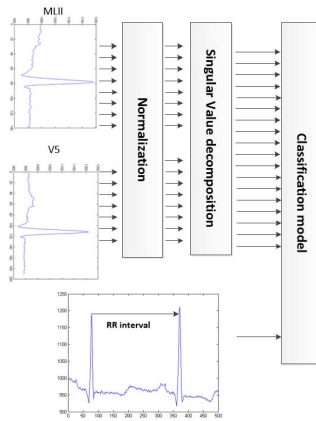
Fig. 2. A proposed classification scheme

left and right side of the R waves for classification and we also added the R-R distance. The mean value was subtracted from each segment. Figure 2 shows principle scheme of the classification. Two ECG channels MLII and V5 were normalized and processed by SVD. As final step the SVD output and R-R distance were used for classification. Two classification approaches were tested: Linear Discriminant Analysis and Support Vector Machine.

The classification performance of the proposed expert system was validated on 48751 beats. The database included 75015 normal beats, 8072 LBBB beats, 7253 RBBB beats and 7128 PVC beats. The database was randomly separated into two subsets: one half of the beats were selected as the training subset and the other half as the testing subset. The performance of the proposed classification method was described by three evaluation techniques: Sensitivity (Equation 1), Specificity (Equation 2) and Accuracy (Equation 3) .

$$Sensitivity(\%) = \frac{TP}{TP + FN} \times 100 \tag{1}$$

$$Specificity(\%) = \frac{TN}{TN + FP} \times 100 \tag{2}$$

$$Accuracy(\%) = \frac{TP}{TP + TN} \times 100 \tag{3}$$

where TP is true positive value, FP is false positive value, TN is true negative value and FN is false negative value.

3.1   Singular Value Decomposition

The Singular Value Decomposition (SVD) is a factorization of a real matrix and is similar to the Principal Component Analysis algorithm. Both of these

algorithms generate eigenvalues. The SVD discovers significant properties of the feature space and represents them as linear combinations of the base vectors. Formally, the matrix of time features is decomposed by singular value decomposition, calculating singular values and singular vectors of this feature matrix. For a better understanding:

Let A be a matrix of all training beats in the time domain, let its size be n x m, where m is number of training beats and n is a number of inputs. The rank r and let the sigma values $\sigma_1.....\sigma_n$ be calculated from the eigenvalues of matrix AA^T as $sigma_i = \sqrt{(\lambda_i)}$

then column-orthonormal matrices U and V can be calculated, where $U^T U = I_n$ and $V^T V = I_m$ and a diagonal matrix $\Sigma = diag(sigma_1.....sigma_n)$, where $sigma_i > 0$ and $sigma_i > sigma_{i+1}$, such that:

$$A = U\Sigma V^T. \tag{4}$$

Now the matrix $RS = U\Sigma$ can be calculated and it serves as the new reduced training set. The second new matrix is $QS = TV$, where T is a original training set. Now, these new matrices can be used in classification tasks as the original testing and training sets. [6]

3.2   Linear Discriminant Analysis

The Linear Discriminant Analysis (LDA) is a technique for pattern recognition. The main goal of this algorithm is to find linear combinations of features, which provide the best separation between classes. These combinations are called the discriminant functions. The classifier was designed by Fischer in 1931. The original form of the algorithm belongs to binnary classifiers. There are n classes. The intra-class matrix can be calculated as

$$\hat{\Sigma}_w = S1 + ... + S_n = \sum_{i=1}^{n} \sum_{x \in c_i} (x - \hat{x}_i)(x - \hat{x}_i)', \tag{5}$$

and inter-class matrix by the equation

$$\hat{\Sigma}_b = \sum_{i=1}^{n} m_i (\bar{x}_i - \bar{x})(\bar{x}_i - \bar{x})', \tag{6}$$

where m_i is the number of samples in each class from the training set, x_i is the mean for each class and $\hat{x}$ is the total mean vector. A linear transformation Φ should be found, in order to maximize the Rayleigh coefficient, which is the ratio of determinants of inter-class and intra-class scatter matrices:

$$J(\Phi) = \frac{\left| \Phi^T \hat{\Sigma}_b \Phi \right|}{\left| \Phi^T \hat{\Sigma}_w \Phi \right|}. \tag{7}$$

The linear transformation Φ can be computed by solving the equation

$$\hat{\Sigma}_b\Phi = \lambda\hat{\Sigma}_w\Phi \tag{8}$$

The last step is the classification itself. The principle is the measurement of metric or cosine distances between new instances and the centroid of classes. The new sample are classified according to the expression:

$$arg\,min\,d\,(z\Phi, \bar{x}_k\Phi)\,. \tag{9}$$

, where $z\Phi$ is a new instance.

3.3 Support Vector Machine

Let N be the number of training samples, where each sample belongs to class $y1 = -1$ or to class $y1 = 1$. The training set has to meet the following condition:

$$\{x_i, y_i\}\,, i = 1..N, x \in R^d, y \in \{-1, 1\} \tag{10}$$

This means that a feature vector contains information about class and each feature has to belong to the set of real numbers. In cases where data is linearly separable, we can use a hyperplane for recognition of these classes. This hyperplane can be expressed by Equation 11.

$$x \cdot w - b = 0, \tag{11}$$

where w is a normal vector to the half plane and b is a constant. A perpendicular distance between the hyperplane and the zero point of the feature vector is possible to express by Equation 12

$$\frac{b}{\|w\|} \tag{12}$$

Now we can define two new hyperplanes, $H1$ and $H2$, which are parallel to the separation hyperplane and which separate both classes and, additionally, there are no samples between them. For the hyperplane $H1$ it is

$$x \cdot w - b = 1 \tag{13}$$

and for the hyperplane $H2$ it is

$$x \cdot w - b = -1 \tag{14}$$

The distance between both planes is called *SVM margin* , which can be computed in the following way:

$$\frac{2}{\|w\|} \tag{15}$$

The main aim of this classifier is to find the hyperplane which maximizes the margin such that the distance between both hyperplanes, $H1$ and $H2$, is the

same. This can be achieved by minimizing $\|w\|$. The data can not lie in the margin. Therefore, the following conditions have to be met:

$$y_i = +1 \Rightarrow x_i \cdot w - b \geq 1 \tag{16}$$

$$y_i = 1 \Rightarrow x_i \cdot w - b \leq -1 \tag{17}$$

These two conditions can be fused into the following condition:

$$y_i (x_i \cdot w - b) - 1 \geq 0 \tag{18}$$

maximizing the margin leads to an optimization problem:

$$min \frac{1}{2} \|w\|^2 \tag{19}$$

,which can be solved by the Langranger multiplicator.

$$L_p (w, b, a) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{N} \alpha_i [y_i (x_i \cdot w - b) - 1], \tag{20}$$

where α is a vector of Lagranger multiplicator. The minimization can be achieved if the partial derivation is made equal to zeros.

$$\frac{\partial L_p}{\partial w} = 0 \Rightarrow \sum_{n=1}^{N} \alpha_i y_i x_i \tag{21}$$

$$\frac{\partial L_p}{\partial b} = 0 \Rightarrow \sum_{n=1}^{N} \alpha_i y_i \tag{22}$$

By substituting equations 21 and 22 back into equation 20, we get a new formulation of the problem, which is transformed into the following form

$$L_d = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j}^{N} \alpha_i y_j \alpha_i y_j x_j x_i \cdot y_j \tag{23}$$

If $y_i y_j x_i \cdot x_j$ is given by the hyperplane H_{ij}, then the optimization problem is reformulated in the equation:

$$min_a \frac{1}{2} \alpha^T H \alpha - \frac{1}{2} \sum_{i=1}^{N} \alpha_i \tag{24}$$

which can be solved by classic quadratic techniques. In case the data is not linearly separable, it is better to change the kernel function. This step transforms the training data to a feature space with a higher dimension where it can be solved by a linear separation. During this modification the scalar product $x_i \cdot x_j$ is replaced by a non-linear kernel function such as a polynomial function or a non linear function with radial basis, which can be expressed by Equation 25.

$$K(x, x') = exp(-\frac{\|x - x'\|_2^2}{2\sigma^2}) \tag{25}$$

4   Results

In this chapter, the achieved results are described by the confusion matrices and by evaluations characteristics. Table 1 describes the results which were reached by the LDA algorithm classifying pure ECG signals in a non-reduced form. Table 1 shows that the classification performance is very poor. The LDA wrongly classified 1136 LBBB which were classified as normal. Classifying of PVC beats achieved similar results. Table 4 shows results which were achieved by the SVM on the same data. The values of Accuracy, Sensitivity and Specificity went up significantly for all classes compared to the LDA algorithm. Only 295 LBBB beats were classified wrong. The number of incorrectly classified PVC was 340 against 963 which were classified wrong by the LDA. The increase of all evaluation characteristics is almost 20 %. The SVD computed 203 new eigenvalues, which brings information about the original signal. Figure 4 shows dependencies between achieved values of sensitivity, specitivity, accuracy and the number of eigenvalues. For small number of eigenvalues these values are very low and grows with the increasing number of eigenvalues. The best values were achieved by the LDA with 70 eigenvalues but if these values are compared with the results in Table 1 it is possible to see that there is no significant improvement. Generally, with the above mentioned settings, the LDA proved to be unsuitable for this task and it can not be used in practise. On the other hand, the SVM already achieved very satisfactory results in classifying the pure ECG. In Figure 3 shows the results which were achieved by the SVM algorithm for different number of eigenvalues. The best values of Sensitivity, Specificity and Accuracy were achieved with 110 eigenvalues. Table 4 describes these values. There it is possible to see that these values are higher than in the previous cases. Compared to the SVM classifying the pure ECG signals, SVM classification on the SVD-reduced dataset provides a higher classification performance. The values of Sensitivity increased from 94.74 % to 96.71 % for LBBB beats, from 98.74 % to 99.09 % for Normal beats, from 92.86 % to 93.05 % for RBBB beats and from 90.47 % to 90.75 % for PVC beats.

Table 1. Confusion matrix of the Linear discriminant anylsis

	LBBB	Normal	RBBB	PVC
LBBB	2855	749	0	263
Normal	1136	36226	399	628
RBBB	1	61	3228	72
PVC	45	482	1	2605
Accuracy	73.82	94.36	96.01	83.14
Sensitivity	70.72	96.55	88.97	73.01
Specificity	97.73	80.74	99.70	98.83

Table 2. Confusion matrix of the SVM

	LBBB	Normal	RBBB	PVC
LBBB	3825	346	0	36
Normal	203	37046	222	294
RBBB	0	21	3369	10
PVC	9	110	37	3228
Accuracy	91.02	98.09	99.08	95.39
Sensitivity	94.74	98.74	92.86	90.47
Specificity	99.15	93.59	99.93	99.65

Table 3. Confusion matrix of the LDA+SVD

	LBBB	Normal	RBBB	PVC
LBBB	2869	766	0	263
Normal	1123	36214	397	626
RBBB	1	60	3230	72
PVC	44	478	1	2607
Accuracy	73.60	94.41	96.05	83.29
Sensitivity	71.07	96.52	89.03	73.07
Specificity	97.70	80.90	99.71	98.84

Table 4. Confusion matrix of the SVM+SVD

	LBBB	Normal	RBBB	PVC
LBBB	3904	276	0	32
Normal	126	37177	247	293
RBBB	0	16	3376	5
PVC	7	49	5	3238
Accuracy	92.69	98.24	99.38	98.15
Sensitivity	96.71	99.09	93.05	90.75
Specificity	99.31	94.07	99.95	99.86
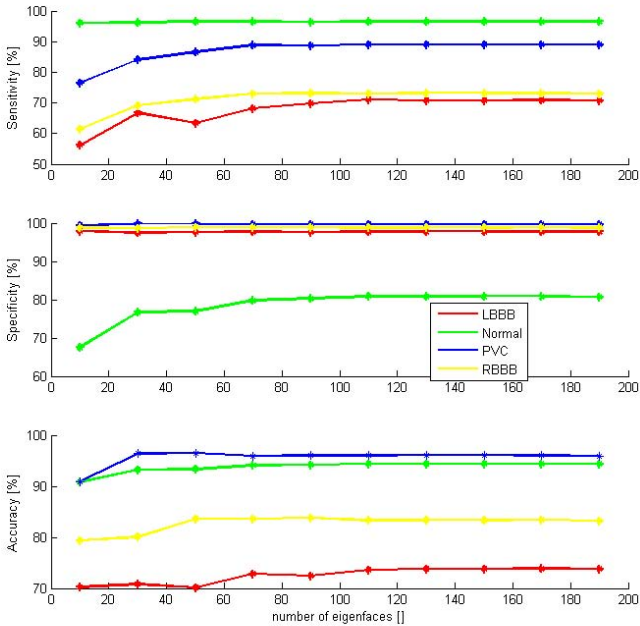
Fig. 3. Sensitivity, specitificity and accuracy versus number of eigenvalues with the LDA classifier
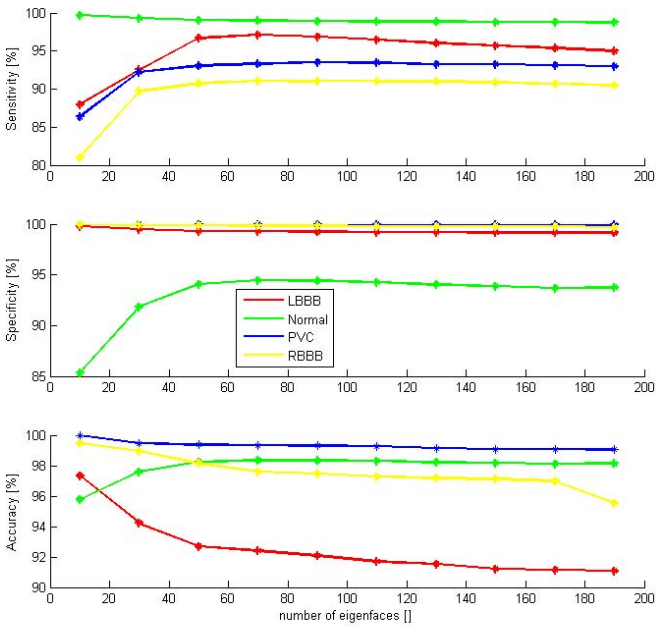
Fig. 4. Sensitivity, specitificity and accuracy versus number of eigenvalues with the SVM classifier

5   Conclusion

In this work, we described an algorithm for recognizing pathological beats in the ECG signals. Our solutions is based on the SVD, which is one of the most powerful matrix factorization techniques. The new reduced dataset was classified by the Support Vector Machine. Combining these two methods proved advantageous over the other mentioned techniques. Comparing to other work, the tests were performed on a relatively high number of testing samples (almost 80000 heartbeats). In our next research we will focus on classifying pathological beats on intra-individual level.

References

1. Malmivuo, J., Plonsey, R.: Bioelectromagnetism-Principles and Applications of Bioelectric and Biomagnetic Fields. Oxford University Press, New York (1995)
2. Penhaker, M., Irnramovsky, M., Tiefenbach, P.: Lekarske Diagnosticke Pristroje VSB-TU Ostrava, Ostrava (2004) ISBN 80-248-0751-3
3. Svatos, J.: Biological Signaly Editing office CVUT, Praha (1995)
4. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.-K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101(23), e215–e220 (2000), Circulation Electronic Pages, http://circ.ahajournals.org/cgi/content/full/101/23/e215
5. Moody, G.B., Mark, R.: The impact of the MIT-BIH Arrhythmia Database. IEEE Eng. in Med. and Biol. 20(3), 45–50 (2001)
6. Moravec, P., Gajdoš, P., Snášel, V., Saeed, K.: Normalization Impact on SVD-Based Iris Recognition. In: Saeed, K., Abraham, A., Porwik, P. (eds.) Proceedings of International Conference on Biometrics and Kansei Engineering, pp. 60–64. IEEE Press (2009) ISBN 978-0-7695-3692-7
7. Vapnik, V., Golowich, S.E., Smola, A.J.: Support Vector Method for Function Approximation, Regression Estimation and Signal Processing. In: NIPS, pp. 281–287 (1996)
8. Vapnik, V.: Statistical learning theory, pp. I–XXIV, 1–736. Wiley (1998)

Author Index